

NIP S 2008

NEURAL
INFORMATION
PROCESSING
SYSTEMS
CONFERENCE


Mini Symposia & Workshops

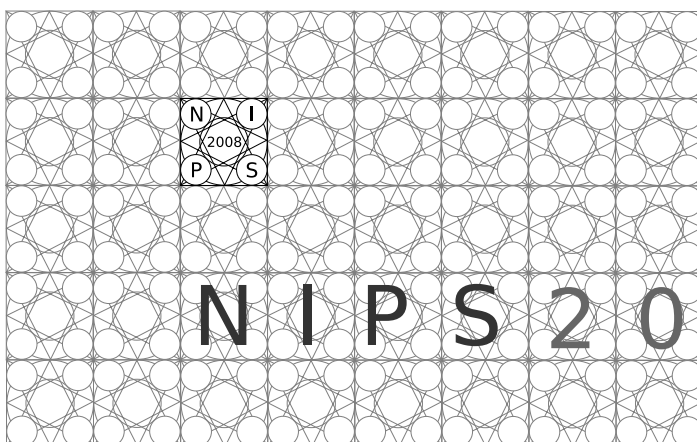
TUTORIALS
December 8, 2008
Hyatt Regency
Vancouver, BC, Canada

CONFERENCE SESSIONS
December 8-11, 2008
Hyatt Regency
Vancouver, BC, Canada

MINI SYMPOSIA
December 11, 2008
Hyatt Regency
Vancouver, BC, Canada

WORKSHOP
December 12-13, 2008
The Westin Resort & Spa
The Hilton Whistler Resort & Spa
Whistler, BC, Canada

™
Neural Information
Processing Systems
Foundation



N I P S 2 0 0 8

Mini Symposia & Workshops

TUTORIALS

December 8, 2008
Hyatt Regency
Vancouver, BC, Canada

CONFERENCE SESSIONS

December 8-11, 2008
Hyatt Regency
Vancouver, BC, Canada

MINI SYMPOSIA

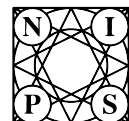
December 11, 2008
Hyatt Regency
Vancouver, BC, Canada

WORKSHOP

December 12-13, 2008
The Westin Resort & Spa
The Hilton Whistler Resort & Spa
Whistler, BC, Canada

Sponsored by the Neural Information Processing Systems Foundation, Inc.

There are 4 Mini Symposia and 25 workshops covering a rather eclectic range of topics from algebra to neuroscience to computational hardware to photography to behavioral science to bioinformatics to learning theory to internet related problems. They provide an exciting preview for future trends in Neural Information Processing Systems and in this way they complement the main conference.



Neural Information
Processing Systems
Foundation

Contents

Organizing Committee	7
Program Committee	7
NIPS Foundation Offices and Board Members	8
Sponsors	9
Core Logistics Team	9
Schedule	11
Workshop Overview	13
MS1 Algebraic methods in machine learning	17
MS2 Computational Photography	19
MS3 Machine Learning in Computational Biology	22
MS4 Principled Theoretical Frameworks for the Perception-Action Cycle	24
WS1 Beyond Search: Computational Intelligence for the Web	26
WS2a Speech and Language: Learning-based Methods and Systems	31
WS2b Speech and Language: Unsupervised Latent-Variable Models	35
WS3 Statistical Analysis and Modeling of Response Dependencies in Neural Populations	38
WS4 Algebraic and combinatorial methods in machine learning	47
WS5 Analyzing Graphs: Theory and Applications	51
WS6 Machine Learning in Computational Biology	59
WS7 Machine Learning Meets Human Learning	61
WS8 Machine Learning Open Source Software	67
WS9 Optimization for Machine Learning	72
WS10 Structured Input - Structured Output	74
WS11 Causality: objectives and assessment	77
WS12 Cost Sensitive Learning	79
WS13 New Challenges in Theoretical Machine Learning: Learning with Data-dependent Concept Spaces	84
WS14 New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis	86
WS15 Cortical Microcircuits and their Computational Functions	90
WS16 Approximate inference - how far have we come?	92
WS17 Kernel Learning: Automatic Selection of Optimal Kernels	94
WS18 Parallel Implementations of Learning Algorithms: What have you done for me lately?	99
WS19 Model Uncertainty and Risk in Reinforcement Learning	101

WS20 Principled Theoretical Frameworks for the Perception-Action Cycle	103
WS21 Probabilistic Programming: Universal Languages and Inference; Systems; and Applications	105
WS22 Stochastic Models of Behaviour	111
WS23 Learning from Multiple Sources	115
Index	123
Maps	129
Notes	132

Organizing Committee

General Chair	DAPHNE KOLLER, Stanford University
Program Co-Chairs	DALE SCHUURMANS, University of Alberta YOSHUA BENGIO, University of Montreal
Tutorial Chair	GEOFF GORDON, Carnegie Mellon University
Workshop Co-Chairs	MANEESH SAHANI, London College University ALEX SMOLA, Yahoo! Research
Demonstration Chair	RALF HERBRICH, Microsoft Research
Publications Chair	LEON BOTTOU, NEC Labs America
Electronic Proceedings Co-Chairs	ARON CULOTTA, Southeastern Louisiana University ANDREW MCCALLUM, University of Massachusetts, Amherst
Publicity Chair	ANTONIO TORRALBA, MIT
Volunteers Chair	YASEMIN ALTUN, Max Planck Institute

Program Committee

Program Co-Chairs	DALE SCHUURMANS, University of Alberta YOSHUA BENGIO, University of Montreal
Program Committee Area Chairs	JEAN-YVES AUDIBERT, Ecole Nationale des Ponts et Chaussées FRANCIS BACH, INRIA & Ecole Normale Supérieure KRISTIN BENNETT, Rensselaer Polytechnic Institute MICHAEL BOWLING, University of Alberta AARON COURVILLE, Université de Montréal KOBY CRAMMER, University of Pennsylvania SANJOY DASGUPTA, University of California, San Diego NATHANIEL DAW, New York University ELEAZAR ESKIN, University of California, Los Angeles DAVID FLEET, University of Toronto PAOLO FRASCONI, Università di Firenze ARTHUR GRETTON, Max Planck Institute TONY JEBARA, Columbia University CHRIS MANNING, Stanford University RON MEIR, Technion NOBORU MURATA, Waseda University ERKKI OJA, Helsinki University of Technology DOINA PRECUP, McGill University STEFAN SCHAAL, University of Southern California FEI SHA, University of Southern California ALAN STOCKER, New York University INGO STEINWART, Los Alamos National Laboratory ERIK SUDDERTH, University of California, Berkeley YEE-WHYE TEH, University College London ANTONIO TORRALBA, Massachusetts Institute of Technology LARRY WASSERMAN, Carnegie Mellon University MAX WELLING, University of California, Irvine

NIPS Foundation Officers and Board Members

President	TERRENCE SEJNOWSKI, The Salk Institute
Vice President for Development	SEBASTIAN THRUN, Stanford University
Treasurer	MARIAN STEWART BARTLETT, University of California, San Diego
Secretary	MICHAEL MOZER, University of Colorado, Boulder
Legal Advisor	PHIL SOTEL, Pasadena, CA
Executive Board	SUE BECKER, McMaster University, Ontario, Canada
	THOMAS G. DIETTERICH, Oregon State University
	JOHN C. PLATT, Microsoft Research
	LAWRENCE SAUL, University of Pennsylvania
	BERNHARD SCHÖLKOPF, Max Planck Institute
	SARA A. SOLLA, Northwestern University Medical School
	YAIR WEISS, Hebrew University of Jerusalem
Advisory Board	GARY BLASDEL, Harvard Medical School
	JACK COWAN, University of Chicago
	STEPHEN HANSON, Rutgers University
	MICHAEL I. JORDAN, University of California, Berkeley
	MICHAEL KEARNS, University of Pennsylvania
	SCOTT KIRKPATRICK, Hebrew University, Jerusalem
	RICHARD LIPPMANN, Massachusetts Institute of Technology
	TODD K. LEEN, Oregon Graduate Institute
	BARTLETT MEL, University of Southern California
	JOHN MOODY, International Computer Science Institute, Berkeley and Portland
	GERALD TESAURO, IBM Watson Labs
	DAVE TOURETZKY, Carnegie Mellon University
Emeritus Members	TERRENCE L. FINE, Cornell University
	EVE MARDER, Brandeis University

Sponsors

NIPS gratefully acknowledges the generosity of those individuals and organizations who have provided financial support for the NIPS 2008 conference. The financial support enabled us to sponsor student travel and participation, the outstanding student paper awards, the mini symposia, the demonstration track and the opening buffet. We gratefully acknowledge Yahoo's support in sponsoring the mini-symposia.

MICROSOFT
GOOGLE
PASCAL
YAHOO
INTEL
BOEING
IBM
WILLOW GARAGE
D.E. SHAW
NUMENTA
TOYOTA RESEARCH
SPRINGER

Core Logistics Team

The running of NIPS would not be possible without the help of many volunteers, students, researchers and administrators who donate their valuable time and energy to assist the conference in various ways. However, there is a core team at the Salk Institute whose tireless efforts make the conference run smoothly and efficiently every year. This year, NIPS would particularly like to acknowledge the exceptional work of:

NELSON LOYOLA, Workflow Master
LEE CAMPBELL
CHRIS HIESTAND
SHERI LEONE
MARY ELLEN PERRY

Schedule

Thursday, December 11th

13.30-16.30	Mini Symposia	Hyatt Vancouver
14.00-18.00	Buses depart Vancouver Hyatt for Westin Resort and Spa	
19.00-22.00	Lift Ticket Sales	Westin lobby
17.00-20.30	Registration	Westin Emerald foyer

Friday, December 12th

6.30-8.00	Breakfast	Westin Emerald
7.00-11.00	Registration	Westin Emerald foyer
7.30-11.30	Workshop sessions	Westin and Hilton
8.00-9.30	Lift Ticket Sales	Westin lobby
14.30-18.30	Workshop sessions continue	Westin and Hilton

Saturday, December 13th

6.30-8.45	Breakfast	Westin Emerald
7.00-11.00	Registration	Westin Emerald foyer
7.30-11.30	Workshop sessions	Westin and Hilton
14.30-18.30	Workshop sessions continue	Westin and Hilton
19.30 - 22.30	Banquet and wrap up meeting	Westin Emerald

**Some workshops run on different schedules.
Please check timings on the subsequent pages.**

Workshop Overview

Thursday, December 11th

Algebraic methods in machine learning

13.30–16.30 Regency A/B **MS1**

Computational Photography

13.30–16.30 Regency C **MS2**

Machine Learning in Computational Biology

13.30–16.30 Regency D **MS3**

Principled Theoretical Frameworks for the Perception-Action Cycle

13.30–16.30 Regency E/F **MS4**

Friday, December 12th

Beyond Search: Computational Intelligence for the Web

07:30–10:30 and 15:30–18:30 Westin: Emerald A **WS1**

Speech and Language: Learning-based Methods and Systems

07:30–10:40 and 15:30–18:30 Hilton: Cheakamus **WS2a**

Statistical Analysis and Modeling of Response Dependencies in Neural Populations

07:30–10:30 and 15:30–18:30 Hilton: Mt. Currie N **WS3**

Algebraic and combinatorial methods in machine learning

07:30–10:30 and 15:30–18:30 Westin: Callaghan **WS4**

Analyzing Graphs: Theory and Applications

07:30–10:30 and 15:30–18:30 Westin: Alpine AB **WS5**

Machine Learning in Computational Biology

07:45–10:30 and 15:45–18:30 Hilton: Mt. Currie S **WS6**

Machine Learning Meets Human Learning

07:30–10:40 and 15:30–18:40 Hilton: Black Tusk **WS7**

Machine Learning Open Source Software

07:30–10:30 and 15:30–18:30 Westin: Alpine CD **WS8**

Optimization for Machine Learning

07:30–10:30 and 15:30–18:30 Hilton: Diamond Head **WS9**

Structured Input - Structured Output

07:30–10:30 and 15:30–18:30 Hilton: Sutcliffe B **WS10**

Causality: objectives and assessment

07:30–10:30 and 16:00–19:00 Westin: Nordic **WS11**

Cost Sensitive Learning

07:30–10:30 and 16:00–19:00 Westin: Alpine E **WS12**

New Challenges in Theoretical Machine Learning: Learning with Data-dependent Concept Spaces

07:30–10:30 and 15:30–18:30 Hilton: Sutcliffe A **WS13**

Saturday, December 13th

Beyond Search: Computational Intelligence for the Web07:30–10:30 and 15:30–18:30 Westin: Emerald A **WS1****Speech and Language: Unsupervised Latent-Variable Models**07:30–10:30 and 15:30–18:30 Hilton: Cheakamus **WS2b****Statistical Analysis and Modeling of Response Dependencies in Neural Populations**07:30–10:30 and 15:30–18:30 Hilton: Mt. Currie N **WS3****New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis**07:30–10:30 and 15:30–18:30 Hilton: Sutcliffe B **WS14****Cortical Microcircuits and their Computational Functions**07:30–10:30 and 16:00–19:00 Hilton: Diamond Head **WS15****Approximate inference - how far have we come?**07:30–10:30 and 16:00–19:00 Westin: Alpine AB **WS16****Kernel Learning: Automatic Selection of Optimal Kernels**07:30–10:30 and 15:30–18:30 Westin: Nordic **WS17****Parallel Implementations of Learning Algorithms: What have you done for me lately?**07:30–10:30 and 3:30–6:30 Hilton: Sutcliffe A **WS18****Model Uncertainty and Risk in Reinforcement Learning**07:30–10:30 and 15:30–18:30 Westin: Callaghan **WS19****Principled Theoretical Frameworks for the Perception-Action Cycle**07:30–10:30 and 15:30–18:30 Hilton: Mt. Currie S **WS20****Probabilistic Programming: Universal Languages and Inference; Systems; and Applications**07:30–10:30 and 15:30–18:30 Westin: Alpine CD **WS21****Stochastic Models of Behaviour**07:30–11:00 and 15:30–18:45 Hilton: Black Tusk **WS22****Learning from Multiple Sources**07:30–10:30 and 15:30–18:30 Westin: Alpine E **WS23**

DECEMBER 11, 2008, 13.30–16.30

REGENCY A/B MS1

Algebraic methods in machine learning

<http://www.gatsby.ucl.ac.uk/~risi/AML08/>

Risi Kondor

GATSBY UNIT, UCL

risi@gatsby.ucl.ac.uk

Guy Lebanon

GEORGIA INSTITUTE OF TECHNOLOGY

lebanon@cc.gatech.edu

Jason Morton

STANFORD UNIVERSITY

jason@math.stanford.edu

Abstract

Preliminary Schedule. There has recently been a surge of interest in algebraic methods in machine learning. This includes new approaches to ranking problems, the budding field of algebraic statistics and various applications of non-commutative Fourier transforms. The aim of the workshop is to bring together these distinct communities, explore connections, and showcase algebraic methods to the machine learning community at large. The symposium is intended to be accessible to researchers with no prior exposure to abstract algebra. The program includes three short tutorials that will cover the basic concepts necessary for understanding cutting edge research in the field.

- | | |
|--------------------|---|
| 13.30-14.00 | Non-commutative harmonic analysis
RISI KONDOR |
| 14.05-14.35 | Modeling distributions on permutations and partial ranking
GUY LEBANON |
| 14.45-15.05 | Algebraic models for multilinear dependence
JASON MORTON |
| 15.10-15.40 | Symmetry Group-based Learning for Regularity Discovery from Real World Patterns
YANXI LIU |
| 15.45-16.15 | Estimation and model selection in stagewise ranking - a representation story
MARINA MEILA |

Non-commutative harmonic analysis

Risi Kondor, GATSBY UNIT, UCL

Fourier analysis is one of the central pillars of applied mathematics. Representation theory makes it possible to generalize Fourier transformation to non-commutative groups, such as permutations and 3D rotations. This talk will survey new applications of this theory in machine learning for problems such as identity management in multi-object tracking, transformation invariant representations of images, and similarity measures between graphs. The talk is intended for a wide audience, no background in group theory or representation theory will be assumed.

Modeling distributions on permutations and partial ranking

Guy Lebanon, GEORGIA INSTITUTE OF TECHNOLOGY

We explore several probabilistic models over the symmetric group of permutations and its cosets - representing partial rankings. We will cover both traditional statistical models such as the Luce-Plackett and Bradley Terry models, as well as more modern ones. Special attention will be given to non-parametric models and their use in modeling and visualizing preference data.

Algebraic models for multilinear dependence

Jason Morton, STANFORD UNIVERSITY

We discuss a new statistical technique inspired by research in tensor geometry and making use of cumulants, the higher order tensor analogs of the covariance matrix. For non-Gaussian data not derived from independent factors, tensor decomposition techniques for factor analysis such as Principal Component Analysis and Independent Component Analysis are inadequate. Seeking a closed space of models which is computable and captures higher-order dependence leads to a proposed extension of PCA and ICA, Principal Cumulant Component Analysis (PCCA). Estimation is performed by maximization over a Grassmannian. Joint work with L.-H. Lim.

Symmetry Group-based Learning for Regularity Discovery from Real World Patterns

Yanxi Liu, PENN STATE

We explore a formal and computational characterization of real world regularity using discrete symmetry groups (hierarchy) as a theoretical basis, embedded in a well-defined Bayesian framework. Our existing work on “Near-regular texture analysis and manipulation” (SIGGRAPH 2004) and “A Lattice-based MRF Model for Dynamic Near-regular Texture Tracking” (TPAMI 2007) already demonstrate the power of such a formalization on a diverse set of real problems, such as texture analysis, synthesis, tracking, perception and manipulation in terms of regularity. Symmetry and symmetry group detection from real world data turns out to be a very challenging problem that has been puzzling computer vision researchers for the past 40 years. Our novel formalization will lead the way to a more robust and comprehensive algorithmic treatment of the whole regularity spectrum, from regular (perfect symmetry), near-regular (approximate symmetry), to various types of irregularities. The proposed method will be justified by several real world applications such as gait recognition, grid-cell clustering, symmetry of dance, automatic geo-tagging and image de-fencing.

Estimation and model selection in stagewise ranking - a representation story

Marina Meila, UNIVERSITY OF WASHINGTON

This talk is another example of the well-known statement that “representation matters”. We describe the code of a permutation, first used in statistics by Fligner and Verducci to define stagewise raking models. The code represents a permutation as a sequence of $n - 1$ independent numbers. This property makes statistical models based on the code have singular advantages w.r.t other probabilistic models over the symmetric group. In particular, the parameters can be better understood and can be estimated more easily. We illustrate this by comparisons to other exponential models of the symmetric group and by describing a suite of algorithms that allow one to estimate stagewise ranking models based on the code under a variety of missing data scenarios. Joint work with Bhushan Madhani and Kobi Abayomi.

DECEMBER 11, 2008, 13.30–16.30

REGENCY C **MS2**

Computational Photography

<http://www.kyb.tuebingen.mpg.de/bs/people/bs/nips-symposium2008.html>

William Freeman

billf@mit.edu

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Bernhard Schölkopf

bernhard.schoelkopf@tuebingen.mpg.de

MPI FOR BIOLOGICAL CYBERNETICS, TÜBINGEN

Abstract

Computation will change photography. The sensor no longer has to record the final image, but only data that can lead to the final image. Computation can solve longstanding photographic problems (e.g., deblurring) and well as open the door for radical new designs and capabilities for image capture, processing, and viewing (e.g., lightfield cameras). Many of these possibilities offer great machine learning problems, and much of the progress in computational photography will rely on solutions to these challenging machine learning problems. We have gathered five leading researchers in this new field to describe their work at the intersection of photography and machine learning.

- | | |
|--------------------|--|
| 13.30-14.00 | Random Projections, Computational Photography, and Computational Neuroscience
YAIR WEISS |
| 14.00-14.30 | Navigating the World's Photographs
STEVE SEITZ |
| 14.30-15.00 | Computational Photography: From Epsilon to Coded Photography
RAMESH RASKAR |
| 15.00-15.30 | Coffee |
| 15.30-16.00 | Computational imaging using camera arrays
WOJCIECH MATUSIK |
| 16.00-16.30 | Bayesian Analysis of Cameras
BILL FREEMAN |

Random Projections, Computational Photography, and Computational Neuroscience

Yair Weiss, HEBREW UNIVERSITY, JERUSALEM, ISRAEL

One way of thinking about computational photography is as an attempt to design the optimal camera to capture visual information. This is closely related to Linsker's InfoMax approach for modeling biological sensory systems. Bell and Sejnowski have shown how the InfoMax approach can be used to predict the Gabor-like receptive fields in V1.

I will describe some of our work on UNDERCOMPLETE InfoMax — when the number of sensors is less than the dimensionality of the input. In this setting, for a surprisingly large number of signal models, it can be shown that random projections are InfoMax optimal. In particular, for whitened natural image patches, random projections are the most informative while searching for the LEAST informative filters gives Gabor filters.

Joint work with Hyung Sung Chang and Bill Freeman.

Navigating the World's Photographs

Steve Seitz, UNIVERSITY OF WASHINGTON, SEATTLE, USA

There's a big difference between looking at a photograph of a place and being there. But what if you had access to every photo ever captured of that place and could conjure up any view at will? With billions of photographs currently available online, the Internet is beginning to resemble such a database, capturing most of the world's significant sites from a huge number of vantage points and viewing conditions. For example, a Google image search for "notre dame" or "grand canyon" each returns more than a million photos, showing the sites from myriad viewpoints, different times of day and night, and changes in season, weather and decade.

This talk explores ways of transforming this massive, unorganized photo collection into visualizations of the world's sites, cities, and landscapes. After a brief recap of our work on Photo Tourism and Photosynth, I will focus on current efforts and newest results that seek to discover human preference and behavior patterns in real world scenes.

Computational Photography: From Epsilon to Coded Photography

Ramesh Raskar, M.I.T., CAMBRIDGE, USA

In this talk, I will focus on Coded Photography. 'Less is more' in Coded Photography. By blocking light over time or space, we can preserve more details about the scene in the recorded single photograph.

1. Coded Exposure <http://raskar.info/deblur>: By blocking light in time, by fluttering the shutter open and closed in a carefully chosen binary sequence, we can preserve high spatial frequencies of fast moving objects to support high quality motion deblurring.
2. Coded Aperture Optical Heterodyning <http://raskar.info/Mask/>: By blocking light near the sensor with a sinusoidal grating mask, we can record 4D light field on a 2D sensor. And by blocking light with a mask at the aperture, we can extend the depth of field and achieve full resolution digital refocussing.
3. Coded Illumination <http://raskar.info/NprCamera>: By observing blocked light at silhouettes, a multi-flash camera can locate depth discontinuities in challenging scenes without depth recovery.
4. Coded Sensors <http://www.umiacs.umd.edu/~Eaagrawal/gradcam/gradcam.html>: By sensing intensities with lateral inhibition, a "Gradient Camera" can record large as well as subtle changes in intensity to recover a high-dynamic range image.
5. Coded Spectrum <http://www.cs.northwestern.edu/~Eamohan/agile/>: By blocking parts of a "rainbow", we can create cameras with digitally programmable wavelength profile.

I will show several applications and describe emerging techniques to recover scene parameters from coded photographs.

Recent joint work with Jack Tumblin, Amit Agrawal, Ashok Veeraraghavan and Ankit Mohan.

Computational imaging using camera arrays

Wojciech Matusik, ADOBE, NEWTON, USA

With the success of digital photography and improvements in digital display technologies during the last years, we have witnessed a transformation in the way photographs and videos are captured, processed, and viewed. The emerging field of computational photography proceeds even further by fundamentally rethinking the design of the hardware, the associated algorithms, and representations for images and video.

In this context, camera arrays emerge as a new type of imaging device with greatly extended capabilities compared to traditional cameras. Camera arrays allow for images that are higher quality, richer, and easily post-processed. However, it is crucial to develop novel algorithms that exploit the wealth of captured information in order to both efficiently solve traditionally difficult problems in computer vision and to allow for new applications that were not considered using current cameras.

In my talk I will present a number of algorithms that showcase unique capabilities of camera arrays. First, I address the problem of automatic, real-time, passive, and robust image segmentation – a long-standing problem in computer vision. I will show how to modify the standard matting equation to work directly with variance measurements captured with a camera array. This leads to development of the first real-time system for video matting. In the second part of my talk, I will present a method to track a 2D object through significant occlusion using a camera array. The proposed method does not require explicit modeling or reconstruction of the scene and enables tracking in complex, dynamic scenes. Finally, I will describe a complete 3-D TV system that allows for real-time acquisition, transmission, and display of dynamic scenes. In this context, I will describe a signal processing framework to address some of the major issues in cinematography for 3-D displays.

Bayesian Analysis of Cameras

Bill Freeman, M.I.T., CAMBRIDGE, USA

Computational approaches to photography have led to many new camera designs: plenoptic, coded aperture, multi-lens, etc. The goal of these cameras may be to reconstruct not just an image but the entire lightfield light rays at every angle at every position. How can we compare how these very different cameras perform at that task? How do the parameters of each camera affect its performance? The breadth of imaging designs requires new tools to understand the tradeoffs between different cameras.

This talk introduces a unified framework for analyzing computational imaging approaches. Each sensor element is modeled as an inner product over the 4D light field. The imaging task is then posed as Bayesian inference: given the observed noisy light field projections and a prior on light field signals, estimate the original light field. Under common imaging conditions, we compare the performance of various camera designs using 2D light field simulations. This framework allows us to better understand the tradeoffs of each camera type and to analyze their strengths and limitations.

Joint work with Anat Levin and Fredo Durand. The manuscript: (ECCV 2008) <http://people.csail.mit.edu/billf/papers/lightfields-Levin-Freeman-Durand-ECCV08.pdf>.

Machine Learning in Computational Biology

<http://www.mlcb.org>

Gal Chechik

GOOGLE RESEARCH (MOUNTAIN VIEW)

gal@ai.stanford.edu

Christina Leslie

MEMORIAL SLOAN-KETTERING CANCER CENTER (NEW YORK CITY)

cleslie@cbio.mskcc.org

Quaid Morris

UNIVERSITY OF TORONTO

quaid.morris@utoronto.ca

William Noble

UNIVERSITY OF WASHINGTON

noble@gs.washington.edu

Gunnar Rätsch

FRIEDRICH MIESCHER LABORATORY, MAX PLANCK SOCIETY (TÜBINGEN)

Gunnar.Raetsch@tuebingen.mpg.de

Abstract

The field of computational biology has seen dramatic growth over the past few years, both in terms of new available data, new scientific questions, and new challenges for learning and inference. In particular, biological data is often relationally structured and highly diverse, well-suited to approaches that combine multiple weak evidence from heterogeneous sources. These data may include sequenced genomes of a variety of organisms, gene expression data from multiple technologies, protein expression data, protein sequence and 3D structural data, protein interactions, gene ontology and pathway databases, genetic variation data (such as SNPs), and an enormous amount of textual data in the biological and medical literature. New types of scientific and clinical problems require the development of novel supervised and unsupervised learning methods that can use these growing resources.

The goal of meeting is to present emerging problems and machine learning techniques in computational biology. It starts off with invited talks by senior computational biologists and continues with a full-day workshop on Friday. The three symposium talks are as follows:

- | | |
|--------------------|---|
| 13:30-14:30 | Modular Biology: the Function and Evolution of Molecular Networks
AVIV REGEV, BROAD INSTITUTE OF MIT & HARVARD, CAMBRIDGE, MA, U.S.A. |
| 14:30-15:30 | Computational Studies Discover an new Mode of Gene Regulation
STEVEN BRENNER, UNIVERSITY OF CALIFORNIA, BERKELY, U.S.A. |
| 15:30-16:30 | Statistical Models for Predicting HIV Phenotypes
THOMAS LENGAUER, MAX PLANCK INSITUTE FOR INFORMATICS, SAARBRÜCKEN, GERMANY |

Modular Biology: the Function and Evolution of Molecular Networks

Aviv Regev, BROAD INSTITUTE OF MIT & HARVARD, CAMBRIDGE, MA, U.S.A.

Molecular networks provide the information processing backbone of cells and organisms, transforming intra- and extra-cellular signals into coherent cellular responses. The qualitative and quantitative understanding of the function and evolution of molecular networks is among the most fundamental questions in biology. Genomics provides powerful tools with which to probe the components and behavior of molecular networks. However, to successfully gain scientific insight from the huge volumes of heterogeneous data they generate requires a combination of experimental design, biological knowledge, and the power of computation. To

address this challenge we focus on the unifying abstraction of the functional module - a collection of biological entities that act in concert to perform an individual identifiable function, such as a molecular machine, a signaling cascade, a regulatory unit, or a biosynthesis pathway.

In this talk I will describe the development and application of computational methods for the reconstruction of the architecture, function and evolution of modules in molecular networks. I will show how we leverage diverse genomics data to discover component modules in yeast, malaria and cancer; identify how relevant information is encoded at different layers of the network and translated into cellular responses; determine how multiple networks are integrated together; and reconstruct how contemporary complex systems have evolved over time to achieve their specific organization and remarkable functionality in organisms from yeast to mammals.

Computational Studies Discover an new Mode of Gene Regulation

Steven Brenner, UNIVERSITY OF CALIFORNIA, BERKELEY, U.S.A.

Statistical Models for Predicting HIV Phenotypes

Thomas Lengauer, MAX PLANCK INSTITUTE FOR INFORMATICS, SAARBRÜCKEN, GERMANY

We describe statistical models for predicting two important phenotypes of HIV, namely HIV resistance to combination drug therapies and HIV tropism.

HIV resistance: Given the relevant portion of an HIV genome, we predict the resistance of HIV to any of a number of antiviral drugs that are in clinical use. Furthermore we rank combination drug therapies with respect to their expected effectiveness against the given HIV variant. This involves a look into the future of the expected evolution of the virus when confronted with the given drug regimen.

HIV tropism: When entering the human host cell, HIV uses one of two coreceptor molecules on the cell surface. Which one the viral variant uses is indicative of the progression of the disease. We present a statistical model that predicts which of the two coreceptors the viral variant uses.

Both models are trained using various linear and nonlinear statistical learning procedures. The training data are carefully assembled databases comprising relevant genotypic, phenotypic and clinical parameters. While the resistance model only incorporates sequence features, one version of the tropism model also involves information on the structure of the relevant portion of the viral gp120 protein that docks to the human cell.

Both models are available via the webserver www.geno2pheno.org. The webserver has been developed in the context of the Aevir consortium, a German National research consortium targeted at the bioinformatical analysis of HIV resistance data, and is currently in prototypical use for research purposes. Members of the consortium and their associated practices treat about two thirds of the AIDS patients in Germany.

Principled Theoretical Frameworks for the Perception-Action Cycle

http://homepages.feis.herts.ac.uk/~comqdp1/NIPS_2008/NIPS_Symposium_Workshop.html

Daniel Polani

UNIVERSITY OF HERTFORDSHIRE

Naftali Tishby

THE HEBREW UNIVERSITY

d.polani@herts.ac.uk

tishby@cs.huji.ac.il

Abstract

A significant emphasis in trying to achieve adaptation and learning in the perception-action cycle of agents lies in the development of suitable algorithms. While partly these algorithms result from mathematical constructions, in modern research much attention is given to methods that mimic biological processes.

However, mimicking the apparent features of what appears to be a biologically relevant mechanism makes it difficult to separate the essentials of adaptation and learning from accidents of evolution. This is a challenge both for the understanding of biological systems as well as for the design of artificial ones. Therefore, recent work is increasingly concentrating on identifying general principles rather than individual mechanisms for biologically relevant information processing.

One advantage is that a small selection of principles can give rise to a variety of — effectively equivalent — mechanisms. The ultimate goal is to attain a more transparent and unified view on the phenomena in question. Possible candidates for such principles governing the dynamics of the perception-action cycle include but are not limited to information theory, Bayesian models, energy-based concepts or principles emerging from neuroscience.

13.30-13.50	Introduction NAFTALI TISHBY, HEBREW UNIVERSITY AND DANIEL POLANI, UNIVERSITY OF HERTFORDSHIRE
13.50-14.40	Encodings the Location of Objects with a Scanning Sensorimotor System DAVID KLEINFELD, UC SAN DIEGO
14.40-14.50	Coffee
14.50-15.40	Dealing with Risk in the Perception-Action Cycle YAEL NIV, PRINCETON UNIVERSITY
15.40-16.30	Perception and Action in Robotics SEBASTIAN THRUN, STANFORD UNIVERSITY
16.30	End of Session

Introduction

Naftali Tishby, HEBREW UNIVERSITY

Daniel Polani, UNIVERSITY OF HERTFORDSHIRE

In view of the variety of models, the need for principled approaches to understand the perception-action loop is increasingly felt. In this mini-symposium as well as in the associated workshop, we intend to present

various approaches towards this goal. The introduction will outline some current issues in the larger context of this complex.

Encodings the Location of Objects with a Scanning Sensorimotor System

David Kleinfeld, UC SAN DIEGO

Sensory perception in natural environments involves the dual challenge to encode external stimuli and manage the influence of changes in body position that alter the sensory field. We examined the mechanisms used to integrate sensory signals elicited by both external stimuli and motor activity through the use of behavioral, electrophysiological, and computation tools in conjunction with the vibrissa system of rat. We show that the location of objects is encoded in an "region-of-interest centered", as opposed to "body-centered", coordinate system. The underlying circuit for this computation is consistent with gating by a shunt, a common motif in cortical circuitry.

Dealing with Risk in the Perception-Action Cycle

Yael Niv, PRINCETON UNIVERSITY

Risk (or outcome variance) is omnipresent in the natural environment, posing a challenge to animal decision-making as well as to artificial agents. Indeed, much empirical research in psychology, economics and ethology has shown that humans and animals are sensitive to risk in their decision-making, often preferring a small but certain outcome to a probabilistic outcome with a higher expected payoff. In light of this it may be surprising that optimal control methods such as reinforcement learning do not explicitly take risk into account. In this talk, I will review some of the literature on behavioral risk sensitivity in humans and in animals. I will then discuss a number of recent studies in which the neural basis of risk sensitivity has been investigated. While it is not surprising that the brain represents risk, the role of risk in decision making is still far from clear. I will argue that risk might play a more integrated role in learning and action selection than was previously postulated, specifically through the mechanism of reinforcement learning. However, this still leaves open the question of what we should learn from this neural solution: are there principled (normative) reasons for the prevalence of risk sensitivity? Does a general solution to optimal action selection have to take risk into account?

Perception and Action in Robotics

Sebastian Thrun, STANFORD UNIVERSITY

This overview talk investigates the perception action cycle from the robotics perspective. The speaker will discuss existing robotic implementation, discuss alternative architectures, and provide insights from the field of robotics. The speaker led the winning DARPA Grand Challenge team and heads Stanford's autonomous car research group. He will provide ample examples from real-time control of self-driving cars in complex traffic situations.

Beyond Search: Computational Intelligence for the Web

http://research.microsoft.com/osa/adCenter/beyond_search/

Anton Schwaighofer
MICROSOFT RESEARCH

antonsc@microsoft.com

Junfeng Pan
GOOGLE

panjf@google.com

Thomas Borchert
MICROSOFT RESEARCH

tborcher@microsoft.com

Olivier Chapelle
YAHOO! RESEARCH

chap@yahoo-inc.com

Joaquin Quiñonero Candela
MICROSOFT RESEARCH

joaquin@microsoft.com

Abstract

The WWW has reached the stage where it can be looked upon as a gigantic information copying and distribution mechanism. But when the problem of distributing and copying information is essentially solved, where do we go next? There are a number of values that can be derived from the mesh, that also have immediate relevancy for the ML community. The goal of the workshop is to link these areas, and encourage cross-boundary thinking and working. The topics will be:

1. Machine learning and probabilistic modeling: Recommendation systems and knowledge extraction are two immediate applications, with research required for large scale inference, modeling languages, and efficient decision making.
2. Game theory and mechanism design: When a large number of contributors is involved, how can tasks and incentive structures be made such that the desired goal is achieved? Research is required for solving very large games, and for mechanism design under uncertainty.
3. Knowledge representation and reasoning: Large parts of the web are currently stored in an unstructured way, making linking and evaluating knowledge a complex problem. Open points are the difficulty of reasoning, the trade-off between efficiency of reasoning and power of the representation, and reasoning under uncertainty.
4. Social networks and collective intelligence: How does information flow in the web? Who is reading what, who is in touch with whom? These networks need to be analyzed, modeled, and made amenable to reasoning.
5. Privacy preserving learning: What can be learned, and how can be learned, whilst only revealing a minimal set of information, or information that does not make users individually identifiable?

- Friday Dec 12** Morning: **Knowledge Representation and Reasoning**
- 7:30-7:40** **Introduction**
THE ORGANIZERS
- 7:40-8:40** **Invited talk: Evolving Today's Web into a Knowledge Web**
AC SURENDRAN, TAREK NAJM, PHANI VADDADI (MICROSOFT)
- 8:40-9:10** **Invited talk: Some Machine Learning Problems Related to Content Optimization**
DEEPAK AGARWAL (YAHOO! LABS)
- 9:10-9:30** Break
- 9:30-10:30** **Invited talk: The Semantic Web**
DOUG LENAT (CYCORP)
- 10:30-11:00** Discussion and wrap-up
- Friday Dec 12** Afternoon: **Social Networks**
- 16:00-17:00** **Invited talk: Scalable Collaborative Filtering Algorithms for Mining Social Networks**
EDWARD CHANG (GOOGLE)
- 17:00-17:20** Break
- 17:20-17:40** **Trust-Enhanced Peer-to-Peer Collaborative Web Search**
BARRY SMYTH, PETER BRIGGS (UC DUBLIN)
- 17:40-18:00** **Collective Wisdom: Information Growth in Wikis and Blogs**
SAMMAY DAS, MALIK MAGDON-ISMAIL (RENSSELAER POLYTECHNIC)
- 18:00-18:30** Discussion and wrap-up
- Saturday Dec 13** Morning: **Machine Learning**
- 7:30-8:30** **Invited talk: Online Search and Advertising, Future and Present**
CHRIS BURGESS (MICROSOFT RESEARCH)
- 8:30-8:50** **Interactively Optimizing Information Systems as a Dueling Bandits Problem**
YISONG YUE, THORSTEN JOACHIMS (CORNELL UNIVERSITY)
- 8:50-9:10** Discussion
- 9:10-9:30** Break
- 9:30-10:30** **Invited talk: Machine Learning for the Web: A Unified View**
PEDRO DOMINGOS (UNIVERSITY OF WASHINGTON)
- 10:30-10:50** **Search Query Disambiguation from Short Sessions**
LILYANA MIHALKOVA, RAYMOND MOONEY (UNIVERSITY OF TEXAS AUSTIN)
- 10:50-11:00** Discussion and wrap-up

Saturday Dec 13	Afternoon: Game Theory and Mechanism Design
16:00-16:30	Invited talk: Optimal Mechanism Design: from the 2007 Nobel Prize in Economics to the Foundations of Internet Algorithms JASON HARTLINE (NORTHWESTERN UNIVERSITY)
16:30-16:40	Break
16:40-17:40	Invited talk: Internet Advertising and Optimal Auction Design MICHAEL SCHWARZ (YAHOO!)
17:40-18:00	Learning Optimally from Self-interested Data Sources in On-line Ad Auctions ONNO ZOETER (XEROX)
18:00-18:30	Discussion and wrap-up

Abstracts

Some Machine Learning Problems related to Content Optimization

Deepak Agarwal, YAHOO! LABS

I will discuss a relatively new problem of selecting the best content to display for a user visit on a site like Yahoo!, MSN, Digg etc. In particular, I will consider scenarios where the content pool to select from is dynamic with short article lifetimes. I will provide an in-depth discussion of modeling challenges (and some of our solutions) that arise in this scenario, viz, a) estimating click-through rates that exhibit both temporal and positional variations b) efficient explore/exploit schemes and c) personalization of content to individual users. Finally, I will end with discussion of some open problems in this area. Throughout, data from a content module published regularly on the Yahoo! Front Page will be used for illustration.

Online Search and Advertising, Future and Present

Chris Burges, MICROSOFT

Search engine companies are gathering treasure troves of user-generated data. It has already been shown that such data can be used to directly improve the user's online experience. I will discuss some ideas as to what online search and advertising might look like a few years hence, in light of the algorithms and data we have now. Moving from future to present, I will outline some recent work done by researchers in the Text Mining, Search and Navigation team at Microsoft Research; the work in TMSN touches many aspects of online search and advertising.

Scalable Collaborative Filtering Algorithms for Mining Social Networks

Edward Chang, GOOGLE

Social networking sites such as Orkut, MySpace, Hi5, and Facebook attract billions of visits a day, surpassing the page views of Web Search. These social networking sites provide applications for individuals to establish communities, to upload and share documents/photos/videos, and to interact with other users. Take Orkut as an example. Orkut hosts millions of communities, with hundreds of communities created and tens of thousands of blogs/photos uploaded each hour. To assist users to find relevant information, it is essential to provide effective collaborative filtering tools to perform recommendations such as friend, community, and ads matching. In this talk, I will first describe both computational and storage challenges to traditional collaborative filtering algorithms brought by aforementioned information explosion. To deal with huge social graphs that expand continuously, an effective algorithm should be designed to 1) run on thousands of parallel machines for sharing storage and speeding up computation, 2) perform incremental retraining and updates for attaining online performance, and 3) fuse information of multiple sources for alleviating information sparseness. In the second part of the talk, I will present algorithms we recently developed including parallel Spectral Clustering, parallel PF-Growth, parallel combinational collaborative filtering, parallel LDA, parallel spectral clustering, and parallel Support Vector Machines.

Collective Wisdom: Information Growth in Wikis and Blogs

Sammay Das, RENSSELAER POLYTECHNIC

Malik Magdon-Ismail, RENSSELAER POLYTECHNIC

Wikis and blogs have become enormously successful media for collaborative information creation. Articles and posts accrue information through the asynchronous editing of users who arrive both seeking information and possibly able to contribute information. Most articles stabilize to high quality, trusted sources of information representing the collective wisdom of all the users who edited the article. We propose a model for information growth which relies on two main observations: (i) as an article's quality improves, it attracts visitors at a faster rate (a rich get richer phenomenon); and, simultaneously, (ii) the chances that a new visitor will improve the article drops (there is only so much that can be said about a particular topic). Our model is able to reproduce many features of the edit dynamics observed on Wikipedia and on blogs collected from LiveJournal; in particular, it captures the observed rise in the edit rate, followed by $(1/t)$ decay.

Machine Learning for the Web: A Unified View

Pedro Domingos, UNIVERSITY OF WASHINGTON

Machine learning and the Web are a technology and an application area made for each other. The Web provides machine learning with an ever-growing stream of challenging problems, and massive data to go with them: search ranking, hypertext classification, information extraction, collaborative filtering, link prediction, ad targeting, social network modeling, etc. Conversely, seemingly just about every conceivable machine learning technique has been applied to the Web. Can we make sense of this vast jungle of techniques and applications? Instead of attempting an (impossible) exhaustive survey, I will instead try to distill a unified view of the field from our experience to date. By using the language of Markov logic networks - which has most of the statistical models used on the Web as special cases - and the state-of-the-art learning and inference algorithms for it, we will be able to cover a lot of ground in a short time, understand the fundamental structure of the problems and solutions, and see how to combine them into larger systems.

Search Query Disambiguation from Short Sessions

Lilyana Mihalkova, U OF TEXAS AUSTIN

Raymond Mooney, U OF TEXAS AUSTIN

Web searches tend to be short and ambiguous. It is therefore not surprising that Web query disambiguation is an actively researched topic. However, most existing work relies on the existence of search engine log data in which each user's search activities are recorded over long periods of time. Such approaches may raise privacy concerns and may be difficult to implement for pragmatic reasons. In this work, we present an approach to Web query disambiguation that bases its predictions only on a short glimpse of user search activity, captured in a brief session of about 5-6 previous searches on average. Our method exploits the relations of the current search session in which the ambiguous query is issued to previous sessions in order to predict the user's intentions and is based on Markov logic. We present empirical results that demonstrate the effectiveness of our proposed approach on data collected from a commercial general-purpose search engine.

Internet Advertising and Optimal Auction Design

Michael Schwarz, YAHOO

This talk describes the optimal (revenue maximizing) auction for sponsored search advertising. We show that a search engine's optimal reserve price is independent of the number of bidders. Using simulations, we consider the changes that result from a search engine's choice of reserve price and from changes in the number of participating advertisers.

Trust-Enhanced Peer-to-Peer Collaborative Web Search

Barry Smyth, UC DUBLIN

Peter Briggs, UC DUBLIN

We spend a lot of our time online using web search services, but even the leading search engines frequently fail to deliver relevant results for the vague queries that are commonplace among today's web searchers. Interestingly, when we look at the search patterns of link-minded searchers (perhaps friends or colleagues) we do find considerable overlap between their queries and result-selections. This motivates a more collaborative approach to web search, one in which the past search experiences of friends and colleagues can be used to

usefully influence our new searches. In this talk we will describe how a novel combination of case-based reasoning, web search, and peer-to-peer networking can be used to develop a platform for personalized web search, which benefits from better quality results, improved robustness against search spam, while offering an increased level of privacy to the individual user.

Evolving Today's Web into a Knowledge Web

AC Surendran, MICROSOFT

Tarek Najm, MICROSOFT

Phani Vaddadi, MICROSOFT

Today's applications like search and social networks, although powerful, are limited in their power due to the shortcoming of their underlying data stores. These information stores are application tuned, silo-ed, and not configured to deeply understand the content. In this talk, we discuss some ideas on evolving the underlying store into a potent knowledge base. We envision how an evolved web might look like in the future - an intelligent knowledge store that integrates information with inference, and a platform that can unleash new killer applications with ease. We outline some practical challenges that have to be solved by the machine learning community to propel this evolution.

Interactively Optimizing Information Systems as a Dueling Bandits Problem

Yisong Yue, CORNELL

Thorsten Joachims, CORNELL

We present an online learning framework tailored towards real-time learning from observed user behavior in search engines and other information access systems. In particular, we only require pairwise comparisons which were shown to be reliably inferred from implicit feedback [4, 3]. We will present an algorithm with theoretical guarantees as well as simulation results.

Learning optimally from self-interested data sources in on-line ad auctions

Onno Zoeter, XEROX

At present all major search engines use an on-line auction to assign advertisements to available slots. A key element in all these auctions is the click-through rate of each advertisement [3, 4, 1]. These click-through rates are not readily observable and need to be learned from usage data. Interestingly enough this is not a standard machine learning problem: the data is not generated by genuinely random, but by self-interested sources. A higher click-through rate estimate is preferable for an advertiser as it leads to a higher position on the results page and a lower cost per click. This motivates a form of advertiser cheating that we refer to as reincarnating: by taking an ad out of the system and reintroducing it as fresh an advertiser can jump from a very low click-through rate estimate to the default starting value. This has a big impact on the system: it represents a loss of opportunity because other, better targeted ads could have been shown, but arguably more importantly it has a negative impact on the experience of the searcher, ads that have been identified as poorly targeted will still be shown instead of correctly filtered out.

DECEMBER 12, 2008, 07:30–10:40 AND 15:30–18:30

HILTON: CHEAKAMUS WS2a

Speech and Language: Learning-based Methods and Systems

<http://research.microsoft.com/~xiaohe/NIPS08/default.aspx>

Xiaodong He
MICROSOFT RESEARCH
Li Deng
MICROSOFT RESEARCH

xiaohe@microsoft.com

deng@microsoft.com

Abstract

This workshop is intended for researchers interested in machine learning methods for speech and language processing and in unifying approaches to several outstanding speech and language processing issues. In the last few years, significant progress has been made in both research and commercial applications of speech and language processing. Despite the superior empirical results, however, there remain important theoretical issues to be addressed. Theoretical advancement is expected to drive greater system performance improvement, which in turn generates the new need of in-depth studies of emerging novel learning and modeling methodologies. The main goal of the proposed workshop is to fill in the above need, with the main focus on the fundamental issues of new emerging approaches and empirical applications in speech and language processing. Another focus of this workshop is on the unification of learning approaches to speech and language processing problems. Many problems in speech processing and in language processing share a wide range of similarities (despite conspicuous differences), and techniques in speech and language processing fields can be successfully cross-fertilized. It is of great interest to study unifying modeling and learning approaches across these two fields. In summary, we hope that this workshop will present an opportunity for intensive discussions of emerging learning methods among speech processing, language processing, and machine learning researchers, and will inspire unifying approaches to problems across the speech and language processing fields.

- | | |
|--------------------|---|
| 7.30-8.10 | Invited Talk: New Multi-level Models for High-dimensional Sequential Data
GEOFFREY HINTON, UNIVERSITY OF TORONTO |
| 8.10-8.50 | Invited Talk: Log-linear Approach to Discriminative Training
RALF SCHLÜTER, RWTH AACHEN |
| 8.50-9.20 | Poster Session
POSTER AND COFFEE BREAK |
| 9.20-10.00 | Invited Talk: On the Role of Local Learning for Language Modeling
MARI OSTENDORF, UNIVERSITY OF WASHINGTON |
| 10.00-10.40 | Invited Talk: Ensemble Machine Learning Methods for Acoustic Modeling of Speech
YUNXIN ZHAO, UNIVERSITY OF MISSOURI, COLUMBIA |
| 10.40-15.30 | Lunch Break
LUNCH BREAK AND POSTER CONTINUES |
| 15.30-16.10 | Invited Talk: Relations Between Graph Triangulation, Stack Decoding, and Synchronous Decoding
JEFF BILMES, UNIVERSITY OF WASHINGTON |

- 16.10-16.50** **Invited Talk: Markov Logic Networks: A Unified Approach to Language Processing**
PEDRO DOMINGOS, UNIVERSITY OF WASHINGTON
- 16.50-17.05** **Coffee Break**
COFFEE BREAK
- 17.05-17.45** **Invited Talk: Some Machine Learning Issues in Discriminative Bilingual Word Alignment**
ROBERT MOORE, MICROSOFT RESEARCH
- 17.45-18.25** **Invited Talk: Machine Learning for Speaker Recognition**
ANDREAS STOLCKE, SRI INTERNATIONAL
- 18.25-18.30** **Conclusion & Close**
WORKSHOP CONCLUSION

New Multi-level Models for High-dimensional Sequential Data

Geoffrey Hinton, UNIVERSITY OF TORONTO

I will describe recent developments in learning algorithms for multilevel nonlinear generative models of sequential data. The models are learned greedily, one layer of features at a time and each additional layer of nonlinear features improves the overall generative model of the data. In earlier work (Taylor et. al. 2006) the basic module used for learning each layer of representation was a restricted Boltzmann machine in which both the hidden and visible units have biases that are dynamically determined by previous frames of data. This simple learning module has now been generalized to allow more complicated, multiplicative interactions so that hidden variables at one level can control the interactions between variables at the level below. These models have not yet been applied to speech but they work well on other data such as broadcast video and sequences of joint-angles derived from motion capture markers. (Joint work with Roland Memisevic, Graham Taylor and Ilya Sutskever).

Log-linear Approach to Discriminative Training

Ralf Schlüter, RWTH AACHEN

The objective of this talk is to establish a log-linear modeling framework in the context of discriminative training criteria, with examples from automatic speech recognition and concept tagging. The talk covers three major aspects. First, the acoustic models of conventional state-of-the-art speech recognition systems conventionally use generative Gaussian HMMs. In the past few years, discriminative models like for example Conditional Random Fields (CRFs) have been proposed to refine acoustic models. This talk addresses to what extent such less restricted models add flexibility to the model compared with the generative counterpart. Certain equivalence relations between Gaussian and log-linear HMMs are established, including context conditional models. Second, it will be shown how conventional discriminative training criteria in speech recognition such as the Minimum Phone Error criterion or the Maximum Mutual Information criterion can be extended to incorporate a margin term. As a result, large-margin training in speech recognition can be performed using the same efficient algorithms for accumulation and optimization and using the same software as for conventional discriminative training. We show that the proposed criteria are equivalent to Support Vector Machines with suitable smooth loss functions, approximating the non-smooth hinge loss function or the hard error (e.g. phone error). Third, CRFs are often estimated using an entropy based criterion in combination with Generalized Iterative Scaling (GIS). GIS offers, upon others, the immediate advantages that it is locally convergent, completely parameter free, and guarantees an improvement of the criterion in each step. Here, GIS is extended to allow for training log-linear models with hidden variables and optimization of discriminative training criteria different from Maximum Entropy/Maximum Mutual Information, including Minimum Phone Error (MPE). Finally, experimental results are provided for different tasks, including the European Parliament Plenary Sessions task as well as Mandarin Broadcasts.

On the Role of Local Learning for Language Modeling

Mari Ostendorf, UNIVERSITY OF WASHINGTON

Local learning methods, such as nearest-neighbor and variants, are known to be very powerful for many problems, particularly for problems where good models are not available. They can also be very useful for problems with a high degree of variability over the input space. In language modeling for speech recognition, local learning has not been particularly useful, in part because of the tremendous power of the n-gram when given large amounts of training data, and in part due to the difficulty of defining distance or similarity measures for word sequences. However, language is quite variable, depending on both topic and genre, such that a model trained in one domain may be of little use in another. With the large amount of data available on the web, and the large number of possible topic/genre combinations, it is of interest to consider local learning for language model adaptation. In this talk, we look at leveraging the similarity function in language model adaptation to benefit from a small neighborhood without losing the power of a large training corpus.

Ensemble Machine Learning Methods for Acoustic Modeling of Speech

Yunxin Zhao, UNIVERSITY OF MISSOURI

Improving recognition accuracy of human speech by computers has been a long standing challenge. Over the past few decades, tremendous research efforts have been made on the optimization of acoustic models. On the other hand, ensemble classifier design is becoming an important direction in machine learning. Different from the commonly adopted approach of optimizing a single classifier, ensemble methods achieve pattern discrimination through synergically combining many classifiers that are complementary in nature. Ensemble methods have shown advantages in classification accuracy and robustness in a variety of application contexts. Aligned with this direction, combining output word hypotheses from multiple speech recognition systems is being increasingly used in ASR for boosting the accuracy performance. Nonetheless, the complexity of speech sound distributions warrants the exploration of using ensemble methods to build robust and accurate acoustic models, where the component models of an ensemble can be combined in computing the acoustic scores during decoding search, for example, at the speech frame level, and thereby a single recognition system would suffice. Recently, some innovative progresses have been made in this direction, producing promising results and revealing attractive properties of ensemble acoustic models. This talk will address several basic issues in ensemble acoustic modeling, including constructing acoustic model ensembles, combining acoustic models in an ensemble, measuring the ensemble quality, etc. Experimental findings will be provided for a conversational speech recognition task, and a discussion will be made regarding research opportunities along this path.

Relations Between Graph Triangulation, Stack Decoding, and Synchronous Decoding

Jeff Bilmes, UNIVERSITY OF WASHINGTON

Speech recognition systems have historically utilized essentially one of two decoding strategies. Stack decoding (also called asynchronous decoding) allows internal decoding hypotheses to exist that have an end-time that spans over a potentially wide range of time frames. Such strategies are amenable to techniques such as A*-search assuming one has available a reasonable continuation heuristic. An alternate decoding strategy is the time-synchronous approach, whereby every active hypothesis has a similar or identical ending time. In this talk, we relate these two decoding strategies to inference procedures in dynamic graphical models (which includes Dynamic Bayesian networks and hidden conditional random fields). In particular, we see that under a hybrid search/belief-propagation inference scheme, the underlying triangulation of the graph determines which of the above two decoding strategies are active. The triangulation, moreover, also suggests decoding strategies that lie somewhere between strictly synchronous and asynchronous approaches.

Markov Logic Networks: A Unified Approach to Language Processing

Pedro Domingos, UNIVERSITY OF WASHINGTON

Language processing systems typically have a pipeline architecture, where errors accumulate as information progresses through the pipeline. The ideal solution is to perform fully joint learning and inference across all stages of the pipeline (part-of-speech tagging, parsing, coreference resolution, semantic role labeling, etc.) To make this possible without collapsing under the weight of complexity, we need a modeling language that provides a common representation for all the stages and makes it easy to combine them. Markov logic networks accomplish this by attaching weights to formulas in first-order logic and viewing them as templates

for features of Markov random fields. In this talk, I will describe some of the main inference and learning algorithms for Markov logic, show how Markov logic can be used to implement an end-to-end NLP system, and present the state-of-the-art results we have obtained with the components we have implemented so far.

Some Machine Learning Issues in Discriminative Bilingual Word Alignment

Robert Moore, MICROSOFT RESEARCH

Bilingual word alignment is the task of identifying the word tokens that are translations of each other in a corpus of sentence pairs that are translations of each other. After being dominated by generative models since the early 1990s, beginning in 2005 this task has been addressed by a number of discriminative approaches, resulting in substantially reduced alignment error rates. In most cases, these discriminative approaches have used a few hundred parallel sentence pairs with word alignments annotated, plus hundreds of thousands of parallel sentence pairs with no word-level annotation, making this task a prime example of semi-supervised learning. In this talk, we will look in detail at some of the machine learning issues in one of the most successful efforts at discriminative word alignment, including benefits of stacking of learners and refinements of the averaged perceptron approach to learning classifiers with structured outputs.

Machine Learning for Speaker Recognition

Andreas Stolcke, SRI INTERNATIONAL

This talk will review some of the main ML techniques employed in state-of-the-art speaker recognition systems, in terms of both modeling and feature design. For modeling, the two main paradigms currently in use are Gaussian mixture models with joint factor analysis, and support vector machines. The latter in particular have enabled a wealth of approaches that model speakers via high dimensional feature vectors drawn from a wide range of observation spaces, including cepstral, phonetic, prosodic, and lexical features. A pervasive problem in feature design is how to collapse a variable-length stream of observations into a fixed-length feature vector. SVM kernels designed for this situation are based on features generated by polynomial expansion, N-gram frequencies, and GMM mixture weights. Miscellaneous other issues include parameter smoothing (prior modeling) and model combination. It is hoped that the talk will give a glimpse into a fascinating application domain for machine learning methods, and instigate ML researchers to contribute to advances in speaker recognition.

Speech and Language: Unsupervised Latent-Variable Models

http://nlp.cs.berkeley.edu/Main.html#NIPS08_Overview

Slav Petrov

UNIVERSITY OF CALIFORNIA, BERKELEY

petrov@cs.berkeley.edu

Aria Haghighi

UNIVERSITY OF CALIFORNIA, BERKELEY

aria42@cs.berkeley.edu

Percy Liang

UNIVERSITY OF CALIFORNIA, BERKELEY

pliang@cs.berkeley.edu

Dan Klein

UNIVERSITY OF CALIFORNIA, BERKELEY

klein@cs.berkeley.edu

Abstract

Natural language processing (NLP) models must deal with the complex structure and ambiguity present in human languages. Because labeled data is unavailable for many domains, languages, and tasks, supervised learning approaches only partially address these challenges. In contrast, unlabeled data is cheap and plentiful, making unsupervised approaches appealing. Moreover, in recent years, we have seen exciting progress in unsupervised learning for many NLP tasks, including unsupervised word segmentation, part-of-speech and grammar induction, discourse analysis, coreference resolution, document summarization, and topic induction. The goal of this workshop is to bring together researchers from the unsupervised machine learning community and the natural language processing community to facilitate cross-fertilization of techniques, models, and applications. The workshop focus is on the unsupervised learning of latent representations for natural language and speech. In particular, we are interested in structured prediction models which are able to discover linguistically sophisticated patterns from raw data. To provide a common ground for comparison and discussion, we will provide a cleaned and preprocessed data set for the convenience of those who would like to participate. This data will contain part-of-speech tags and parse trees in addition to raw sentences. An exciting direction in unsupervised NLP is the use of parallel text in multiple languages to provide additional structure on unsupervised learning. To that end, we will provide a bilingual corpus with word alignments, and encourage the participants to push the state-of-the-art in unsupervised NLP.

- | | |
|--------------------|---|
| 7.30-7.35 | Welcome |
| 7.35-8.35 | Invited Talk: Unsupervised Learning with Side Information
ANDREW MCCALLUM |
| 8.35-9.30 | Poster Session & Coffee Break |
| 9.30-10.30 | Invited Talk: Climbing the Tower of Babel: Advances in Unsupervised Multilingual Learning
REGINA BARZILAY |
| 15.30-16.30 | Invited Talk: A Bayesian Approach to Learning Linguistic Structure
SHARON GOLDWATER |
| 16.30-17.30 | Poster Session & Coffee Break |
| 17.30-18.30 | Panel Discussion
REGINA BARZILAY, SHARON GOLDWATER, ANDREW MCCALLUM, HAL DAUME III |

Invited Talk: Unsupervised Learning with Side Information**Andrew McCallum**, UNIVERSITY OF MASSACHUSETTS, AMHERST

Even though we use the term “unsupervised learning”, we often have a particular goal or use-case in mind beyond simply reproducing the input data. I will review several approaches that leverage additional side information to guide learning without the use of traditional labeling, including recent work at UMass on Generalized Expectation (GE) criteria.

Invited Talk: A Bayesian Approach to Learning Linguistic Structure**Sharon Goldwater**, UNIVERSITY OF EDINBURGH

Natural language is characterized by highly skewed (often Zipfian) distributions over both observed and latent variables. This leads to ubiquitous problems with sparse data at the tails of the distribution, and often means that generalizations are heavily influenced by a few very frequent forms that are not representative of the range of forms to be expected. In this talk, I show how three ideas from Bayesian statistics can help with these problems: 1) *hierarchical models* allow sharing of information between similar cases, providing principled back-off methods, 2) *integrating over parameters* allows the use of priors favoring skewed distributions, and 3) *adaptor processes* such as the Chinese restaurant process and Pitman-Yor process provide a way to model the Zipfian distribution on frequencies, allowing more subtle (but more interesting) linguistic structure to be identified. I discuss examples from part-of-speech tagging, morphology, and word segmentation.

Invited Talk: Climbing the Tower of Babel: Advances in Unsupervised Multilingual Learning**Regina Barzilay**, MIT

For most natural language processing tasks, unsupervised methods significantly underperform their supervised counterparts. In this talk, I will demonstrate that multilingual learning can narrow this gap. The key insight is that joint learning from several languages reduces uncertainty about the linguistic structure of individual languages. These methods exploit the deep structural connections between languages, connections that have driven many important discoveries in anthropology and historical linguistics.

I will present multilingual unsupervised models for morphological segmentation and part-of-speech tagging. Multilingual data is modeled as arising through a combination of language-independent and language-specific probabilistic processes. This approach allows the model to identify and learn from recurring cross-lingual patterns, ultimately to improve prediction accuracy in each language. I will also discuss ongoing work on unsupervised decoding of ancient Ugaritic tablets using data from related Semitic languages. This is joint work with Benjamin Snyder, Tahira Naseem and Jacob Eisenstein.

Learning Latent-Variable Models for Mapping Sentences to Logical Form**Luke Zettlemoyer**, MIT**Michael Collins**, MIT**The Shared Logistic Normal Distribution for Grammar Induction****Shay Cohen**, CMU**Noah Smith**, CMU**Hidden Topic Models for Hierarchical Segmentation****Jacob Eisenstein**, UIUC**Sparse Topic Models****Chong Wang**, PRINCETON**David Blei**, PRINCETON**Multilingual Topic Models****Jordan Boyd-Graber**, PRINCETON

David Blei, PRINCETON

A Hierarchical Dirichlet Process Prior for a Conditional Model of Phrase Alignment

John DeNero, UC BERKELEY

Alexandre Bouchard-Cote, UC BERKELEY

Statistical Analysis and Modeling of Response Dependencies in Neural Populations

<http://ni.cs.tu-berlin.de/projects/nips2008/>

Klaus Obermayer

BERNSTEIN CENTER FOR COMPUTATIONAL NEUROSCIENCE AND TECHNISCHE UNIVERSITÄT BERLIN

Valentin Dragoi

UNIVERSITY OF TEXAS-HOUSTON MEDICAL SCHOOL

Arno Onken

BERNSTEIN CENTER FOR COMPUTATIONAL NEUROSCIENCE AND TECHNISCHE UNIVERSITÄT BERLIN

Steffen Grünewälder

TECHNISCHE UNIVERSITÄT BERLIN

Denise Berger

FREIE UNIVERSITÄT BERLIN

oby@cs.tu-berlin.de

v.dragoi@uth.tmc.edu

aonken@cs.tu-berlin.de

gruenew@cs.tu-berlin.de

d.berger@gmx.net

Abstract

It is well known that sensory and motor information is represented in the activity of large populations of neurons. Encoding and decoding this information is a matter of active debate. One way to study population coding is to analyze response dependencies. Rate and temporal coding are opposing theories of neural coding. Dependency concepts for these theories are rate covariance and temporal spike coordination. In the typical theoretical framework, response dependencies are characterized by correlation coefficients and cross-correlograms. The main goal of this workshop is to challenge the dependency concepts that are typically applied and to disseminate more sophisticated concepts to a wider public. It will bring together experts from different fields and encourage exchange of insights between experimentalists and theoreticians.

Day 1

07:30-07:45	Introduction
07:45-08:15	Beyond correlations: modeling neural dependencies with copulas PIETRO BERKES
08:15-08:45	Copula-based point process models of neural dependence RICK JENISON
08:45-09:00	Coffee break
09:00-09:30	Dual coding by spiking neural networks NAOKI MASUDA
09:30-10:00	Low-dimensional single-trial analysis of neural population activity BYRON YU
10:00-10:30	Discussion
10:30-15:30	Break
15:30-16:15	Inter-neuronal connectivity and correlation in a model of V1 simple cells WYETH BAIR

- 16:15-16:45** **The flashlight transformation for mixture copula based modeling of spike-counts**
ARNO ONKEN
- 16:45-17:00** Coffee break
- 17:00-17:40** **Adaptive coding in visual cortical networks**
VALENTIN DRAGOI
- 17:40-18:10** **Optimal inference from population of neurons in macaque primary visual cortex**
ARNULF GRAF
- 18:10-18:30** Open discussion

Day 2

- 07:30-08:00** **Identifying assemblies in massively parallel spike trains by higher-order synchrony**
SONJA GRÜN
- 08:00-08:30** **Spatially organized higher-order spike synchrony in cat area 17**
DENISE BERGER
- 08:30-08:45** Coffee break
- 08:45-09:30** **Relating response variability across stages of cortical processing**
ADAM KOHN
- 09:30-10:00** **How pairwise correlations shape the statistical structure of population activity**
JAKOB MACKE
- 10:00-10:30** Discussion
- 10:30-15:30** Break
- 15:30-16:15** **Modelling dependent count data**
DIMITRIS KARLIS
- 16:15-16:30** Coffee break
- 16:30-17:00** **Capacity of a single spiking neuron for temporal and rate coding**
SHIRO IKEDA
- 17:00-17:30** **State-space analysis on time-varying higher-order spike correlations**
HIDEAKI SHIMAZAKI
- 17:30-18:30** Open discussion

Beyond correlations: modeling neural dependencies with copulas

Pietro Berkes, BRANDEIS UNIVERSITY

An important open problem in systems neuroscience is to develop models that accurately describe the joint activity of multiple neurons. The specific challenge we address is to formulate a parametric model with the flexibility to capture both the discrete, non-negative distribution of single-neuron firing rates, and the strong

inter-neuron dependencies that arise from shared input and network interactions.

Copula models are statistical objects that combine a set of marginal distributions into a joint distribution with arbitrary dependency structure. This makes them an ideal tool for modeling neural data. Copulas are insensitive to nonlinear transformations of the individual variables, allowing one to define strong dependency measures that generalize correlation coefficients. Moreover, one can use different parametric families of copulas to model different dependency structures.

I will review the basic theoretical results behind copula models, derive a new Maximum Likelihood estimation method for parametric copula families with discrete marginals, and show results for pairs of neurons in pre-motor cortex. I will discuss the issues that still need to be solved to scale copula models to larger groups of neurons, and sketch the most promising future directions.

Copula-based point process models of neural dependence

Rick Jenison, UNIVERSITY OF WISCONSIN-MADISON

One of the most important questions in computational neuroscience is what role coordinated activity of ensemble neurons plays in the neural coding of sensory information. One approach to understanding this role is to formally model the ensemble responses as multivariate probability distributions. We have previously introduced alternatives to linear assumptions of gaussian dependence for spike timing in neural ensembles using the probabilistic copula approach, as well as demonstrating its utility in computing multi-information. In probability theory the copula “couples” marginal distributions to form flexible multivariate distribution functions for characterizing ensemble behavior. I will provide a basic introduction to Sklar’s theorem in my talk, which motivated our use of copulas for modeling ensemble spike timing. In principle, a copula construction is only appropriate for continuous random variables and not for stochastic processes such as point processes. However, latent diffusion processes underlying observed dependent point processes are amenable to copula constructions.

The first-passage time of a diffusion process through a constant or variable boundary has been the focus of many stochastic models of neuronal membrane potential dynamics. Diffusion processes have been used extensively to model a latent process that may only be observable through consequent point process events. The mathematical relationship between inter-spike intervals and the first-passage time of simple diffusion models is well-known, however this relationship becomes increasingly more complex as the diffusion models become more physiologically realistic, and when multivariate diffusion processes are no longer considered to be independent. The probability density of a diffusing particle position at a particular point in time $p(x,t)$ as defined by the Fokker-Planck equation can be solved, under suitable conditions, using the method of images. I’ll show how the method of images can be extended to a multivariate probability density constructed from marginal densities modeling simple individual spiking neurons using a copula construction that factors out the correlated (dependent) noise structure. This in turn provides a method for estimating multivariate spike survival and conditional intensity (hazard) functions from simultaneously recorded single unit activity and, indirectly, the ensemble diffusion noise. The utility of the approach in estimating neural response dependencies will be demonstrated with simulations as well as simultaneously recorded single-unit activity from Heschl’s gyrus (primary auditory cortex) in patients with pharmacologically intractable epilepsy.

Dual coding by spiking neural networks

Naoki Masuda, UNIVERSITY OF TOKYO

In addition to firing rates, more intricate features of spikes, such as synchrony, relative spike timing, and reproducible spatiotemporal spike patterns, have been suggested to carry neural information. They may represent external inputs, cognitive states, motor signal, and so on, cooperatively with or independently of firing rates. The code used probably depends on neural systems. However, how rates and such spike patterns comodulate and which aspects of inputs are effectively encoded, particularly in the presence of dynamical inputs, are elusive. Here I discuss the possibility of dual or multiple codes, in which different coding strategies are used simultaneously or separately in one neural system. I focus on spatially homogeneous networks of spiking neurons and show how input information can be coded onto firing rates or synchrony patterns in different conditions. Extending the results beyond mere synchrony and modeling experimental data are warranted for future work.

Low-dimensional single-trial analysis of neural population activity

Byron Yu, STANFORD UNIVERSITY

John Cunningham, STANFORD UNIVERSITY

Mark Churchland, STANFORD UNIVERSITY

Gopal Santhanam, STANFORD UNIVERSITY

Stephen Ryu, STANFORD UNIVERSITY

Krishna Shenoy, STANFORD UNIVERSITY

Maneesh Sahani, GATSBY COMPUTATIONAL NEUROSCIENCE UNIT, UCL

The response of a neuron is traditionally characterized by a peri-stimulus time histogram (PSTH), constructed by averaging the spike trains observed across repeated experimental trials. A fundamental assumption when constructing a PSTH is that the timecourse of the neuron's response is identical from one trial to the next. However, there is mounting evidence (spanning visual and motor areas) that a neuron's response can evolve quite differently on repeated trials, even in well-controlled experimental paradigms. In such settings, it is critical that the neural data not be averaged across trials, but instead be analyzed on a trial-by-trial basis. By leveraging multi-electrode recordings (comprising tens to hundreds of simultaneously-recorded neurons), we consider methods for extracting a smooth, low-dimensional "neural trajectory" summarizing the high-dimensional recorded activity on a single trial. Beyond basic data visualization, such trajectories can offer insight into the dynamics of the neural circuitry underlying the recorded activity. We applied these methods to neural activity recorded in macaque premotor and motor cortices during reach planning and execution. This yielded the first direct view of single-trial trajectories converging during reach planning, suggestive of attractor dynamics. We also show how such methods can be a powerful tool for relating the spiking activity across a neural population to the subject's behavior on a single-trial basis.

Inter-neuronal connectivity and correlation in a model of V1 simple cells

Wyeth Bair, UNIVERSITY OF OXFORD

Peter Keating, UNIVERSITY OF OXFORD

Inter-neuronal correlation is an important feature of cortical activity, providing an additional source of constraint for models of neuronal circuitry. Employing a modelling approach, we investigated the potential implications of inter-neuronal correlation, both within V1 layer 4C and between V1 layer 4C and the LGN, for models of thalamocortical (TC) and corticocortical (CC) connectivity. Our modelling framework consists of four populations of conductance-based integrate-and-fire units: ON and OFF cells in the LGN and excitatory and inhibitory simple cells in cortex. We explored a variety of TC connectivity schemes, and tested two previously proposed schemes for CC connectivity, one with anti-phase inhibition and one with isotropic inhibition. Using the model, we examined CC correlation for pairs of nearby cortical cells that had similar receptive fields (RFs) and TC correlation for pairs in which the LGN cell provided synaptic input to the cortical cell. Dense and sparse TC connectivity regimes were compared. In the dense regime, cortical cells sampled a large fraction of the LGN cells that matched their RFs, whereas in the sparse regime, they sampled only a small fraction of plausible LGN inputs. Dense connectivity yielded CC correlation that was consistent with experimental data, but the resulting TC correlation was weaker than expected. Sparse TC connectivity yielded stronger TC correlation, but produced CC correlation that was too weak. To achieve realistic TC and CC correlation strength simultaneously, it seems necessary to invoke either common CC input or correlation among LGN spike trains. Investigating the dependency of CC correlation strength on stimulus orientation, we found that our anti-phase and isotropic inhibition models show opposite behaviour, with the latter showing stronger correlations between a pair of similarly-tuned cells when the stimulus orientation matches the cells' preference and the former showing stronger correlations when the stimulus orientation deviated from preferred. Thus inter-neuronal correlation may importantly constrain functional circuitry. Supported by The Wellcome Trust.

The flashlight transformation for mixture copula based modeling of spike-counts

Arno Onken, BERNSTEIN CENTER FOR COMPUTATIONAL NEUROSCIENCE AND TECHNISCHE UNIVERSITÄT BERLIN

Steffen Grünewälder, TECHNISCHE UNIVERSITÄT BERLIN

Klaus Obermayer, BERNSTEIN CENTER FOR COMPUTATIONAL NEUROSCIENCE AND TECHNISCHE UNI-

VERSITÄT BERLIN

Copula based models of neural spike counts provide a way to model a rich set of dependence structures together with appropriate distributions for single neuron variability. In this talk we discuss three topics:

Firstly, results of applying the Farlie-Gumbel-Morgenstern (FGM) copula family are presented. This family has separate parameters for all higher order correlations. We used it to tackle the long-standing problem of analyzing the impact of higher order interactions.

Secondly, we present a novel copula transformation with interpretations for the underlying neural connectivity. The so-called flashlight transformation is a generalization of the copula survival transformation and makes it possible to move the tail dependence of a copula into arbitrary corners of the distribution. We discuss several interpretations with respect to inhibitory and excitatory connections of projecting populations and demonstrate their validity on integrate and fire population models. A mixture approach enables us to combine the advantages of this transformation with the FGM family. Inference can be performed by subsequent expectation maximization. The method is applied to data from macaque prefrontal cortex.

Finally, we talk about problems associated with the current approach: 1) The computational complexity restricts the number of neurons that can be analyzed. 2) Typically, not many samples are available for model inference - hence overfitting is an issue. We will discuss a potential solution to both problems: The approximation of copula based distributions by exponential families. Exponential families allow efficient inference in terms of computation time and sample size, whereas copula based models are well suited for describing and interpreting distributions of spike counts.

This work was supported by BMBF grant 01GQ0410.

Adaptive coding in visual cortical networks

Valentin Dragoi, UNIVERSITY OF TEXAS

It is increasingly being realized that the neural code is adaptive, that is sensory neurons change their responses and selectivity dynamically to match the changes in the statistics of the input stimuli. Understanding how rapid adaptation changes information processing by cortical networks is essential for understanding the relationship between sensory coding and behavior. Whether and how adaptation impacts information coding in neural populations is unknown. We examined how brief adaptation (on the time scale of visual fixation) influences the structure of neuronal correlations and the accuracy of population coding in macaque primary visual cortex (V1). We found that brief adaptation to a stimulus of fixed structure reorganizes the distribution of correlations across the entire network and improves the efficiency of the population code to optimize neuronal performance during natural viewing. We further examined whether and how rapid adaptation influences network processing in mid-level visual cortex, such as area V4. We investigated the degree of synchronization between the responses of individual neurons and the local populations by measuring the changes in spike-LFP coherence after adaptation. We found that the spike-triggered average LFP power was increased after adaptation in the gamma frequency band to indicate that rapid adaptation influences the communication between neuronal ensembles in V4. Altogether, these results suggest that adaptation improves both the accuracy of the population code and information transmission in early and mid-level visual cortical networks to facilitate visual perception.

Optimal inference from population of neurons in macaque primary visual cortex

Arnulf Graf, NEW YORK UNIVERSITY

Inferring the state of the world from the responses of sensory neurons is a central problem in neural computation. We studied the activity of groups of 40-70 neurons recorded simultaneously from the superficial layers of primary visual cortex of anesthetized, paralyzed macaque monkeys. We presented optimized high-contrast drifting sinusoidal gratings of 36 orientations. To explore the best method for extracting information from this activity, we estimated the orientation of the gratings with four different decoding strategies. A population vector method, in which each neurons “votes” for its preferred orientation in proportion to its actual response, yielded relatively inaccurate estimates. We obtained somewhat better performance by optimizing the weights in the population vector calculation to minimize the squared error of the estimates. These “direct” methods estimate orientation without creating an intermediate representation, but the theory of Bayesian statistics suggests that performance could be improved by introducing an intermediate representation of sensory activity to encode the likelihood that each orientation gave rise to the pattern of activity

observed. We therefore studied the performance of decoders that explicitly represent this likelihood function. The simple form of this decoder assumes that neuronal spike counts are statistically independent and Poisson distributed. This decoder, like the population vector, can be built directly from measured tuning properties, and yielded more accurate estimates than the previous decoders; taking advantage of the probabilistic nature of the neural response can therefore improve performance. Finally, we optimized this decoder by empirically deriving its parameters from the neuronal data using statistical learning theory. This empirical decoder does not assume either Poisson variability or independence, and uses the structure of the data to increase accuracy. The performance of this empirical decoder was the best of the four, and the main reason for its superiority lay in its ability to adjust parameters to take advantage of interneuronal correlations. Our results suggest that inference from neuronal populations is best achieved by using an intermediate representation of stimulus likelihood that is empirically constructed from a weighted sum of neuronal responses. We speculate that the computation of likelihood functions to approximate Bayesian optimal inference is an important function of the hierarchical cascade of connections among the visual areas of the extrastriate cortex.

Grants: The Swartz Foundation: EY07158, EY02017, EY04440.

Identifying assemblies in massively parallel spike trains by higher-order synchrony

Sonja Grün, RIKEN BRAIN SCIENCE INSTITUTE

The cell assembly hypothesis (Hebb, 1949) postulates interacting groups of neurons as building blocks of cortical information processing. Synchronized spiking across neuronal groups was later suggested as a potential signature for active assemblies (Abeles, 1991). Due to the rapid progress in recording technology the massively parallel data required to search for such signatures are now becoming available (e.g. Csicsvari et al, 2003; Euston et al, 2007). Although mere pairwise analysis may give indications of groups of intercorrelated neurons (Berger et al, 2007), it may not conclude on the existence of higher-order synchrony (HOS). Existing tools are severely limited by the combinatorial explosion in the number of spike patterns to be considered (e.g. Martignon et al, 1995; Nakahara & Amari, 2002; Grün et al, 2002). Therefore, population measures need to be constructed reducing the number of tests, potentially for the price of being able to answer only a restricted set of questions (e.g. Schrader et al, 2008).

We are currently following different approaches to tackle this problem. One is based on the investigation of the population histogram, i.e. composed of the sums of spike activities across neurons. The statistical features of the resulting amplitude distribution of this histogram ('complexity distribution') expresses correlations between neuronal activities. Independent of neuron identity it describes the probability to observe a particular number of synchronous spikes. On the basis of stochastic models (Kuhn et al, 2003; Grün et al, submitted) we illustrate that in the presence of higher-order synchrony the complexity distribution exhibits characteristic deviations from expectation (Grün et al, 2008). We devised a statistical test that identifies the presence of HOS and their order based on cumulants estimated from the complexity distribution (Stauder et al, to be submitted).

Another approach identifies higher-order synchrony based on the accretion method (Gerstein et al, 1978). The basic idea of accretion is to analyze pairs of spike trains for significant correlation which then are reduced to new point processes containing only synchronized spikes. These processes are in turn correlated with single neuron spike trains and so on, until the maximal order of correlation is found. In order to reduce the complexity of the search we make use of data mining approaches, in particular frequent itemset mining. As a result we get strings of neuron ids expressing higher-order synchrony between these neurons. By use of graph theoretical tools redundancy of the strings is eliminated to finally extract assemblies composed of neurons expressing higher-order synchrony.

Spatially organized higher-order spike synchrony in cat area 17

Denise Berger, FREIE UNIVERSITÄT BERLIN AND BCCN BERLIN

Christian Borgelt, EUROPEAN CENTER FOR SOFT COMPUTING

Markus Diesmann, RIKEN BRAIN SCIENCE INSTITUTE

George Gerstein, UNIVERSITY OF PENNSILVANIA

Sonja Grün, RIKEN BRAIN SCIENCE INSTITUTE

It is still an open question, how the concept of maps is related to temporal coding. Therefore we analyzed parallel spike recordings from cat visual cortex (10 x 10 grid, 3.6 x 3.6 mm) for spike correlation. Application

of pairwise cross-correlation analysis onto all (100) recorded multi-unit activities (MUA) allowed us to extract all significantly correlated MUA pairs. Graph theoretical analysis revealed a decomposition into a small number (4) of distinct clusters of inter-correlated MUAs, which also correspond to segregated clusters in cortical space. Their spatial scale is in agreement with the scale of orientation tuning maps. However, due to the limitation of the applied pairwise analysis, cell assemblies composed of larger groups of neurons could not be conclusively identified. Therefore, a test which extracts higher-order synchrony (HOS) needs to be used.

Therefore, we developed a new method for the detection of HOS, which combines the accretion method with frequent itemset mining (FIM). Spike synchrony is detected by the accretion approach: pairs of spike trains are tested for significant correlation and then reduced to new point processes containing only synchronized events. These processes are in turn correlated with further, single neuron spike trains and so on, until the maximal order of correlation is found. Ideas from FIM algorithms help to search the space of all neuron subsets efficiently. However, such algorithms usually rely on a minimum support criterion to prune the search. However, HOS does not necessarily imply frequent occurrence of spike patterns, and we are rather interested to extract spike patterns that occur significantly more often than expected given by the firing rates. Therefore we designed a FIM related algorithm (AFIM), that processes large sets of data efficiently. Using this approach we re-analyzed the above mentioned data. To account for non-stationarity in time we segmented the data into quasi-stationary segments, and analyzed these separately. In different time segments different sets of MUAs exhibit higher-order spike synchrony in groups of up to order 7. Within a time segment, the groups exhibiting HOS are highly overlapping, and are therefore combined into supersets. Most interestingly, the supersets of the different time segments do not overlap, but reveal the same, separate clusters of MUAs that were identified by the pairwise analysis. We verified our results by shuffling the trials of the individual MUAs, which did not reveal any HOS in any of the time segments. Thus, our results show strong evidence, that similar to the dynamic occurrence of activity patterns found in optical imaging, spatially segregated groups of higher-order correlated neurons are alternatingly active. Partially funded by BCCN Berlin (01GQ0413) and Helmholtz Alliance Systems Biology.

Relating response variability across stages of cortical processing

Adam Kohn, ALBERT EINSTEIN COLLEGE OF MEDICINE

The responses of cortical neurons to repeated presentations of a stimulus are variable and this variability is correlated between cells. Theoretical work has shown that correlated variability can strongly affect the ability of populations of neurons to encode sensory information, with the impact depending on both the structure of correlation and the algorithm used to decode the population response. Ultimately, however, the impact of correlated variability in a cortical area lies in its relationship to and impact on responses in downstream networks. To relate correlated variability at multiple stages of the visual system, we recorded simultaneously from a population of neurons in macaque primary visual cortex (V1, using implanted arrays of 100 microelectrodes) and their downstream targets in the input layers of area V2. We measured responses to repeated presentations of drifting gratings and evaluated how well we could predict trial-to-trial fluctuations in V2 responsiveness by monitoring population activity in V1. We fit a generalized linear model (GLM) to a subset of the trials, and tested its ability to predict responses on a novel subset of the data. We found we could predict a substantial portion of trial-to-trial fluctuations in V2 by monitoring V1 responses, with a performance level that was similar to our ability to predict a V1 neuron's response by monitoring its nearby neighbors; that is, it was as if V1 and V2 neurons were embedded in a single network. The V1 neurons weighted most heavily in the model were those whose spatial receptive fields were most similar to that of the V2 cell; relative orientation preference of the V1 and V2 cells played relatively little role. Our results suggest that a substantial portion of V2 variability is related to, and potentially inherited from, fluctuations in the response of populations of V1 neurons.

How pairwise correlations shape the statistical structure of population activity

Jakob Macke, MPI FOR BIOLOGICAL CYBERNETICS

Manfred Opper, TECHNISCHE UNIVERSITÄT BERLIN

Matthias Bethge, MPI FOR BIOLOGICAL CYBERNETICS

Simultaneously recorded neurons often exhibit correlations in their spiking activity. These correlations shape

the statistical structure of the population activity, and can lead to substantial redundancy across neurons. Knowing the amount of redundancy in neural responses is critical for our understanding of the neural code. Here, we study the effect of pairwise correlations on global population statistics, such as the redundancy. We model correlated activity as arising from common Gaussian inputs into simple threshold neurons. This model is equivalent to the Dichotomized Gaussian distribution, a flexible statistical model for correlated binary random variables. In population models with exchangeable correlation structure, one can analytically calculate the distribution of synchronous events across the whole population, and the joint entropy (and thus the redundancy) of the neural responses. We investigate the scaling of the redundancy as the population size is increased, and characterize its phase transitions for increasing correlation strengths. We compare the asymptotic redundancy in our models to the corresponding maximum- and minimum entropy models. Although this model must exhibit more redundancy than the maximum entropy model, we do not find a dramatic increase in redundancy when increasing the population size.

Modelling dependent count data

Dimitris Karlis, ATHENS UNIVERSITY

While methods for modelling dependent continuous data are abundant in the statistical literature, this is not the case when treating dependent count data. The talk aims at exploiting different ideas for modelling such data. There will be a discussion on how one can incorporate dependence on the data. The case of using mixtures will be examined as well as models based on sharing a common parameter to induce dependence. Traditional ideas like the GEE will be also mentioned.

Then we will consider the case of time series for count data. The two broad categories of observations and parameters driven models will be discussed. The similarities with the corresponding continuous time series models will be considered.

Finally we will discuss a recent idea that uses copulas to create multivariate discrete distributions. While copulas are very fashionable for continuous data the literature for discrete data is sparse for certain reasons. This will be emphasized.

A comparison of the pros and cons of different approaches will be also part of the talk.

State-space analysis on time-varying higher-order spike correlations

Hideaki Shimazaki, RIKEN BRAIN SCIENCE INSTITUTE

Shun-ichi Amari, RIKEN BRAIN SCIENCE INSTITUTE

Emery Brown, MASSACHUSETTS GENERAL HOSPITAL

Sonja Grün, RIKEN BRAIN SCIENCE INSTITUTE

Precise spike coordination in the spiking activities of a neuronal population is discussed as an indication of coordinated network activity in form of a cell assembly relevant for information processing. Supportive evidence for its relevance in behavior was provided by the existence of excess spike synchrony occurring dynamically in relation to behavioral context [e.g. Riehle et. al., *Science* (278) 1950-1953, 1997]. This finding was based on the null-hypothesis of full independence. However, one can assume that neurons jointly involved in assemblies express higher-order correlation (HOC) between their activities. Previous work on HOC assumed stationary condition. Here we aim at analyzing simultaneous spike trains for time-dependent HOCs to trace active assemblies.

We suggest to estimate the dynamics of HOCs by means of a state-space analysis with a log-linear observation model. A log-linear representation of the parallel spikes provides a well-defined measure of HOC based on information geometry (Amari, *IEEE Trans. Inf. Theory* (47) 1701-1711, 2001). In order to do that, we developed a nonlinear recursive filtering algorithm by applying a log-quadratic approximation to the filter distribution. Together with a fixed-interval smoothing algorithm, smoothed estimates of the time-dependent log-linear parameters are obtained. The time-scales of each parameter and their covariation are automatically optimized via the EM-algorithm under the maximum likelihood principle.

To obtain the most predictive model, we compare the goodness-of-fit of hierarchical log-linear models with different order of interactions using the Akaike information criterion (AIC; Akaike, *IEEE Trans. Autom. Control* (19) 716-723, 1974). While the inclusion of increasingly higher-order interaction terms into the log-linear state-space model improves the model accuracy, the estimation of higher-order parameters may suffer from large variances due to the paucity of synchronous spikes in the data. This bias-variance trade-off

is optimally resolved with the model that minimizes the AIC. The complexity of the model is thus selected based on the sample size of the data and the prominence of the higher-order structure.

Application of the proposed method to simultaneous recordings of neuronal activity is expected to provide us with new insights into the dynamics of assembly activities, their composition, and behavioral relevance.

DECEMBER 12, 2008, 07:30–10:30 AND 15:30–18:30

WESTIN: CALLAGHAN WS4

Algebraic and combinatorial methods in machine learning

<http://www.gatsby.ucl.ac.uk/~risi/AML08/>**Risi Kondor**

GATSBY UNIT, UCL

risi@gatsby.ucl.ac.uk

Guy Lebanon

GEORGIA INSTITUTE OF TECHNOLOGY

lebanon@cc.gatech.edu

Jason Morton

STANFORD UNIVERSITY

jason@math.stanford.edu

Abstract

There has recently been a surge of interest in algebraic methods in machine learning. In no particular order, this includes: new approaches to ranking problems; the budding field of algebraic statistics; and various applications of non-commutative Fourier transforms. The aim of the workshop is to bring together these distinct communities, explore connections, and showcase algebraic methods to the machine learning community at large. AML'08 is intended to be accessible to researchers with no prior exposure to abstract algebra. The program includes three short tutorials that will cover the basic concepts necessary for understanding cutting edge research in the field.

- | | |
|--------------------|--|
| 7.30-8.00 | Algebraic statistics for random graph models: Markov bases and their uses
STEPHEN E. FIENBERG |
| 8.05-8.35 | Algebraic statistics and contingency tables
ADRIAN DOBRA |
| 8.40-9.10 | Toric Modification on Mixture Models
KEISUKE YAMAZAKI |
| 9.15-9.25 | Coffee break |
| 9.25-9.55 | Learning Parameters in Discrete Naive Bayes Models by Computing Fibers of the Parametrization map
VINCENT AUVRAY |
| 10.00-10.30 | Stationary Subspace Analysis
PAUL VON BÜNAU |
| 15.30-16.00 | Alternatives to the Discrete Fourier Transform
DORU BALCAN |
| 16.05-16.35 | Graph Helmholtzian and rank learning
LEK-HENG LIM |
| 16.40-17.10 | Identity Management On Homogeneous spaces
XIAOYE JIANG |
| 17.15-17.25 | Coffee break |
| 17.25-17.55 | Exploiting Probabilistic Independence for Permutations
CARLOS GUESTRIN |

17.55-18.25 Consistent structured estimation for weighted bipartite matching
TIBERIO CAETANO

Algebraic statistics for random graph models: Markov bases and their uses

Stephen E. Fienberg, CARNEGIE MELLON UNIVERSITY

Sonja Petrović, UNIVERSITY OF ILLINOIS

Alessandro Rinaldo, CARNEGIE MELLON UNIVERSITY

We use algebraic geometry to study a statistical model for the analysis of networks represented by graphs with directed edges due to Holland and Leinhardt, known as p_1 , which allows for differential attraction (popularity) and expansiveness, as well as an additional effect due to reciprocation. In particular, we attempt to derive Markov bases for p_1 and to link these to the results on Markov bases for working with log-linear models for contingency tables. Because of the contingency table representation for p_1 we expect some form of congruence. Markov bases and related algebraic geometry notions are useful for at least two statistical problems: (i) determining condition for the existence of maximum likelihood estimates, and (ii) using them to traverse conditional (given minimal sufficient statistics) sample spaces, and thus generating “exact” distributions useful for assessing goodness of fit. We outline some of these potential uses for the algebraic representation of p_1 .

Algebraic statistics and contingency tables

Adrian Dobra, UNIVERSITY OF WASHINGTON

In this talk I will give an overview of the role of algebraic statistics in the statistical analysis of contingency tables. I will survey major areas in which algebraic methods proved to be crucial and provided a fertile ground for novel research directions: computation of sharp integer bounds for cell entries, existence of maximum likelihood estimates, simulation from probability distributions on spaces of tables, Markov bases, high-dimensional sparse tables with structural zeros, log-linear model selection. I will give examples that illustrate this methodology and talk about open problems.

Toric Modification on Mixture Models

Keisuke Yamazaki, TOKYO INSTITUTE OF TECHNOLOGY

Sumio Watanabe, TOKYO INSTITUTE OF TECHNOLOGY

In the Bayes estimation, it was pointed out that resolution of singularity provides an algorithm to elucidate the generalization performance of learning machines. However, there is no effective procedure to find the resolution map. This presentation proposes a new method to find it based on the toric modification, using Newton diagram. By the proposed method, learning curves of several hierarchical models are clarified.

Learning Parameters in Discrete Naive Bayes Models by Computing Fibers of the Parametrization map

Vincent Auvray, UNIVERSITY OF LIÉGE

Louis Wehenkel, UNIVERSITY OF LIÉGE

Discrete Naive Bayes models are usually defined parametrically with a map from a parameter space to a probability distribution space. First, we present two families of algorithms that compute the set of parameters mapped to a given discrete Naive Bayes distribution satisfying certain technical assumptions. Using these results, we then present two families of parameter learning algorithms that operate by projecting the distribution of observed relative frequencies in a dataset onto the discrete Naive Bayes model considered. They have nice convergence properties, but their computational complexity grows very quickly with the number of hidden classes of the model.

Stationary Subspace Analysis

Paul von Büna, TU BERLIN

Frank C. Meinecke, TU BERLIN

Klaus-Robert Müller, TU BERLIN

Non-stationarities are an ubiquitous phenomenon in real-world data, yet they challenge standard Machine

Learning methods: if training and test distributions differ we cannot, in principle, generalise from the observed training sample to the test distribution. This affects both supervised and unsupervised learning algorithms. In a classification problem, for instance, we may infer spurious dependencies between data and label from the the training sample that are mere artefacts of the non-stationarities. Conversely, identifying the sources of non-stationary behaviour in order to better understand the analyzed system often lies at the heart of a scientific question. To this end, we propose a novel unsupervised paradigm: Stationary Subspace Analysis (SSA). SSA decomposes a multi-variate time-series into a stationary and a non-stationary subspace. We derive an efficient algorithm that hinges on an optimization procedure in the Special Orthogonal Group. By exploiting the Lie group structure of the optimization manifold, we can explicitly factor out the inherent symmetries of the problem and thereby reduce the number of parameters to the exact degrees of freedom. The practical utility of our approach is demonstrated in an application to Brain Computer-Interfacing (BCI).

Alternatives to the Discrete Fourier Transform

Doru Balcan, CARNEGIE MELLON UNIVERSITY

Aliaksei Sandryhaila, CARNEGIE MELLON UNIVERSITY

Jonathan Gross, CARNEGIE MELLON UNIVERSITY

Markus Püschel, CARNEGIE MELLON UNIVERSITY

It is well-known that the discrete Fourier transform (DFT) of a finite length discrete-time signal samples the discrete-time Fourier transform of the same signal at equidistant points on the unit circle. Hence, as the signal length goes to infinity, the DFT approaches the DTFT. Associated with the DFT are circular convolution and a periodic signal extension. In this paper we identify a large class of alternatives to the DFT using the theory of polynomial algebras. Each of these Fourier transforms approaches the DTFT just as the DFT does, but has its own signal extension and notion of convolution, which therefore are not periodic. Furthermore, these Fourier transforms have Vandermonde structure, which enables their computation via fast $O(n \log^2(n))$ algorithms.

Graph Helmholtzian and rank learning

Lek-Heng Lim, UNIVERSITY OF CALIFORNIA, BERKELEY

The graph Helmholtzian is the graph theoretic analogue of the Helmholtz operator or vector Laplacian, in much the same way the graph Laplacian is the analogue of the Laplace operator or scalar Laplacian. We will see that a decomposition associated with the graph Helmholtzian provides a way to learn ranking information from incomplete, imbalanced, and cardinal score-based data. In this framework, an edge flow representing pairwise ranking is orthogonally resolved into a gradient flow (acyclic) that represents the L2-optimal global ranking and a divergence-free flow (cyclic) that quantifies the inconsistencies. If the latter is large, then the data does not admit a statistically meaningful global ranking. A further decomposition of the inconsistent component into a curl flow (locally cyclic) and a harmonic flow (locally acyclic) provides information on the validity of small- and large-scale comparisons of alternatives. This is joint work with Xiaoye Jiang, Yuan Yao, and Yinyu Ye.

Identity Management On Homogeneous spaces

Xiaoye Jiang, STANFORD UNIVERSITY

Leonidas J. Guibas, STANFORD UNIVERSITY

We consider the identity management problem, where the identities are classified into two classes, red and blue. The purpose here is to make predictions of the two class identities when confusions arise among identities. In this work, we propose a principle to maintain probability distributions over homogeneous space which provides a mechanism valid for taking into account of any desired degree of approximation. Markov models are used to formulate the two class identity management problem which tries to compactly summarize distributions on homogeneous spaces. Projecting down and lifting up information on different order of statistics can be achieved by using Radon transformations. The commutative property of Markov updating with Radon transform enable us to maintain exact information over different order of statistics. Thus, accurate classification predictions can be made based on the low order statistics we maintained. We evaluate the performance of our algorithms on a real camera network data and show effectiveness of our scheme.

Exploiting Probabilistic Independence for Permutations**Jonathan Huang**, CARNEGIE MELLON UNIVERSITY**Carlos Guestrin**, CARNEGIE MELLON UNIVERSITY**Xiaoye Jiang**, STANFORD UNIVERSITY**Leonidas J. Guibas**, STANFORD UNIVERSITY

Permutations are ubiquitous in many real world problems, such as voting, rankings and data association. Representing uncertainty over permutations is challenging, since there are $n!$ possibilities. Recent Fourier-based approaches can be used to provide a compact representation over low-frequency components of the distribution. Though polynomial, the complexity of these representations grows very rapidly, especially if we want to maintain reasonable estimates for peaked distributions. In this talk, we first characterize the notion of probabilistic independence for distribution over permutations. We then present a method for factoring distributions into independent components in the Fourier domain and use our algorithms to decompose large problems into much smaller ones. Building on this method, we describe an algorithm that detects independence and adapts the choice of representation according to the complexity of the underlying problem. We demonstrate that our method provides very significant improvements in terms of running time, on real tracking data.

Consistent Structured Estimation for Weighted Bipartite Matching**James Petterson**, NICTA CANBERRA**Tiberio Caetano**, NICTA CANBERRA**Julian McAuley**, NICTA CANBERRA

Given a weighted bipartite graph, the assignment problem consists of finding the heaviest perfect match. This is a classical problem in combinatorial optimization, which is solvable exactly and efficiently by standard methods such as the Hungarian algorithm, and is widely applicable in real-world scenarios. We give an exponential family model for the assignment problem. Edge weights are obtained from a suitable composition of edge features and a parameter vector, which is learned so as to maximize the likelihood of a sample consisting of training graphs and their labeled matches. The resulting consistent estimator contrasts with existing max-margin structured estimators, which are inconsistent for this problem.

DECEMBER 12, 2008, 07:30–10:30 AND 15:30–18:30

WESTIN: ALPINE AB WS5

Analyzing Graphs: Theory and Applications

<http://research.yahoo.com/workshops/nipsgraphs2008/>

Edoardo Airoldi

PRINCETON UNIVERSITY

eairoldi@princeton.edu

David Blei

PRINCETON UNIVERSITY

blei@cs.princeton.edu

Jake Hofman

YAHOO-INC.

hofman@yahoo-inc.com

Tony Jebara

COLUMBIA UNIVERSITY

jebara@cs.columbia.edu

Eric Xing

CARNEGIE MELLON UNIVERSITY

epxing@cs.cmu.edu

Abstract

Recently, statistics and machine learning have seen the proliferation of both theoretical and computational tools for analyzing graphs to support progress in applied domains such as social sciences, biology, medicine, neuroscience, physics, finance, and economics. This workshop actively promotes a concerted effort to address statistical, methodological and computational issues that arise when modeling and analyzing large collection of data which are primarily represented as static and/or dynamic graphs. Presentations include (but are not limited to) novel graph models, the application of established models to new domains, theoretical and computational issues, limitations of current graph methods and directions for future research. The workshop aims to bring together researchers from applied disciplines such as sociology, economics, medicine and biology with researchers from mathematics, physics, statistics and computer science.

07.30-07.35	Opening remarks
07.35-08.15	Invited talk. Graph-based methods for open information extraction WILLIAM COHEN, CARNEGIE MELLON UNIVERSITY
08.15-08.30	Connections between the lines: Extracting social networks from text JONATHAN CHANG, JORDAN BOYD-GRABER, DAVID BLEI
08.30-08.45	Gibbs sampling for logistic normal topic models with graph based priors DAVID MIMNO, HANNA M. WALLACH, ANDREW MCCALLUM
08.45-08.55	Coffee Break
08.55-09.35	Invited talk. From here to eternity: Developing dynamic network models STEPHEN FIENBERG, CARNEGIE MELLON UNIVERSITY
09.35-10.30	Poster Session
10.30-03.00	Skiing / poster session continued
03.30-04.10	Invited talk. A statistical perspective on large-scale network data: The blending of inference and algorithms for analysis PATRICK WOLFE, HARVARD UNIVERSITY
04.10-04.25	Maximum likelihood graph structure estimation with degree distributions BERT HUANG, TONY JEBARA

- 04.25-04.40** **Probabilistic graph models for debugging software**
LAURA DIETZ, VALENTIN DALLMEIER
- 04.40-04.50** Coffee Break
- 04.50-05.30** **Invited talk. Size matters: Benefits from studying large networks**
JURE LESKOVEC, CORNELL UNIVERSITY
- 05.30-05.45** **Time Varying Ising Models**
MLADEN KOLAR, ERIC XING
- 05.45-06.00** **Uncovering latent structure in valued graphs: A variational approach**
MAHENDRA MARIADASSOU, STEPHANE ROBIN, CORRINE VACHER
- 06.00-06.30** Panel Discussion

Graph-based methods for open information extraction

William Cohen, CARNEGIE MELLON UNIVERSITY AND GOOGLE

Traditional information extraction (IE) uses supervised learning to add structure to information in free text. Recent work in "open IE" uses unsupervised or lightly-supervised learning methods for extracting information from text, with a focus on extracting information from large redundant corpora (such as the web). For example, traditional IE techniques might learn to extract entities of specific, pre-determined types, such as people, locations, or protein names, using large annotated-corpora; in contrast, open IE techniques might learn to extract instances of types like "reality TV shows" or "rugby teams" using unlabeled documents from the web and minimal type-specific information from a user - perhaps a handful of seed instances and/or the name of the type. In my talk, I will survey recent work on open IE using graphs constructed from text. I will show that open IE can be used to obtain highly accurate entity lists from dozens of entity types and three languages, using graphs built from semi-structured documents on the web. I will show that experimentally, certain variants of the personalized PageRank similarity measure appear to be especially well-suited to this problem. I will also report on recent work in using adaptively similarity metrics for the more difficult task of open IE from dependency-parsed free text. This is joint work with Einat Minkov and Richard C. Wang.

From here to eternity: Developing dynamic network models

Stephen Fienberg, DEPARTMENT OF STATISTICS AND MACHINE LEARNING DEPARTMENT, CARNEGIE MELLON UNIVERSITY

Much of the recent literature on the modeling of network data has focused on snapshots of networks, often accumulated over periods of time. More interesting are dynamic network models but these are often simplistic or focus on selected network characteristics. In this presentation we attempt to provide a common framework for the modeling of networks evolving over time and we discuss how different strategies that appear in the literature fit within the framework.

A statistical perspective on large-scale network data: The blending of inference and algorithms for analysis

Patrick Wolfe, STATISTICS AND INFORMATION SCIENCES LABORATORY, HARVARD UNIVERSITY

Modern science and engineering applications give rise to vast quantities of network data, and in this talk we provide a statistical perspective on the challenges and opportunities that these data sets present. We begin by relating notions of classical statistics to the context of graph-valued data sets, with a particular focus on formal hypothesis testing for network structure based on exchangeability. The exact inference problem is quickly seen to be impractical, however, and traditional approaches singularly fail to scale. We subsequently introduce practical solutions by way of algorithms for data reduction, along with bounds and performance guarantees. We then stress connections to induced subgraphs and matrix completion problems, and conclude with several open questions for the future of network inference.

Size matters: Benefits from studying large networks

Jure Leskovec, CORNELL UNIVERSITY

With the rise of the internet and the web large amounts of human social interaction data became available. This offered great opportunities to study social behaviors and information dynamics at scales and magnitudes never possible before. In this talk I will present two such examples where studying large networks lead us to novel findings and observations that would practically be impossible when working with small data sets. First, I will present our work on the “planetary scale” dynamics and the structure of the full Microsoft Instant Messenger communication network that contains 240 million people, with more than 255 billion exchanged messages per month, which makes it the largest social network studied to date. We will investigate the “6.6 degrees of separation” of Messenger and examine how to best search and navigate world’s social network. Second example will investigate the properties of community structure in networks, where we find that there is a natural size scale to network cluster size and the absence of large well-defined clusters. We employ approximation algorithms for the graph partitioning problem to characterize statistical and structural properties of partitions of graphs. We observe tight communities that are barely connected to the rest of the network at very small size scales (up to 100 nodes); and communities of size scale beyond 100 nodes gradually “blend into” the expander-like core of the network and thus become less “community-like”. We will then investigate modeling questions and implications on the structure of large networks.

A path following algorithm for the graph matching problem

Mikhail Zaslavskiy,

Francis Bach,

Jean-Philippe Vert,

We propose a convex-concave programming approach for the labeled weighted graph matching problem. The convex-concave programming formulation is obtained by rewriting the weighted graph matching problem as a least-square problem on the set of permutation matrices and relaxing it to two different optimization problems: a quadratic convex and a quadratic concave optimization problem on the set of doubly stochastic matrices. The concave relaxation has the same global minimum as the initial graph matching problem, but the search for its global minimum is also a hard combinatorial problem. We therefore construct an approximation of the concave problem solution by following a solution path of a convex-concave problem obtained by linear interpolation of the convex and concave formulations, starting from the convex relaxation. This method allows to easily integrate the information on graph label similarities into the optimization problem, and therefore to perform labeled weighted graph matching. The algorithm is compared with some of the best performing graph matching methods on three datasets: simulated graphs, QAPLib and handwritten chinese characters. In all cases, the results are competitive with the state-of-the-art.

A simple infinite topic mixture for rich graphs and relational data

Janne Sinkkonen,

Juuso Parkkinen,

Janne Aukia,

Samuel Kaski,

We propose a simple component or topic model for relational data, that is, for heterogeneous collections of co-occurrences between categorical variables. Graphs are a special case, as collections of dyadic co-occurrences (edges) over a set of vertices. The model is especially suitable for finding global components from collections of massively heterogeneous data, where encoding all the relations to a more sophisticated model becomes cumbersome, as well as for quick-and-dirty modeling of graphs enriched with, e.g., link properties or nodal attributes. The model is here estimated with collapsed Gibbs sampling, which allows sparse data structures and good memory efficiency for large data sets. Other inference methods should be straightforward to implement. We demonstrate the model with various medium-sized data sets (scientific citation data, MovieLens ratings, protein interactions), with brief comparisons to a full relational model and other approaches.

Adjusting for network size and composition effects in exponential random graphs

Pavel Krivitsky,

Exponential random graph models (ERGMs) provide a principled way to model and simulate features common in social networks, particularly those of people, such as homophily and friend-of-a-friend dynamics. We

show that these models have trouble realistically modeling effects of changes in network size and composition, and suggest an offset model, which we argue produces realistic behavior asymptotically.

Connections between the lines: Extracting social networks from text

Jonathan Chang,
Jordan Boyd-Graber,
David Blei,

Relational data is ubiquitous, encoding collections of relationships between entities such as people, places, genes, or corporations. For some entities of interest, graphs of their connections are explicitly collected and represented. For many other collections, however, the graph of connections is implicitly encoded in a text corpus. In this paper we develop a new machine learning framework for analyzing such texts to find the hidden graph of connections between their entities. Our method is based on a novel probabilistic model that uncovers the underlying graph and provides keywords to describe the salient characteristics of its nodes and edges. We study three corpora with our technique: the Bible, Wikipedia, and scientific abstracts to uncover the connections between biblical characters, notable people, and genes. We report qualitative and quantitative results.

Gaussian process models for colored graphs

Zhao Xu,
Kristian Kersting,
Volker Tresp,

Many real-world domains can naturally be represented as a complex graph, i.e., in terms of entities (nodes) and relations (edges) among them. In domains with multiple relations, represented as colored graphs, we may further improve the quality of a model by exploiting information from one relation while modeling another. To this end, we develop a multi-relational Gaussian process (MRGP) model. MRGPs treat relations and node colors as the non-parameterized function of all other related information. We do not simply integrate all information into a single GP, instead, several latent variables are drawn from different GPs: one for each type of entities and relations. These latent variables respectively represent the profile of and hidden causes among the entities in the domain. To couple the GPs together, the relations depend on the linear combination of related latent values. We give an analysis of the MRGP model for bipartite, directed and undirected univariate relations.

Gibbs sampling for logistic normal topic models with graph based priors

David Mimno,
Hanna Wallach,
Andrew McCallum,

Previous work on probabilistic topic models has either focused on models with relatively simple conjugate priors that support Gibbs sampling or models with non-conjugate priors that typically require variational inference. Gibbs sampling is more accurate than variational inference and better supports the construction of composite models. We present a method for Gibbs sampling in non-conjugate logistic normal topic models, and demonstrate it on a new class of topic models with arbitrary graph-structured priors that reflect the complex relationships commonly found in document collections, while retaining simple, robust inference.

Improved algorithm and data structures for modularity analysis of large networks

Alexandre Francisco,

Graph clustering is an important problem in the analysis of computer networks, social networks, biological networks and many other natural and artificial networks. These networks are in general very large and, thus, finding hidden structures and functional modules is a very hard task. In this paper we propose new data structures and make available a new implementation of a well known agglomerative greedy algorithm to find community structure in large networks. The experimental results show that the improved data structures speedup the method by a large factor, for very large networks.

Large-scale stochastic relational models

Kai Yu,

Shenghuo Zhu,

Stochastic relational models (SRMs) are a family of models for learning and predicting dyadic data between two sets of entities. The models generalize matrix factorization to a supervised learning problem that utilizes attributes of entities in a hierarchical Bayesian framework. Previously variational Bayes inference was applied for SRMs, which is, however, not scalable when the size of either entity set grows to tens of thousands. In this paper, we introduce a Markov chain Monte Carlo (MCMC) algorithm for equivalent models of SRMs in order to scale the computation to very large dyadic data sets. Both superior scalability and predictive accuracy are demonstrated on a collaborative filtering problem, which involves tens of thousands users and half million items.

”Maximum likelihood graph structure estimation with degree distributions**Bert Huang,****Tony Jebara,**

We describe a generative model for graph edges under specific degree distributions which admits an exact and efficient inference method for recovering the most likely structure. This binary graph structure is obtained by reformulating the inference problem as a generalization of the polynomial time combinatorial optimization problem known as b-matching, which recovers a degree constrained maximum weight subgraph from an original graph. After this mapping, the most likely graph structure can be found in cubic time with respect to the number of nodes using max flow methods. Furthermore, in some instances, the combinatorial optimization problem can be solved exactly in cubic time by loopy belief propagation and max product updates. Empirical results show the method’s ability to recover binary graph structure with appropriate degree distributions from partial or noisy information.

Predicting gene function in a hierarchy**Sara Mostafavi,****Quaid Morris,**

We present a method for hierarchical multilabel classification using network based input data and apply it to the prediction of gene function. In this setting, genes are represented by nodes and their similarities (association) are represented by the edges of the network. Although genes have multiple functions, most previous approaches predict gene functions as independent classification problems. Here we extend a graph-based semi-supervised learning algorithm to predict multiple gene functions according to a hierarchy (e.g. Gene Ontology). Our results demonstrate a considerable improvement over state-of-the-art approaches.

Probabilistic graph models for debugging software**Laura Dietz,****Valentin Dallmeier,**

Of all software development activities, debugging, locating the defective source code statements that cause a failure can be by far the most time-consuming. We employ probabilistic modeling to support programmers in finding defective code. Most defects are identifiable in control flow graphs of software traces. A trace is represented by a sequence of code positions (line numbers in source filenames) that are executed when the software runs. The control flow graph represents the finite state machine of the program, in which states depict code positions and arcs indicate valid follow up code positions. In this work, we extend this definition towards an n-gram control flow graph, where a state represents a fragment of subsequent code positions, also referred to as an n-gram of code positions. We devise a probabilistic model for such graphs in order to infer code positions in which anomalous program behavior can be observed. This model is evaluated on real world data obtained from the open source AspectJ project and compared to the well known multinomial and multi-variate Bernoulli model.

Re-weighting graph links for quantifying difference**Yu-Shi Lin,****Chung-Chi Lin,****Yuh-Show Tsai,****Tien-Chuan Ku,****Yi-Hung Huang,**

Chun-Nan Hsu,

In this paper, we describe a new approach to quantifying difference between sets of high dimensional data points in a Euclidean space. In this method, data points are connected with their neighbors to form a graph and graph transition energy is used to quantify the difference. To stabilize quantification against sub-sampling variances, we apply Cheeger's constant regularization in the spectral graph theory to re-weight links based on labeled training examples. The regularization also allows us to measure how well a given set of features can quantify the difference. We empirically show that the method is stable for images of hand-written digits and report a real application on quantifying differences of microscopic cell images for drug discovery. The method may potentially be applied to other fields of studies where data points are given as a multi-graph with weighted links.

Selection of regularization parameter in sparse MRF learning: A Bayesian approach**Narges Asadi,****Irina Rish,****Katya Scheinberg,**

Recently proposed l_1 -regularized maximum-likelihood optimization methods for learning sparse Markov networks result into convex problems that can be solved optimally and efficiently. However, the accuracy of such methods can be very sensitive to the choice of regularization parameter, and optimal selection of this parameter remains an open problem. Herein, we propose a Bayesian approach that investigates the effect of a prior on the regularization parameter, and yields promising empirical results on both synthetic data and real-life application such as brain imaging data (fMRI).

Sparse multiscale regression for graphical models**Justin Guinney,****Simon Lunagomez,****Mauro Maggioni,****Sayan Mukherjee,**

We present a novel framework for multiscale regression of high-dimensional data when *a priori* local dependencies of the variables are known. The multiscale analysis adapts to nonlinear structure in the data and complex dependencies among the variables and it yields sparse representations for large classes of functions. Due to localization properties and the hierarchical structure imposed by the multiscale model we find these models to be more interpretable than other manifold models based on spectral methods. We show promising results on a variety of highdimensional classification problems with respect to accuracy, sparsity, as well as interpretability when we compare our approach to other methods.

Temporally-evolving mixed membership stochastic blockmodels: Exploring the Enron e-mail database**Seungil Huh,****Stephen Fienberg,**

The e-mail database for the Enron Corp. linked to the prosecution of a number of its senior executives poses an interesting challenge for researchers interested in network modeling. In this paper we adapt the mixed membership stochastic blockmodel approach to a time-varying setting and apply this extension to a version of the Enron database.

The information in one prior relative to another**Michael Evans,****Gun Ho Jang,**

A question of some interest is how to characterize the amount of information that a prior puts into a statistical analysis. Rather than a general characterization of this quantity, we provide here an approach to characterizing the amount of information a prior puts into an analysis, when compared to another base prior. The base prior is considered to be the prior that best reflects the current available information. Our purpose then, is to characterize priors that can be used as conservative inputs to an analysis, relative to the base prior, in the sense that they put less information into the analysis. The characterization that we provide is in terms of *a priori* measures of prior-data conflict.

Time varying Ising models

Mladen Kolar,

Eric Xing,

In this paper, we propose a nonparametric method for estimating of the structure of a discrete undirected graphical models from data. We assume that the distribution generating the data smoothly evolves over time and that the given sample is not identically distributed. Under the assumption that the underlying graphical model is sparse, our method recovers the structure consistently even the high dimensional case where the ambient dimension is larger than the size of the sample.

Topic Models for Hypertext: How many words is a single link worth?

Amit Gruber,

Michal Rosen-Zvi,

Yair Weiss,

Latent topic models have been successfully applied as an unsupervised learning technique on various types of data such as text documents, images and biological data. In recent years, with the rapid growth of the Internet, these models have also been adapted to hypertext data. Explicitly modeling the generation of both words and links has been shown to improve inferred topics and open a new range of applications for topic models. However, it remains unclear how to balance the information contributed by links with the information contributed by words. In this paper we enrich the Latent Topic HypertextModel [7] with a parameter that stands for importance of links in the model. Specifically, we quantitatively explore whether a single link is more indicative of the topic mixture in the document it points to than a single word. We show that putting an emphasis on the topics associated with links leads to better link prediction results.

Towards understanding network dynamics

Vladimir Marbukh,

This paper discusses possible approaches and challenges of modeling networks in market economy as a non-cooperative game with users adjusting their demand in attempt to maximize their net utilities and providers adjusting their supply and pricing in attempt to maximize their profit. It is natural to assume that the network dynamics follows the evolutionary/learning algorithms approaching Nash equilibria of this game. Due to typical multiplicity of the Nash equilibria, this dynamics is highly non-linear and exhibits complex behavior very sensitive to the initial conditions. In a case of perfect competition, when providers charge users only the marginal resource cost, the network evolution results in “social welfare” maximization, which may be interpreted as “entropy” maximization due to similarities of these two concepts.

Uncovering latent structure in valued graphs: A variational approach

Mahendra Mariadassou,

Stephane Robin,

Corrine Vacher,

As more and more network-structured datasets are available, the statistical analysis of valued graphs has become a common place. Looking for a latent structure is one of the many strategies used to better understand the behavior of a network. Several methods already exist for the binary case. We present a model-based strategy to uncover groups of nodes in valued graphs. This framework can be used for a wide span of parametric random graphs models. Variational tools allow us to achieve approximate maximum likelihood estimation of the parameters of these models. We provide a simulation study showing that our estimation method performs well over a broad range of situations. We apply this method to analyze interaction networks of tree and fungal species.

Visualizing graphs with structure preserving embedding

Blake Shaw,

Tony Jebara,

Structure Preserving Embedding (SPE) is a method for embedding graphs in lowdimensional Euclidean space such that the embedding preserves the graph’s global topological properties. Specifically, topology is preserved if a connectivity algorithm can recover the original graph from only the coordinates of its nodes after embedding. Given an input graph and an algorithm for linking embedded nodes, SPE learns a low-

rank kernel matrix by means of a semidefinite program with linear constraints that captures the connectivity structure of the input graph. The SPE cost function ensures that the learned kernel is low-rank and thus the resulting embedding uses low-dimensional coordinates for each node that reproduce the original graph when processed by a connectivity algorithm (such as k-nearest neighbors, or b-matching). SPE provides significant improvements in terms of visualization and lossless compression of graphs, outperforming popular methods such as spectral embedding and spring embedding. Furthermore, we find that many classical graphs and networks can be properly embedded using only a few dimensions.

Community detection: Model fitting, comparison, and utility

Jake Hofman,

Much recent work has focused on community detection, or the task of identifying sets of similar nodes from network topology. Underlying this work is the implicit assumption that inferred communities inform node attributes or function in a meaningful and useful sense. We investigate these ideas by phrasing community detection as Bayesian inference, which provides a scalable and efficient algorithm for fitting and comparing network models, and applying the resulting algorithm to a university e-mail data set that includes both topology (who e-mailed whom) and node attributes (age, gender, academic affiliation, etc.). We study the relationship between the identified topological communities and the node attributes and discuss implications for community detection as a tool for network analysis.

The exchangeable graph model

Edoardo Airoldi,

Collections of pairwise measurements arise in a number of settings in the biological sciences (e.g., www.yeastgenome.org), with collections of scientific publications (e.g., www.jstor.org) and other hyper-linked resources (e.g., www.wikipedia.org), and in social networks (e.g., www.linkedin.com). In this talk we introduce the exchangeable graph model, a simple extension of the random graph model by Erdos & Renyi (1959) and Gilbert (1959). The exchangeable graph model can instantiate realistic connectivity patterns, is amenable to mathematical analysis, and preserves phenomena such as the emergence of a giant component. We demonstrate the utility of the exchangeable graph model in solving two open problems in statistical network analysis: 1. model selection among very different statistical models of pairwise measurements, and 2. assessing the statistical significance associated with the observed overlap of independent cliques in a graph.

Program Committee: David Banks (Duke University), Peter Bearman (Columbia University), Joseph Blitzstein (Harvard University), Kathleen Carley (Carnegie Mellon University), Jonathan Chang (Princeton University), Aaron Clauset (Santa Fe Institute), William Cohen (Carnegie Mellon University), Stephen Fienberg (Carnegie Mellon University), Paolo Frasconi (Universita degli Studi di Firenze), Lise Getoor (University of Maryland), Peter Hoff (University of Washington), Eric Horvitz (Microsoft Research), Alan Karr (National Institute of Statistical Sciences), Jure Leskovec (Cornell University), Kevin Murphy (University of British Columbia), Eugene Stanley (Boston University), Lyle Ungar (University of Pennsylvania), Chris Wiggins (Columbia University).

DECEMBER 12, 2008, 07:45–10:30 AND 15:45–18:30

HILTON: MT. CURRIE S **WS6**

Machine Learning in Computational Biology

<http://www.mlcb.org>

Gal Chechik

GOOGLE RESEARCH

Christina Leslie

MEMORIAL SLOAN-KETTERING CANCER CENTER

Quaid Morris

UNIVERSITY OF TORONTO

William Noble

UNIVERSITY OF WASHINGTON

Gunnar Raetsch

FRIEDRICH MIESCHER LABORATORY, MAX PLANCK SOCIETY (TUEBINGEN)

gal@ai.stanford.edu

cleslie@cbio.mskcc.org

quaid.morris@utoronto.ca

noble@gs.washington.edu

Gunnar.Raetsch@tuebingen.mpg.de

Abstract

The field of computational biology has seen dramatic growth over the past few years, both in terms of new available data, new scientific questions, and new challenges for learning and inference. In particular, biological data is often relationally structured and highly diverse, well-suited to approaches that combine multiple weak evidence from heterogeneous sources. These data may include sequenced genomes of a variety of organisms, gene expression data from multiple technologies, protein expression data, protein sequence and 3D structural data, protein interactions, gene ontology and pathway databases, genetic variation data (such as SNPs), and an enormous amount of textual data in the biological and medical literature. New types of scientific and clinical problems require the development of novel supervised and unsupervised learning methods that can use these growing resources. The goal of this workshop is to present emerging problems and machine learning techniques in computational biology. We invited several speakers from the biology/bioinformatics community who will present current research problems in bioinformatics, and we invite contributed talks on novel learning approaches in computational biology. We encourage contributions describing either progress on new bioinformatics problems or work on established problems using methods that are substantially different from standard approaches. Kernel methods, graphical models, feature selection and other techniques applied to relevant bioinformatics problems would all be appropriate for the workshop.

Morning session

7.45-8.10

Learning Temporal Sequence of Biological Networks

LE SONG AND ERIC XING

8.10-8.35

Switching Regulatory Models of Cellular Stress Response

GUIDO SANGUINETTI, ANDREAS RUTTOR, MANFRED OPPER AND CEDRIC ARCHAMBEAU

8.35-9.00

Detecting the Presence and Absence of Causal Relationships Between Expression of Yeast Genes with Very Few Samples

EUN YONG KANG, ILYA SHPITSER, HYUN MIN KANG, CHUN YE AND ELEAZAR ESKIN

9.00-9.15

Coffee

9.15-9.40

KIRMES: Kernel-based Identification of Regulatory Modules in Euchromatic Sequences

SEBASTIAN J. SCHULTHEISS, WOLFGANG BUSCH, JAN LOHMANN, OLIVER KOHLBACHER
AND GUNNAR RÄTSCH

9.40-10.05 **Approximate Substructure Matching for Biological Sequence Classification**

PAVEL KUKSA AND VLADIMIR PAVLOVIC

10.05-10.30 **Predicting Binding Affinities of MHC Class II Epitopes Across Alleles**
NICO PFEIFER AND OLIVER KOHLBACHER

Afternoon session

3.45-4.10 **Inside the black box: Identifying causal genetic factors of drug resistance**
BO-JUEN CHEN, HELEN CAUSTON, ETHAN PERLSTEIN AND DANA PEER

4.10-4.35 **Full Bayesian Survival Models for Analyzing Human Breast Tumors**
VOLKER ROTH, THOMAS FUCHS, SUDHIR RAMAN, PETER WILD, EDGAR DAHL
AND JOACHIM BUHMANN

4.35-5.00 **Probabilistic assignment of formulas to mass peaks in metabolomics experiments**
SIMON ROGERS, RICHARD A. SCHELTEMA, MARK GIROLAMI AND RAINER BREITLING

5.00-5.15 Coffee

5.15-5.40 **Learning “graph-mer” motifs that predict gene expression trajectories in development**
XUEJING LI, CHRIS WIGGINS, VALERIE REINKE AND CHRISTINA LESLIE

5.40-6.05 **On the relationship between DNA periodicity and local chromatin structure**
SHEILA REYNOLDS, JEFF BILMES AND WILLIAM STAFFORD NOBLE

6.05-6.30 Discussion

DECEMBER 12, 2008, 07:30–10:40 AND 15:30–18:40

HILTON: BLACK TUSK WS7

Machine Learning Meets Human Learning

<http://pages.cs.wisc.edu/~jerryzhu/nips08.html>

Nathaniel Daw

NEW YORK UNIVERSITY

daw@gatsby.ucl.ac.uk

Tom Griffiths

UNIVERSITY OF CALIFORNIA, BERKELEY

tom_griffiths@berkeley.edu

Josh Tenenbaum

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

jbt@mit.edu

Xiaojin Zhu

UNIVERSITY OF WISCONSIN-MADISON

jerryzhu@cs.wisc.edu

Abstract

Can statistical machine learning theories and algorithms help explain human learning? Broadly speaking, machine learning studies the fundamental laws that govern all learning processes, including both artificial systems (e.g., computers) and natural systems (e.g., humans). It has long been understood that theories and algorithms from machine learning are relevant to understanding aspects of human learning. Human cognition also carries potential lessons for machine learning research, since people still learn languages, concepts, and causal relationships from far less data than any automated system. There is a rich opportunity to develop a general theory of learning which covers both machines and humans, with the potential to deepen our understanding of human cognition and to take insights from human learning to improve machine learning systems. The goal of this workshop is to bring together the different communities that study machine learning, cognitive science, neuroscience and educational science. We will investigate the value of advanced machine learning theories and algorithms as computational models for certain human learning behaviors, including, but not limited to, the role of prior knowledge, learning from labeled and unlabeled data, learning from active queries, and so on. We also wish to explore the insights from the cognitive study of human learning to inspire novel machine learning theories and algorithms. It is our hope that the NIPS workshop will provide a venue for cross-pollination of machine learning approaches and cognitive theories of learning to spur further advances in both areas.

- | | |
|------------------|---|
| 7.30-7.35 | Welcome
WORKSHOP ORGANIZERS |
| 7.35-7.55 | Training Deep Architectures: Inspiration from Humans
YOSHUA BENGIO |
| 7.55-8.15 | Stochastic programs as a framework for clustering and causation
NOAH GOODMAN |
| 8.15-8.35 | Learning abstract causal knowledge: a case study in human and machine learning
JOSH TENENBAUM |
| 8.35-8.55 | Rational Approximations to Rational Models of Categorization
ADAM SANBORN |
| 8.55-9.10 | Coffee |
| 9.10-9.30 | The role of prior knowledge in human reconstructive memory
MARK STEYVERS |

- 9.30-9.50** **Normative models of multiple interacting memory systems**
MATE LENGYEL
- 9.50-10.10** **Compositional Logic Learning**
ALAN YUILLE
- 10.10-10.40** Panel Discussion 1: Probabilistic models of cognition
- 10.40-** Poster or Ski
- 15.30-15.50** **Reconciling reinforcement learning and risk sensitivity: a model-based fMRI study**
YAEL NIV
- 15.50-16.10** **Goal-directed decision making as structured probabilistic inference**
MATTHEW BOTVINICK
- 16.10-16.30** **Reward bonuses for efficient, effective exploration**
MICHAEL LITTMAN
- 16.30-16.50** **Human Semi-Supervised Learning and Human Active Learning**
XIAOJIN ZHU
- 16.50-17.05** Coffee
- 17.05-17.25** **Where am I and what should I do next? Overcoming perceptual aliasing in sequential tasks**
TODD GURECKIS
- 17.25-17.45** **Using reinforcement learning models to interpret human performance on Markov decision problems**
MICHAEL MOZER
- 17.45-18.05** **A Bayesian Algorithm for Change Detection with Identification: Rational Analysis and Human Performance**
JUN ZHANG
- 18.05-18:40** Panel Discussion 2: Decision and reward
- 18.40-** Poster

Training Deep Architectures: Inspiration from Humans

Yoshua Bengio, UNIVERSITE DE MONTREAL

Theoretical results in circuit complexity suggest that deep architectures are necessary to efficiently represent highly-varying functions, which may be needed for many AI tasks. However, training deep architectures is not only non-convex, but the optimization difficulty seems to increase for deeper architectures. Can we get inspiration from how humans manage to learn complicated concepts and high-level abstractions? The first successful algorithms for training deep architectures suggest a principle is at work: first optimizing something easier (learning concepts that can be represented with shallower architectures), and gradually increasing the difficulty (increasing depth), in such a way as to guide the optimization towards better basins of attraction of a local optimization procedure. Another related principle that we are exploring involves breaking from the traditional iid dataset methodology, and breaking training into gradually more difficult phases (and data streams), where each phase allows the learner to learn more complex concepts, exploiting previously learned concepts. Other inspiration from how humans learn complicated concepts will be discussed.

Stochastic programs as a framework for clustering and causation

Noah Goodman, MIT

I will consider a series of concept learning problems faced by people in everyday life. These will proceed from simple clustering problems to the problem of learning latent events underlying a stream of input and the causal relations amongst these events. Each learning problem will be formulated as a stochastic program in the Church language, and I will argue that this framework permits flexible and rapid investigation of learning problems for both cognitive science and machine learning.

Learning abstract causal knowledge: a case study in human and machine learning

Josh Tenenbaum, MIT

TBA

Rational Approximations to Rational Models of Categorization

Adam Sanborn, GATSBY, UNIVERSITY COLLEGE LONDON

Rational models have been successfully used to explain behavior as the optimal solution to a computational problem in many areas of cognition, including memory, reasoning, generalization, and causal induction. While these models can be used to explore the assumptions people make in a particular task, the computation required to produce the optimal solution is often intractable and thus not a reasonable model of the computations performed by people. To make working with rational models practical, computer scientists have developed approximation algorithms with asymptotic convergence guarantees, such as Gibbs sampling and particle filtering. We propose to use these same algorithms to generate rational process models from rational models of cognition – making the assumption that cognition utilizes these statistical algorithms to approximate intractable rational models. In particular, we show that a particle filter approximation to the Rational Model of Categorization (RMC; Anderson, 1990) can reproduce human data, including more human-like order effects than are produced by the RMC.

The role of prior knowledge in human reconstructive memory

Mark Steyvers, UNIVERSITY OF CALIFORNIA, IRVINE

Prior knowledge and expectations about events are known to influence recall in human memory, but the specific interactions of memory and knowledge are unclear. We propose hierarchical Bayesian models of reconstructive memory in which prior knowledge is combined with noisy memory representations at multiple levels of abstraction. We present empirical evidence from studies where participants reconstruct the sizes of objects, recall objects in scenes and draw handwritten digits from memory. These studies demonstrate the hierarchical influences of prior knowledge and the beneficial effects of utilizing prior knowledge in recall.

Normative models of multiple interacting memory systems

Mate Lengyel, UNIVERSITY OF CAMBRIDGE

In this talk I will demonstrate how ideas from machine learning, namely unsupervised learning, reinforcement learning, and information theory, can be used to understand fundamental aspects of semantic, episodic, and working memory, respectively, and the interaction of these memory systems in particular. We developed a normative theory of learning about meaningful chunks in visual scenes, and of the way such statistically optimal representations on long-term memory should affect short-term retention of visual scenes in working memory. We also investigated why and how even such a seemingly optimal system might still be beaten by a much simpler episodic memory-based system when one considers the ultimate use of memories for decision making. Most of the work I will present also includes experimental data, collected by collaborators, that test key predictions of the theories.

Compositional Logic Learning

Alan Yuille, UCLA

The paper describes a new method for learning conditional probabilities from binary-valued labeled data. We represent the distributions in noisy-logical form (Yuille and Lu 2008) which is motivated by experiments in Cognitive Science and which offers an alternative to the sigmoid regression representation used (implicitly) by methods like AdaBoost. We specify algorithms for learning these distributions by composing them from elementary structures. Our experimental results show that we obtain experimental results which are slightly better than AdaBoost but which are of far simpler forms.

Reconciling reinforcement learning and risk sensitivity: a model-based fMRI study

Yael Niv, PRINCETON

Which of these would you prefer: getting \$10 with certainty or tossing a coin for a 50% chance to win \$20? Whatever your answer, you probably were not indifferent between these two options. In general, human choice behavior is influenced not only by the expected reward value of options, but also by their variance, with subjects differing in the degree to which they are risk-averse or risk-seeking. Traditional reinforcement learning (RL) models of action selection, however, rely on temporal difference methods that learn the mean value of an option, ignoring risk. These models have been strongly linked to learning via prediction errors conveyed by dopaminergic neurons, and to BOLD signals reflecting prediction errors in the nucleus accumbens. Here, in an fMRI study of decision making, we set forth to reconcile the behavioral results and computational theory by inquiring whether the neural implementation of RL is indeed risk-neutral or whether it shows sensitivity to risk. We used the neural signature of RL in the nucleus accumbens to compare between four qualitatively different computational models of how risk can influence decision making. Our results reveal that choice behavior is better accounted for by incorporating risk-sensitivity into reinforcement learning, and, furthermore, that the BOLD correlates of prediction error learning in the brain indeed reflect subjective risk-sensitivity.

Goal-directed decision making as structured probabilistic inference

Matthew Botvinick, PRINCETON UNIVERSITY

Within psychology and neuroscience, there is growing interest in the mechanisms underlying "goal-directed" decision making: the selection of actions based on 1) knowledge of action-outcome contingencies, and 2) knowledge of the incentive value associated with specific outcomes. In formulating theories of how humans and other animals accomplish this kind of decision making, it is natural to look to classical methods for solving Markov decision problems. However, some additional leverage may be gained by considering a more recent approach, which translates the dynamic programming task into a problem of structured probabilistic inference. I'll describe one version of this approach, involving recursive Bayesian inference within graphical models. The components of the underlying graphs align with a set of key functional anatomical systems, allowing the theory to make contact with cognitive neuroscientific data. The approach also gives rise to novel predictions concerning human choice behavior, some of which we have been testing through experimental work.

Reward bonuses for efficient, effective exploration

Michael Littman, RUTGERS UNIVERSITY

Children must strike a balance between taking the time to perfectly understand their environment and taking advantage of what they already know. Viewed mathematically, solving this exploration-exploitation dilemma is computationally difficult, even in the case in which the environment simply consists of two unknown values (so-called 'bandit' problems). Natural environments present an even more challenging problem because the number of possible events to consider learning about vastly outnumbers the opportunities to explore. In practice, children can never completely experience their world, but nonetheless need to understand it well enough to navigate, make predictions, and explain the events around them.

The exploration-exploitation dilemma has long been recognized in the engineering disciplines as a problem that learning systems must face. Recent work in computer science has highlighted the importance of retreating from perfect optimality and settling for 'good enough' solutions. This talk will survey some new developments in machine learning that introduce reward bonuses for insufficiently explored states and show that the resulting learning algorithms balance exploration and exploitation while remaining computationally tractable. They can also search hypotheses spaces, even given noisy experience, to find rules that allow them to make predictions in the absence of exhaustive experience. These computationally tractable solutions from the machine-learning community could provide insight on the potential limitations, constraints, and mechanisms that may shape children's exploration and understanding of the world.

Human Semi-Supervised Learning and Human Active Learning

Xiaojin Zhu, UNIVERSITY OF WISCONSIN-MADISON

We explore the connections between machine learning and human learning in two settings: semi-supervised learning and active learning. Both are well studied in statistical machine learning. In our experiments,

humans replace learning algorithms to assume the role of the learner in a category learning (i.e., classification) task. In semi-supervised learning, subjects are given additional unlabeled data. In active learning, subjects are given the ability to choose which items to query for label. Our results indicate that humans can perform semi-supervised learning and active learning. Quantitatively their performance also differs from learning theory predictions in interesting ways.

Where am I and what should I do next? Overcoming perceptual aliasing in sequential tasks

Todd Gureckis, NEW YORK UNIVERSITY

A critical challenge facing learners in a changing environment is correctly representing the current state of the world and appreciating how it may influence future outcomes. My talk considers recent work in my lab looking at issues of state representation and generalization in sequential decision making by humans. The experiments and models I describe are principally inspired by recent advances in machine learning which address how artificial agents may learn from experience in complex task domains. Overall, the goal of this work is to establish connections between this foundational computational work and issues of mental representation, categorization, stimulus generalization, and decision making traditionally studied in cognitive science/psychology.

Using reinforcement learning models to interpret human performance on Markov decision problems

Michael Mozer, UNIVERSITY OF COLORADO AT BOULDER

Theories of learning by reinforcement have been used to interpret data from individuals performing one-step choice tasks (e.g., the Iowa gambling task), and data from animals performing temporally extended behaviors, but not, to our knowledge, data from individuals performing sequential decision tasks. We tested participants in a temporally extended task that involved exploring an unfamiliar environment. The environment consisted of rooms, each containing two doors leading to other rooms. The participant's task was to select a sequence of doors to enter. Rewards were associated with state-action pairs. One question we address is whether formal theories of reinforcement learning (Q learning, Q policy gradient, and model based approaches) are suitable for characterizing the behavior of participants. We obtained a maximum likelihood fit of Q learning parameters to the pattern of choices made by individual participants. The parameters include: exploration strategy (epsilon-greedy versus normalized exponential), control of the exploration-exploitation trade off, the discounting rate (gamma), the backup parameter of the eligibility trace (lambda), and a learning rate. We report mixed results fitting participant data to the models. Beyond using reinforcement-learning models to fit data, the data has the potential to inform theories of reinforcement learning. These theories are neutral with regard to how the model parameters are set. Thus, a second question we address is: how do task variables and cognitive constraints modulate parameter settings? To explore this question, we performed experimental manipulations such as varying the time allotted for choosing an action, and varying a concurrent working-memory load. We find that these manipulations can be interpreted in terms of their influence on model parameters.

A Bayesian Algorithm for Change Detection with Identification: Rational Analysis and Human Performance

Jun Zhang, UNIVERSITY OF MICHIGAN-ANN ARBOR

We consider the problem of change detection along with identification in multi-hypotheses setting, where the state-of-world changes from H_0 to H_i ($i = 1, 2, \dots, N$) under known prior distributions. A Bayesian sequential updating equation is derived, along with the usual boundary-crossing stopping rule. The algorithm has the property that the value of an absorbing boundary, when overshoot is ignored, equals the hit rate of a decision-maker conditioned on that response. Computer simulation reveals that the algorithm shares many similarities with human performance in stimulus detection/identification experiments.

POSTERS

1. A psychophysical investigation of clustering. Joshua Lewis, UCSD
2. The Hierarchical Dirichlet Process as a model of Human Categorization. Kevin Canini, Berkeley

3. Kernels and Exemplar Models. Frank Jäkel, MIT
4. Learning Object-based Attention Control. Ali Borji, Majid N. Ahmadabadi and Babak N. Araabi, Institute for Studies in Theoretical Physics and Mathematics, Iran
5. A Hebbian Learning Rule for Optimal Decision Making. Michael Pfeiffer, Bernhard Nessler, and Wolfgang Maass, Graz University of Technology, Austria
6. Modeling Word Association Data using Multiple Maps. Laurens van der Maaten and Geoffrey Hinton, Tilburg University and University of Toronto
7. Integrating Statistics from the World and from Language to Learn Semantic Representations. Mark Andrews, Gabriella Vigliocco, and David P. Vinson, University College London
8. Bayesian modeling of intuitive pedagogical reasoning. Patrick Shafto and Noah Goodman, University of Louisville.
9. Learning from actions and their consequences: Inferring causal variables from continuous sequences of human action. Daphna Buchsbaum and Tom Griffiths, University of California Berkeley
10. Translation-invariant sparse deep belief networks for scalable unsupervised learning of hierarchical representation. Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y. Ng, Stanford University
11. Machine learning in the service of understanding human learning: an ideal observer-based analysis of the learning curve. Ferenc Huszar, Uta Noppeney, and Mate Lengyel. Budapest University of Technology and Economics, MPI Tuebingen, and University of Cambridge

DECEMBER 12, 2008, 07:30–10:30 AND 15:30–18:30

WESTIN: ALPINE CD **WS8**

Machine Learning Open Source Software

<http://mloss.org/workshop/nips08/>

Soeren Sonnenburg

FRAUNHOFER FIRST

Mikio Braun

TECHNISCHE UNIVERSITÄT BERLIN

Cheng Soon Ong

ETH ZÜRICH

Soeren.Sonnenburg@first.fraunhofer.de

mikio@cs.tu-berlin.de

chengsoon.ong@inf.ethz.ch

Abstract

We believe that the wide-spread adoption of open source software policies will have a tremendous impact on the field of machine learning. The goal of this workshop is to further support the current developments in this area and give new impulses to it. Following the success of the inaugural NIPS-MLOSS workshop held at NIPS 2006, the Journal of Machine Learning Research (JMLR) has started a new track for machine learning open source software initiated by the workshop's organizers. Many prominent machine learning researchers have co-authored a position paper advocating the need for open source software in machine learning. Furthermore, the workshop's organizers have set up a community website mloss.org where people can register their software projects, rate existing projects and initiate discussions about projects and related topics. This website currently lists 156 such projects including many prominent projects in the area of machine learning. The main goal of this workshop is to bring the main practitioners in the area of machine learning open source software together in order to initiate processes which will help to further improve the development of this area. In particular, we have to move beyond a mere collection of more or less unrelated software projects and provide a common foundation to stimulate cooperation and interoperability between different projects. An important step in this direction will be a common data exchange format such that different methods can exchange their results more easily.

7.30-7.45	Introduction THE ORGANIZERS
7.45-8.30	Octave JOHN W. EATON
8.30-8.50	Torch RONAN COLLOBERT, SAMY BENGIO, LEON BOTTOU, JASON WESTON, AND IAIN MELVIN
8.50-9.10	Shark CHRISTIAN IGEL, TOBIAS GLASMACHERS, AND VERENA HEIDRICH-MEISNER
9.10-9.30	kernlab ALEXANDROS KARATZOGLOU, ALEX SMOLA, AND KURT HORNIK
9.30-9.50	Machine Learning Py (mlpy) DAVIDE ALBANESE, STEFANO MERLER, ROBERTO VISINTAINER, CESARE FURLANELLO
9.50-10.00	MDP – Modular toolkit for Data Processing PIETRO BERKES, NIKO WILBERT, TIZIANO ZITO
10.00-10.30	Discussion: What is a good mloss project? • Data exchange standards

- review criteria for JMLR mloss
 - interoperable software
 - test suites
- 15.30-16.15** **matplotlib**
JOHN D. HUNTER
- 16.15-16.35** **Disco**
NOKIA RESEARCH CENTER AND THE DISCO OPEN-SOURCE PROJECT
- 16.35-16.55** **Nieme**
FRANCIS MAES
- 16.55-17.05** **libDAI**
JORIS MOOIJ
- 17.05-17.15** **BCPy2000**
JEREMY HILL, THOMAS SCHREINER, CHRISTIAN PUZICHA, AND JASON FAR-
QUHAR
- 17.15-17.25** **Model Monitor**
TROY RAEDER, AND NITESH CHAWLA
- 17.25-17.45** **RL Glue and Codecs Glue**
BRIAN TANNER, ADAM WHITE, AND RICHARD S. SUTTON
- 17.45-17.50** **Experiment Databases for Machine Learning**
JOAQUIN VANSCHOREN, AND HENDRIK BLOCKEEL
- 17.50-17.55** **BenchMarking Via Weka**
PETER REUTEMANN, AND GEOFF HOLMES
- 17.55-18.30** **Discussion: Reproducible research**
- Shall datasets be open too? How to provide access to data sets.
 - Reproducible research, next step beyond UCI datasets.

Octave (Invited Talk)

John W. Eaton, UNIVERSITY OF WISCONSIN

GNU Octave is a high-level language, primarily intended for numerical computations. It provides a convenient command line interface for solving linear and nonlinear problems numerically, and for performing other numerical experiments using a language that is mostly compatible with Matlab. It may also be used as a batch-oriented language.

Torch

Ronan Collobert, NEC LABORATORIES AMERICA

Torch provides a Matlab-like environment for state-of-the-art machine learning algorithms. It is easy to use and very efficient, thanks to a simple-yet-powerful fast scripting language (Lua), and a underlying C/C++ implementation. Torch is easily extensible and has been shown to scale to very large applications.

Shark

Tobias Glasmachers, RUHR-UNIVERSITÄT-BOCHUM, GERMANY

Shark is a C++ machine learning library. Tutorials and html documentation make Shark easy to learn. The installation of Shark is straightforward, it does not depend on any third party software and compiles under Linux, Solaris, MacOS, and Windows. Various example programs serve as starting points for own projects. Shark provides methods for linear and nonlinear optimization, in particular evolutionary and gradient-based algorithms. It comes with different types of artificial neural networks ranging from standard

feed-forward architectures to recurrent networks, with support vector machines relying on a competitive SMO implementation, and with various other machine learning techniques. The focus of the library is on algorithms. We feel that data visualization should not be hard coded into a machine learning library. However, graphical example programs using QT are available for analysis and demonstration purposes.

kernlab

Alexandros Karatzoglou, LITIS LAB, INSA DE ROUEN, FRANCE

kernlab is an R package providing kernel-based machine learning functionality. It is designed to provide tools for kernel algorithm development but also includes a range of popular machine learning methods for classification, regression, clustering, novelty detection, quantile regression and dimensionality reduction. Among other algorithms included in the package are Support Vector Machines, Spectral Clustering, Kernel PCA, a QP solver and a range of kernels (Gaussian, Laplacian, string kernels etc.).

Machine Learning Py (mlpy)

Giuseppe Jurman, FBK-MPBA, TRENTO, ITALY

We introduce mlpy, a high-performance Python package for predictive modeling. It makes extensive use of NumPy to provide fast N-dimensional array manipulation and easy integration of C code. Mlpy provides high level procedures that support, with few lines of code, the design of rich Data Analysis Protocols (DAPs) for predictive classification and feature selection. Methods are available for feature weighting and ranking, data resampling, error evaluation and experiment landscaping. The package includes tools to measure stability in sets of ranked feature lists, of special interest in bioinformatics for functional genomics, for which large scale experiments with up to 106 classifiers have been run on Linux clusters and on the Grid.

MDP – Modular toolkit for Data Processing

Tiziano Zito, BERNSTEIN CENTER FOR COMPUTATIONAL NEUROSCIENCE, BERLIN, GERMANY

Modular toolkit for Data Processing (MDP) is a Python data processing framework. From the user's perspective, MDP is a collection of supervised and unsupervised learning algorithms and other data processing units that can be combined into data processing sequences and more complex feed-forward network architectures. From the scientific developer's perspective, MDP is a modular framework, which can easily be expanded. The implementation of new algorithms is easy and intuitive. The new implemented units are then automatically integrated with the rest of the library. The base of available algorithms is steadily increasing and includes, to name but the most common, Principal Component Analysis (PCA and NIPALS), several Independent Component Analysis algorithms (CuBICA, FastICA, TDSEP, and JADE), Slow Feature Analysis, Gaussian Classifiers, Restricted Boltzmann Machine, and Locally Linear Embedding.

matplotlib (Invited Talk)

John D. Hunter, TRADELINK, QUANTITATIVE STRATEGIES

matplotlib is a python 2D plotting library which produces publication quality figures in a variety of hardcopy formats and interactive environments across platforms. *matplotlib* can be used in python scripts, the python and ipython shell (ala matlab or mathematica), web application servers, and works with six graphical user interface toolkits. *matplotlib* tries to make easy things easy and hard things possible. You can generate plots, histograms, power spectra, bar charts, error charts, scatter plots, etc, with just a few lines of code. For the power user, you have full control of line styles, font properties, axes properties, etc, via an object oriented interface or via a handle graphics interface familiar to matlab users.

Disco

Ville Tuulos, NOKIA RESEARCH CENTER, PALO ALTO

Disco is an open-source implementation of the Map-Reduce framework for distributed computing. As the original framework, Disco supports parallel computations over large data sets on unreliable cluster of computers. You don't need a cluster to use Disco – a script is provided that installs Disco automatically to the Amazon's EC2 computing cloud where you get computing resources on demand basis.

Nieme

Francis Maes, UNIVERSITY OF PARIS 6, FRANCE

Nieme is a machine learning library for large-scale classification, regression and ranking. It relies on the

framework of energy-based models which unifies several learning algorithms ranging from simple perceptrons to recent models such as the Pegasos support vector machine or L1-regularized maximum entropy models. This framework also unifies batch and stochastic learning which are both seen as energy minimization problems. Nieme can hence be used in a wide range of situations, but is particularly interesting for very-large-scale learning tasks where both the examples and the features are processed incrementally. Being able to deal with new incoming features at any time within the learning process is another key feature of the Nieme toolbox. Nieme is released under the GPL license. It is efficiently implemented in C++ and works on Linux, MacOS and Windows. Interfaces are available for C++, Java and Python.

libDAI

Joris Mooij, MAX-PLANCK-INSTITUTE FOR BIOLOGICAL CYBERNETICS, TÜBINGEN, GERMANY

libDAI is a free and open source C++ library (licensed under GPL) that provides implementations of various (approximate) inference methods for discrete graphical models. libDAI supports arbitrary factor graphs with discrete variables; this includes discrete Markov Random Fields and Bayesian Networks. The library is targeted at researchers; to be able to use the library, a good understanding of graphical models is needed. Currently, libDAI supports the following (approximate) inference methods: exact inference by brute force enumeration, exact inference by junction-tree methods, Mean Field, Loopy Belief Propagation, Tree Expectation Propagation, Generalized Belief Propagation, Double-loop GBP, and various variants of Loop Corrected Belief Propagation. Planned extensions are Gibbs sampling and IJGP, as well as various methods for obtaining bounds on the partition sum and on marginals (Bound Propagation, Box Propagation, Tree-based Reparameterization).

BCPy2000

Jeremy Hill, MAX-PLANCK-INSTITUTE FOR BIOLOGICAL CYBERNETICS, TÜBINGEN, GERMANY

BCPy2000 provides a platform for rapid, flexible development of experimental Brain-Computer Interface systems based on the BCI2000.org project. From the developer’s point of view, the implementation is carried out in Python, taking advantage of various high-level packages: VisionEgg for stimulus presentation, NumPy and SciPy for signal processing and classification, and IPython for interactive debugging. BCPy2000 implements a lot of infrastructure allowing you to get new experiments up and running quickly. It also contains a set of optional tools, which are still a work in progress but which are rapidly turning into a kind of “standard library” of object-oriented signal-processing and stimulus widgets. These features make it a flexible platform for developers of new NumPy/SciPy-based machine-learning algorithms in the field of realtime biosignal analysis.

Model Monitor

Troy Raeder, UNIVERSITY OF NOTRE DAME

Common practice in Machine Learning often implicitly assumes a stationary distribution, meaning that the distribution of a particular feature does not change over time. In practice, however, this assumption is often violated and real-world models have to be retrained as a result. It would be helpful, then, to be able to anticipate and plan for changes in distribution in order to avoid this retraining. Model Monitor is a Java toolkit that addresses this problem. It provides methods for detecting distribution shifts in data, comparing the performance of multiple classifiers under shifts in distribution, and evaluating the robustness of individual classifiers to distribution change. As such, it allows users to determine the best model (or models) for their data under a number of potential scenarios. Additionally, Model Monitor is fully integrated with the WEKA machine learning environment, so that a variety of commodity classifiers can be used if desired.

RL Glue and Codecs Glue

Brian Tanner, UNIVERSITY OF ALBERTA

RL-Glue is a protocol and software implementation for evaluating reinforcement learning algorithms. Our system facilitates the comparison of alternative algorithms and can greatly accelerate research progress as the UCI database has accelerated progress in supervised machine learning. Creating a comparable benchmarking resource for reinforcement learning is challenging because of the temporal nature of reinforcement learning. Reinforcement learning agents interact with a dynamic process (the environment) which generates observations and rewards. The observations and rewards received by the learning agent depend on the

actions; training data cannot simply be stored in a file as they are in supervised learning. Instead, the reinforcement learning agent and environment must be interacting programs. RL-Glue agents and environments can be written in Java, C/C++, Matlab, Python, and Lisp and can all run on one machine, or can connect across the Internet. In this seminar, we will introduce the design principles that helped shape RL-Glue and demonstrate some of the interesting extensions that have been created by the reinforcement learning community.

Experiment Databases for Machine Learning

Joaquin Vanschoren, UNIVERSITY OF LEUVEN, BELGIUM

Experiment Databases for Machine Learning is a large public repository of machine learning experiments as well as a framework for producing similar databases for specific goals. This project aims to bring the information contained in many machine learning experiments together and organize it a way that allows everyone to investigate how learning algorithms have performed in previous studies. To share such information with the world, a common language is proposed, dubbed ExpML, capturing the basic structure of a large range of machine learning experiments while remaining open for future extensions. This language also enforces reproducibility by requiring links to the used datasets and algorithms and by storing all details of the experiment setup. All stored information can then be accessed by querying the database, creating a powerful way to collect and reorganize the data, thus warranting a very thorough examination of the stored results. The current publicly available database contains over 500,000 classification and regression experiments, and has both an online interface, at <http://expdb.cs.kuleuven.be>, as well as a stand-alone explorer tool offering various visualization techniques. This framework can also be integrated in machine learning toolboxes to automatically stream results to a global (or local) experiment database, or to download experiments that have been run before.

BenchMarking Via Weka

Peter Reutemann, UNIVERSITY OF WAIKATO, NEW ZEALAND

BenchMarking Via Weka is a client-server architecture that supports interoperability between different machine learning systems. Machine learning systems need to provide mechanisms for processing data and evaluating generated models. In our system, the server hosts all the data and performs all the statistical analyses, while the client performs all the pre-processing and model building. This separation of tasks opens up the possibility of offering a cross-platform and cross-language framework. By performing statistical analyses on the host, we avoid unnecessary exchange and conversion of generated results.

Optimization for Machine Learning

<http://opt2008.kyb.tuebingen.mpg.de/>

Suvrit Sra

MAX PLANCK INSTITUTE TÜBINGEN

Sebastian Nowozin

MAX PLANCK INSTITUTE TÜBINGEN

S V N Vishwanathan

PURDUE UNIVERSITY

suvrit@tuebingen.mpg.de

sebastian.nowozin@tuebingen.mpg.de

vishy@stat.purdue.edu

Abstract

Classical optimization techniques have found widespread use in machine learning. Convex optimization has occupied the center-stage and significant effort continues to be still devoted to it. New problems constantly emerge in machine learning, e.g., structured learning and semi-supervised learning, while at the same time fundamental problems such as clustering and classification continue to be better understood. Moreover, machine learning is now very important for real-world problems with massive datasets, streaming inputs, the need for distributed computation, and complex models. These challenging characteristics of modern problems and datasets indicate that we must go beyond the traditional optimization approaches common in machine learning. What is needed is optimization tuned for machine learning tasks. For example, techniques such as non-convex optimization (for semi-supervised learning, sparsity constraints), combinatorial optimization and relaxations (structured learning), stochastic optimization (massive datasets), decomposition techniques (parallel and distributed computation), and online learning (streaming inputs) are relevant in this setting. These techniques naturally draw inspiration from other fields, such as operations research, polyhedral combinatorics, theoretical computer science, and the optimization community.

7.30-7.35	Opening remarks ORGANIZERS
7.35-8.25	Invited talk 1. Optimization in Machine Learning: Recent Developments and Current Challenges STEPHEN WRIGHT
8.25-8.35	Coffee break
8.35-8.55	Talk 1, Online and Batch Learning Using Forward-Looking Subgradients JOHN DUCHI, UC BERKELEY
8.55-9.15	Talk 2, Robustness and Regularization of Support Vector Machines HUAN XU, MCGILL
9.15-9.35	Talk 3, An Improved Branch-and-Bound Method for Maximum Monomial Agreement NOAM GOLDBERG, RUTCOR, RUTGERS UNIVERSITY
9.35-9.40	Break
9.40-10.30	Invited talk 2. Polyhedral Approximations in Convex Optimization DIMITRI BERTSEKAS
10.30-3.30	Break

3.30-4.20	Invited talk 3. Large-scale Machine Learning and Stochastic Algorithms LEON BOTTOU
4.20-4.40	Talk 4, Training a Binary Classifier with the Quantum Adiabatic Algorithm HARTMUT NEVEN, GOOGLE
4.40-4.50	Coffee break
4.50-5.10	Talk 5, Optimization on a Budget: A Reinforcement Learning Approach PAUL RUVOLO, UCSD
5.10-5.30	Talk 6, Online Optimization in X-Armed Bandits SEBASTIEN BUBECK, INRIA LILLE
5.30-6.10	Panel discussion
6.10 Onwards	Poster presentations

Optimization in Machine Learning: Recent Developments and Current Challenges

Stephen Wright, UNIVERSITY OF WISCONSIN MADISON

The use of optimization as a framework for formulating machine learning problems has become much more widespread in recent years. In some cases, the demands of the machine learning problems go beyond the scope of traditional optimization paradigms. While existing optimization formulations and algorithms serve as a good starting point for the solution strategies, important work must be carried out at the interface of optimization and machine learning to devise strategies that exploit the special features of the application and that perform well on very large data sets. This talk reviews recent developments from an optimization perspective, focusing on activity during the past three years, and looking in particular at problems where the machine learning application has motivated novel algorithms or analysis in the optimization domain. We also discuss some current challenges, highlighting several recent developments in optimization that may be useful in machine learning applications.

Polyhedral Approximations in Convex Optimization

Dimitri Bertsekas, MIT

We propose a unifying framework for solution of convex programs by polyhedral approximation. It includes classical methods, such as cutting plane, Dantzig-Wolfe decomposition, bundle, and simplicial decomposition, but also includes refinements of these methods, as well as new methods that are well-suited for important large-scale types of problems, arising for example in network optimization.

Large-scale Machine Learning and Stochastic Algorithms

Leon Bottou, NEC LABORATORIES AMERICA

TBA

Structured Input - Structured Output

<http://agbs.kyb.tuebingen.mpg.de/wikis/bg/iso2008/FrontPage>

Karsten Borgwardt

UNIVERSITY OF CAMBRIDGE, MPIS DEVELOPMENTAL BIOLOGY AND BIOLOGICAL CYBERNETICS

kmb51@cam.ac.uk

Koji Tsuda

MPI BIOLOGICAL CYBERNETICS

koji.tsuda@tuebingen.mpg.de

S V N Vishwanathan

PURDUE UNIVERSITY

vishy@purdue.edu

Xifeng Yan

IBM T. J. WATSON RESEARCH CENTER

xifengyan@us.ibm.com

Abstract

Structured data emerges rapidly in a large number of disciplines: bioinformatics, systems biology, social network analysis, natural language processing and the Internet generate large collections of strings, graphs, trees, and time series. Designing and analysing algorithms for dealing with these large collections of structured data has turned into a major focus of machine learning over recent years, both in the input and output domain of machine learning algorithms, and is starting to enable exciting new applications of machine learning. The goal of this workshop is to bring together experts on learning with structured input and structured output domains and its applications, in order to exchange the latest developments in these growing fields. The workshop will feature keynotes by Prof. Eric Xing from Carnegie Mellon University and by Dr Yasemin Altun from the MPI for Biological Cybernetics.

7.30-7.35	Welcome and Introduction of Keynote Speaker KARSTEN BORGWARDT
7.35-8.30	1st Keynote Speech ERIC XING
8.30-8.50	Coffee Break
8.50-9.15	Integrating Ontological Prior Knowledge into Relational Learning STEFAN RECKOW, VOLKER TRESP
9.15-9.40	Relation-Prediction in Multi-Relational Domains using Matrix-Factorization CHRISTOPH LIPPERT, STEFAN-HAGEN WEBER, YI HUANG, VOLKER TRESP, MATTHIAS SCHUBERT, HANS-PETER KRIEDEL
9.40-10.05	Logistic Regression for Graph Classification NINO SHERVASHIDZE, KOJI TSUDA
10.05-10.30	Graphical Multi-Task Learning DANIEL SHELDON
10.30-15.30	Snow Break
15.30-16.30	2nd Keynote Speech YASEMIN ALTUN
16.30-16.50	Coffee Break

- 16.50-17.15** **Learning to Predict Combinatorial Structures**
 THOMAS GAERTNER, SHANKAR VEMBU
- 17.15-17.40** **Joint Kernel Support Estimation for Structured Prediction**
 CHRISTOPH LAMPERT, MATTHEW BLASCHKO
- 17.40-18.05** **Learning Optimal Subsets with Implicit User Preferences**
 YUNSONG GUO, CARLA GOMES
- 18.05-18.30** **Learning Structural SVMs with Latent Variables**
 CHUN-NA YU, THORSTEN JOACHIMS

Integrating Ontological Prior Knowledge into Relational Learning

Stefan Reckow, MPI PSYCHIATRY

Volker Tresp, SIEMENS AG

Ontologies represent an important source of prior information which lends itself to the integration into statistical modeling. This paper discusses approaches towards employing ontological knowledge for relational learning. Our analysis is based on the IHRM model that performs relational learning by including latent variables that can be interpreted as cluster variables of the entities in the domain. We apply our approach to the modeling of yeast genomic data and demonstrate that the inclusion of ontologies as prior knowledge in relational learning can lead to significantly improved results and to better interpretable clustering structures.

Relation-Prediction in Multi-Relational Domains using Matrix-Factorization

Christoph Lippert, MPIS DEVELOPMENTAL BIOLOGY AND BIOLOGICAL CYBERNETICS

Stefan-Hagen Weber, SIEMENS AG

Yi Huang, SIEMENS AG

Volker Tresp, SIEMENS AG

Matthias Schubert, LMU MUENCHEN

Hans-Peter Kriegel, LMU MUENCHEN

The paper is concerned with relation prediction in multi-relational domains using matrix factorization. While most past predictive models focussed on one single relation type between two entity types, in the paper a generalized model is presented that is able to deal with an arbitrary number of relation types and entity types in a domain of interest. The novel multi-relational matrix factorization is domain independent and highly scalable. We validate the performance of our approach using two real-world data sets, i.e. user-movie recommendations and gene function prediction.

Logistic Regression for Graph Classification

Nino Shervashidze, MPIS DEVELOPMENTAL BIOLOGY AND BIOLOGICAL CYBERNETICS

Koji Tsuda, MPI BIOLOGICAL CYBERNETICS

In this paper we deal with graph classification. We propose a new algorithm for performing sparse logistic regression for graphs, which is comparable in accuracy with other methods of graph classification and produces probabilistic output in addition. Sparsity is required for the reason of interpretability, which is often necessary in domains such as bioinformatics or chemoinformatics.

Graphical Multi-Task Learning

Daniel Sheldon, CORNELL UNIVERSITY

We investigate the problem of learning multiple tasks that are related according to a network structure, using the multi-task kernel framework proposed in (Evgeniou et al., 2006). Our method combines a graphical task kernel with an arbitrary base kernel. We demonstrate its effectiveness on a real ecological application that inspired this work.

Learning to Predict Combinatorial Structures

Thomas Gaertner, FRAUNHOFER IAIS

Shankar Vembu, FRAUNHOFER IAIS

We consider the problem of predicting combinatorial structures such as directed cycles, partially ordered sets, and other graph classes. Assumptions made by existing structured prediction algorithms preclude their applicability to this problem. We present an algorithm that overcomes these limitations.

Joint Kernel Support Estimation for Structured Prediction

Christoph Lampert, MPI BIOLOGICAL CYBERNETICS

Matthew Blaschko, MPI BIOLOGICAL CYBERNETICS

We present a new technique for structured prediction that works in a hybrid generative/discriminative way, using a one-class support vector machine to model the joint probability of (input, output)-pairs in a joint reproducing kernel Hilbert space. Compared to discriminative techniques, like conditional random fields or structured output SVMs, the proposed method has the advantage that its training time depends only on the number of training examples, not on the size of the label space. Due to its generative aspect, it is also very tolerant against ambiguous, incomplete or incorrect labels. Experiments on realistic data show that our method works efficiently and robustly in situations that discriminative techniques have problems with or that are computationally infeasible for them.

Learning Optimal Subsets with Implicit User Preferences

Yunsong Guo, CORNELL UNIVERSITY

Carla Gomes, CORNELL UNIVERSITY

In this paper we study the problem of learning an optimal subset from a larger ground set of items, where the optimality criterion is defined by an unknown preference function. We model the problem as a discriminative structural learning problem and solve it using a Structural Support Vector Machine (SSVM) that optimizes a “set accuracy” performance measure representing set similarities. Our approach departs from previous approaches since we do not explicitly learn a pre-defined preference function. Experimental results on both a synthetic block selection problem and a real-world face image subset selection problem show that our method significantly outperforms previous approaches.

Learning Structural SVMs with Latent Variables

Chun-Na Yu, CORNELL UNIVERSITY

Thorsten Joachims, CORNELL UNIVERSITY

It is well known in statistics and machine learning that the combination of latent variables and observed variables offer more expressive power than models with observed variables alone. Structural SVMs have excellent performance in many structured prediction tasks, and yet currently they do not support the use of latent variables. In this work we extend Structural SVMs to include latent variables, and provide an efficient algorithm for solving the optimization problem of our proposed formulation. We apply our new algorithm to the problem of discriminative motif finding in yeast DNA and some initial results are presented.

DECEMBER 12, 2008, 07:30–10:30 AND 16:00–19:00

WESTIN: NORDIC WS11

Causality: objectives and assessment

<http://clopinet.com/isabelle/Projects/NIPS2008>**Isabelle Guyon**

CLOPINET

isabelle@clopinet.com

Dominik Janzing

MPI FOR BIOLOGICAL CYBERNETICS, TUEBINGEN

dominik.janzing@tuebingen.mpg.de

Bernhard Schölkopf

MPI FOR BIOLOGICAL CYBERNETICS TUEBINGEN

bernhard.schoelkopf@tuebingen.mpg.de

Abstract

Machine learning has traditionally been focused on prediction. Given observations that have been generated by an unknown stochastic dependency, the goal is to infer a law that will be able to correctly predict future observations generated by the same dependency. Statistics, in contrast, has traditionally focused on data modeling, i.e., on the estimation of a probability law that has generated the data. During recent years, the boundaries between the two disciplines have become blurred and both communities have adopted methods from the other, however, it is probably fair to say that neither of them has yet fully embraced the field of causal modeling, i.e., the detection of causal structure underlying the data. Since the Eighties there has been a community of researchers, mostly from statistics and philosophy, who have developed methods aiming at inferring causal relationships from observational data. While this community has remained relatively small, it has recently been complemented by a number of researchers from machine learning. The goal of this workshop is to discuss new approaches to causal discovery from empirical data, their applications and methods to evaluate their success. Emphasis will be put on the definition of objectives to be reached and assessment methods to evaluate proposed solutions. The participants are encouraged to participate in a competition pot-luck in which datasets and problems will be exchanged and solutions proposed.

Morning Session. Chair: Bernhard Schölkopf

7.30 - 8.00	Welcome and program presentation, short overview over the posters DOMINIK JANZING
8.00 - 9.00	Tutorial / overview: Causal Inference as Computational Learning JUDEA PEARL, UCLA
9.00 - 9.15	Competition Results ISABELLE GUYON, FOR THE CAUSALITY WORKBENCH TEAM
9.15 - 9.30	Benchmarks, wikis, and open-source causal discovery PATRIK HOYER, UNIVERSITY OF HELSINKI
9.30 - 10.30	Poster viewing, coffee, informal discussions

Afternoon Session. Chair: Dominik Janzing

- 4.00 - 4.30** **Keynote talk: Causal Structure Search: Philosophical Foundations and Future Problems**
RICHARD SCHEINES AND PETER SPIRITES
- 4.30 - 4.45** **Contributed talk: Best results in the pot-luck challenge.**
- 4.45 - 5.00** **Contributed talk: Best proposed task in the pot-luck challenge.**
- 5.00 - 5.15** **Causal models as conditional density models**
KEVIN MURPHY, UNIVERSITY OF BRITISH COLUMBIA
- 5.15 - 5.30** **Analysis of the binary instrumental variable model**
THOMAS RICHARDSON , UNIVERSITY OF WASHINGTON
- 5.30 - 5.45** **Beware of the DAG!**
PHIL DAWID, UNIVERSITY OF CAMBRIDGE
- 6.00 - 7.00** **Plenary discussion**
ISABELLE GUYON, MODERATOR

DECEMBER 12, 2008, 07:30–10:30 AND 16:00–19:00

WESTIN: ALPINE E WS12

Cost Sensitive Learning

http://www.cs.iastate.edu/~oksayakh/csl/cslworkshop_nips2008.html

Balaji Krishnapuram

SIEMENS MEDICAL SOLUTIONS

balaji.krishnapuram@siemens.com

Shipeng Yu

SIEMENS MEDICAL SOLUTIONS

shipeng.yu@siemens.com

Oksana Yakhnenko

IOWA STATE UNIVERSITY

oksayakh@cs.iastate.edu

Bharat Rao

SIEMENS MEDICAL SOLUTIONS

bharat.rao@siemens.com

Lawrence Carin

DUKE UNIVERSITY

lcarin@ee.duke.edu

Abstract

Cost-sensitive learning aims to minimize the data acquisition cost while maximizing the accuracy of the learner/predictor. Many sub-fields in machine learning such as semi-supervised learning, active label/feature acquisition, cascaded classification, and inductive transfer are motivated by the need to minimize the cost of data acquisition in various application domains. These approaches typically attempt to minimize data acquisition costs under strong simplifying assumptions – e.g., features vectors are assumed to have zero cost in semi-supervised learning. Although all of these areas have felt the need for a principled solution to minimize data costs, until recently the acquisition cost has rarely been modeled directly. Despite some recent work in this area, much more research is needed on this important topic. It is also important to ensure that the theoretical work addresses the practical needs of several application communities such as computer aided medical diagnosis, signal processing, remote sensing, computer vision, etc. We hope to bring together researchers from semi-supervised learning, active label/feature acquisition, inductive transfer learning, cascaded classification and other theoretical areas with practitioners from various application domains. We welcome both novel theory/algorithms and contributions that draw attention to open problems and challenges in real-world applications which call for cost-sensitive learning.

7.30-8.15	Workshop Introduction SHIPENG YU
8.15-8.35	Contributed Talk 1 BURR SETTLES
8.35-8.55	Contributed Talk 2 ADRIANA BIRLUTIU
8.55-9.05	Coffee break and poster session
9.05-9.40	Invited Talk LAWRENCE CARIN
9.40-10.00	Contributed Talk 3 ALEXANDER LIU
10.00-10.30	Panel discussion
4.00-4.35	Invited Talk VOLKER TRESP

4.35-4.55	Contributed Talk 4 PREM MELVILLE
4.55-5.05	Poster Spotlight
5.05-5.15	Coffee break and poster session
5.15-5.50	Invited Talk JOHN SHAWE-TAYLOR
5.50-6.10	Contributed Talk 5 ROBBIE HAERTEL
6.10-6.30	Contributed Talk 6 JASON EISNER
6.30-7.00	Panel discussion and Closing Remarks

INVITED TALKS

With a little help from some friendly models

Volker Tresp, UNIVERSITY OF MUNICH

Inductive transfer learning, hierarchical modeling and multitask learning are among the statistical approaches that enable a sharing of strengths between models trained for different but related tasks. Particular flexibility is achieved by using nonparametric models, i.e., Gaussian processes and Dirichlet processes. In our presentation we relate the various approaches to each other and demonstrate their effectiveness in different applications.

Cost-Sensitive Feature Acquisition and Classification

Lawrence Carin, DUKE UNIVERSITY

There are many sensing challenges for which one must balance the effectiveness of a given measurement with the associated sensing cost. For example, when performing a diagnosis a doctor must balance the cost and benefit of a given test (measurement), and the decision to stop sensing (stop performing tests) must account for the risk to the patient and doctor (malpractice) for a given diagnosis based on observed data. This motivates a cost-sensitive classification problem in which the features (sensing results) are not given *a priori*; the algorithm determines which features to acquire next, as well as when to stop sensing and make a classification decision based on previous observations (accounting for the costs of various types of errors, as well as the rewards of being correct). We formally define the cost-sensitive classification problem and solve it via a partially observable Markov decision process (POMDP). While the POMDP constitutes an intuitively appealing formulation, the intrinsic properties of classification tasks resist application of it to this problem. We circumvent the difficulties of the POMDP via a myopic approach, with an adaptive stopping criterion linked to the standard POMDP. The myopic algorithm is computationally feasible, easily handles continuous features, and seamlessly avoids repeated actions. Experiments with several benchmark datasets show that the proposed method yields state-of-the-art performance, and importantly our method uses only a small fraction of the features that are generally used in competitive approaches.

PAC-Bayes Analysis of Stochastic Differential Equation Modeling: Reducing Learning Costs by Incorporating Complex Prior Knowledge

John Shawe-Taylor, UNIVERSITY COLLEGE LONDON

When labeling training data is costly, a natural approach to reducing the amount required is to incorporate prior knowledge into the learning. We consider the case where prior probability over possible time series is encoded in the form of a non-linear stochastic differential equation. Examples of applications are climate modeling, weather prediction and cell pathway dynamics. Bayesian inference in this prior can be approximated efficiently using variational approximation, but the quality of the approximation is difficult to assess.

We present an extension of PAC-Bayes analysis that enables us to lower bound the quality of generalisation of the inferred model.

CONTRIBUTED TALKS

Active Learning with Real Annotation Costs

Burr Settles, UNIVERSITY WISCONSIN, MADISON

Mark Craven, UNIVERSITY WISCONSIN, MADISON

Lewis Friedland, UNIVERSITY WISCONSIN, MADISON

The goal of active learning is to minimize the cost of training an accurate model by allowing the learner to choose which instances are labeled for training. However, most research in active learning to date has assumed that the cost of acquiring labels is the same for all instances. In domains where labeling costs may vary, a reduction in the number of labeled instances does not guarantee a reduction in cost. To better understand the nature of actual labeling costs in such domains, we present a detailed empirical study of active learning with annotation costs in four real-world domains involving human annotators.

Optimal experimental design in a hierarchical setting for probabilistic choice models

Adriana Birlutiu, RADBOUD UNIVERSITY NIJMEGEN

Tom Heskes, RADBOUD UNIVERSITY NIJMEGEN

We propose a new criterion for experimental design in the context of preference learning. This new criterion makes direct use of the data available from a group of subjects for which the preferences were already learned. Furthermore, we show the connections between this criterion and the standard criteria used in experimental design. Empirical results on a real audiological data set, show a factor of two speed-up for learning user preferences relative to random selection.

Active Learning with Spatially Sensitive Labeling Costs

Alexander Liu, UNIVERSITY OF TEXAS

Goo Jun, UNIVERSITY OF TEXAS

Joydeeo Ghosh, UNIVERSITY OF TEXAS

In active learning, it is typically assumed that all instances require the same amount of effort to label and that the cost of labeling an instance is independent of other selected instances. In spatially distributed data such as hyperspectral imagery for land-cover classification, the act of labeling a point (i.e., determining the land-type) may involve physically traveling to a location and determining ground truth. In this case, both assumptions about label acquisition costs made by traditional active learning are broken, since costs will depend on physical locations and accessibility of all the visited points as well as the order of visitations. This paper formulates and analyzes the novel problem of performing active learning on spatial data where label acquisition costs are proportional to distance traveled.

Prediction-time Active Feature-Value Acquisition for Cost-Effective Customer Targeting

Pallika Kanani, UNIVERSITY OF MASSACHUSETTS, AMHERST

Prem Melville, IBM T.J. WATSON RESEARCH CENTER

In general, the prediction capability of classification models can be enhanced by acquiring additional relevant features for instances. However, in many cases, there is a significant cost associated with this additional information — driving the need for an intelligent acquisition strategy. Motivated by real-world customer targeting domains, we consider the setting where a fixed set of additional features can be acquired for a subset of the instances at test time. We study different acquisition strategies of selecting instances for which to acquire more information, so as to obtain the most improvement in prediction performance per unit cost. We apply our methods to various targeting datasets and show that we can achieve a better prediction performance by actively acquiring features for only a small subset of instances, compared to a random-sampling baseline.

Return on Investment for Active Learning

Robbie Haertel, BRIGHAM YOUNG UNIVERSITY

Kevin D. Seppi, BRIGHAM YOUNG UNIVERSITY

Eric K. Ringger, BRIGHAM YOUNG UNIVERSITY

James L. Carroll, BRIGHAM YOUNG UNIVERSITY

Active Learning (AL) can be defined as a selectively supervised learning protocol intended to present those data to an oracle for labeling which will be most enlightening for machine learning. While AL traditionally accounts for the value of the information obtained, it often ignores the cost of obtaining the information thus causing it to perform sub-optimally with respect to total cost. We present a framework for AL that accounts for this cost and discuss optimality and tractability in this framework. Using this framework we motivate Return On Investment (ROI), a new, practical, cost-sensitive heuristic that can be used to convert existing algorithms into cost-conscious active learners. We demonstrate the validity of ROI in a simulated AL part-of-speech tagging task on the Penn Treebank in which ROI achieves as high as a 73% reduction in hourly cost over random selection.

Machine Learning with Annotator Rationales to Reduce Annotation Cost

Omar Zaidan, JOHNS HOPKINS UNIVERSITY

Jason Eisner, JOHNS HOPKINS UNIVERSITY

Christine D. Piatko, JOHNS HOPKINS UNIVERSITY

We review two novel methods for text categorization, based on a new framework that utilizes richer annotations that we call annotator rationales. A human annotator provides hints to a machine learner by highlighting contextual "rationales" in support of each of his or her annotations. We have created a dataset with substring rationales from an existing sentiment classification dataset (Pang and Lee, 2004). We describe here two methods, one discriminative (Zaidan et al., 2007) and one generative (Zaidan and Eisner, 2008), that use these rationales during training to obtain significant accuracy improvements over two strong baselines. Our generative model in particular could be adapted to help learn other kinds of probabilistic classifiers for quite different tasks. Based on a small study of annotation speed, we posit that for some tasks, providing rationales can be a more fruitful use of an annotator's time than annotating more examples.

POSTER SESSION

Empirical Evaluation of Support Vector Machine Active Learning Algorithms for Imbalanced Data Sets

Michael Bloodgood, UNIVERSITY OF DELAWARE

K. Vijay-Shanker, UNIVERSITY OF DELAWARE

First we consider the importance of addressing class imbalance during Active Learning (AL) with SVMs (AL-SVM). AL brings out the need to modify passive learning approaches and our InitPA method outperforms previous methods for addressing imbalance during AL. Three leading selection strategies that can be used in conjunction with InitPA are evaluated on multiple datasets for relation extraction and text classification. Margin-based selection with InitPA for handling imbalance is shown to outperform the other approaches in a variety of ways and possible explanations for this are proposed.

Explicit Utility in Supervised Learning

James Carroll, BRIGHAM YOUNG UNIVERSITY

Neil Toronto, BRIGHAM YOUNG UNIVERSITY

Kevin Seppi, BRIGHAM YOUNG UNIVERSITY

Robbie Haertel, BRIGHAM YOUNG UNIVERSITY

We use a graphical model of the supervised learning problem together with the principles of decision theory to explore the theoretical effect of utility in the form of end use and sample cost on supervised learning, no free lunch, sample complexity, and active learning.

Cost-sensitive learning based on Bregman divergences

Jesús Cid-Sueiro, UNIVERSIDAD CARLOS III DE MADRID, SPAIN

Rocío Alaiz-Rodríguez, UNIVERSIDAD DE LEÓN, SPAIN

Alicia Guerrero-Curieses, UNIVERSIDAD REY JUAN CARLOS, MADRID, SPAIN

This paper analyzes the application of a particular class of Bregman divergences to design cost-sensitive classifiers for multiclass problems. We show that these divergence measures can be used to estimate posterior probabilities with maximal accuracy for the probability values that are close to the boundary decisions.

Asymptotically, the proposed divergence measures provide classifiers minimizing the sum of decision costs in non-separable problems, and maximizing a generalized margin in separable problems.

New Challenges in Theoretical Machine Learning: Learning with Data-dependent Concept Spaces

<http://www.kuleuven.be/wehys/>

Maria-Florina Balcan

MICROSOFT RESEARCH

Shai Ben-David

UNIVERSITY OF WATERLOO

Avrim Blum

CARNEGIE MELLON UNIVERSITY

Kristiaan Pelckmans

K.U.LEUVEN

John Shawe-Taylor

UNIVERSITY COLLEGE LONDON

ninamf@cs.cmu.edu

shai@cs.uwaterloo.ca

avrim@cs.cmu.edu

Kristiaan.Pelckmans@esat.kuleuven.be

jst@cs.ucl.ac.uk

Abstract

This workshop aims at collecting theoretical insights in the design of data-dependent learning strategies. Specifically we are interested in how far learned prediction rules may be characterized in terms of the observations themselves. This amounts to capturing how well data can be used to construct structured hypothesis spaces for risk minimization strategies - termed *empirical hypothesis spaces*. Classical analysis of learning algorithms requires the user to define a proper hypothesis space before seeing the data. In practice however, one often decides on the proper learning strategy or the form of the prediction rules of interest after inspection of the data. This theoretical gap constitutes exactly the scope of this workshop.

- | | |
|------------------|---|
| 7.30-8.15 | Avrim Blum
SEMI-SUPERVISED LEARNING AND LEARNING VIA SIMILARITY FUNCTIONS: TWO KEY SETTINGS FOR DATA-DEPENDENT CONCEPT SPACES |
| 8.15-8.19 | Mugizi Rwebangira
LEARNING BY COMBINING NATIVE FEATURES WITH SIMILARITY FUNCTIONS |
| 8.19-8.23 | Zakria Hussain, John Shawe-Taylor
THEORY OF MATCHING PURSUIT IN KERNEL DEFINED FEATURE SPACES |
| 8.23-8.27 | Liva Ralaivola, Marie Szafranski, Guillaume Stempfel
CHROMATIC PAC-BAYES BOUNDS FOR NON-IID DATA |
| 8.27-8.31 | Doru Balcan, Maria-Florina Balcan, Avrim Blum
SAMPLE COMPLEXITY FOR MULTIREOLUTION ICA |
| 8.31-8.35 | Mark Herbster, Guy Lever, Massi Pontil
ONLINE PREDICTION ON LARGE DIAMETER GRAPHS |
| 8.35-8.40 | Nicoló Cesa-Bianchi, Claudio Gentile, Fabio Vitale
ONLINE GRAPH PREDICTION WITH RANDOM TREES |
| 8.40-8.55 | Coffee + Posters |
| 8.55-9.40 | Shai Ben-David
REPRESENTATION OF PRIOR KNOWLEDGE - FROM BIAS TO 'META-BIAS' |

- 9.40-10.00** **Ali Rahimi, Eric Garcia, Maya Gupta**
GENERALIZATION BOUNDS FOR INDEFINITE KERNEL MACHINES
- 10.00-10.30** **Discussion**
POSTERS
- 15.30-16.15** **Claudio Gentile**
FROM ON-LINE ALGORITHMS TO DATA-DEPENDENT GENERALIZATION
- 16.15-16.35** **Amit Dhurandhar, Alin Dobra**
STUDY OF CLASSIFICATION ALGORITHMS USING MOMENT ANALYSIS
- 16.35-16.50** Coffee + Posters
- 16.50-17.35** **Csaba Szepesvári**
THE USE OF UNLABELED DATA IN SUPERVISED LEARNING: THE MANIFOLD DOSSIER
- 17.35-18.15** **Jean-Yves Audibert**
TRANSDUCTIVE LEARNING AND COMPUTER VISION
- 17.15-18.30** **Discussion**

New Directions in Statistical Learning for Meaningful and Reproducible fMRI Analysis

<http://www.cs.princeton.edu/mlneuro/nips08>

Melissa K. Carroll

PRINCETON UNIVERSITY

mkc@princeton.edu

Irina Rish

IBM T.J. WATSON RESEARCH CENTER

rish@us.ibm.com

Francisco Pereira

PRINCETON UNIVERSITY

fpereira@cs.cmu.edu

Guillermo Cecchi

IBM T.J. WATSON RESEARCH CENTER

gcecchi@us.ibm.com

Abstract

Statistical learning methods have become mainstream in the analysis of Functional Magnetic Resonance Imaging (fMRI) data, spurred on by a growing consensus that meaningful neuro-scientific models built from fMRI data should be capable of accurate predictions of behavior or neural functioning. These approaches have convinced most neuroscientists that there is tremendous potential in the decoding of brain states using statistical learning. Along with this realization, though, has come a growing recognition of the limitations inherent in using black-box prediction methods for drawing neuro-scientific interpretations. The primary challenge now is how best to exploit statistical learning to answer scientific questions by incorporating domain knowledge and embodying hypotheses about cognitive processes into our models. Further advances will require resolution of many open questions, including:

1. Variability/Robustness: to what extent do patterns in fMRI replicate across trials, subjects, tasks, and studies? To what extent are processes that are observable through the fMRI BOLD response truly replicable across these different conditions? How similar is the neural functioning of one subject to another?
2. Representation: the most common data representation continues to consider voxels as static and independent, and examples are i.i.d.; however, activation patterns almost surely do not lie in voxel space. What are the true, modular activation structures? What is the relationship between similarity in cognitive state space and similarity in fMRI activation space? Can causality be inferred from fMRI?

This workshop will engage leaders in the field in a debate about these issues while providing an opportunity for presentation of cutting-edge research addressing these questions.

7.30-8.20	Tutorials: Statistical Learning for fMRI - Current and Future Directions FRANCISCO PEREIRA AND GUILLERMO CECCHI
8.20-9.00	Invited talk: Predictive performance and brain map reproducibility LARS KAI HANSEN
9.00-9.10	Coffee
9.10-9.30	Modeling Trial Based Neuroimaging Data MORTEN MØRUP, KRISTOFFER HOUGAARD MADSEN, LARS KAI HANSEN

- 9.30-9.50** **Increasing Robustness of Sparse Regression with the Elastic Net**
MELISSA K. CARROLL, GUILLERMO CECCHI, IRINA RISH, RAHUL GARG, A. RAVI RAO
- 9.50-10.10** **Selecting and Identifying Regions of Interest using Groupwise Regularization**
MARCEL VAN GERVEN, ATUSKO TAKASHIMA, TOM HESKES
- 10.10-10.30** Panel Discussion: Variability and Robustness — Speakers
- 10.30-3.30** Break
- 3.30-4.10** **Invited talk: Analysis of Inter-Subject fMRI Data: Finding Adapted Spatial Representations for Group Inference**
BERTRAND THIRION
- 4.10-4.30** **Functional Holography (FH) Analysis Applied to fMRI Data**
Yael Jacob, Amir Rapson, Michal Kafri, Itay Baruchi, Talma Hendler, Eshel Ben-Jacob
- 4.30-4.50** **A Genetic Programming Approach to fMRI-Decoding**
RAFAEL RAMIREZ, PATRICIA SANZ
- 4.50-5.10** **Spatiotemporal Compression for fMRI Classification with K-Means**
VICENTE L. MALAVE
- 5.10-5.20** Coffee
- 5.20-5.40** Panel Discussion: Data Representations — Speakers
- 5.40-6.20** General Group Discussion
- 6.20-6.30** Closing Remarks — Organizers

Tutorials: Statistical Learning for fMRI - Current and Future Directions

Francisco Pereira, PRINCETON UNIVERSITY

Guillermo Cecchi, IBM TJ WATSON RESEARCH CENTER

Predictive performance and brain map reproducibility

Lars Kai Hansen, DENMARK TECHNICAL UNIVERSITY

Modeling Trial Based Neuroimaging Data

Morten Morup, TECHNICAL UNIVERSITY OF DENMARK

Kristoffer Hougaard Madsen, TECHNICAL UNIVERSITY OF DENMARK

Lars Kai Hansen, TECHNICAL UNIVERSITY OF DENMARK

We will demonstrate in EEG and fMRI data how the proposed convCP model indeed alleviates CPdegeneracy and improves the identification of the consistent activities across trials despite variability in latency and shape. The approach generalizes to the identification of consistent activities across other types of measuring modalities such as subjects and conditions and can also here be used to model the inevitable delay and shape variation. Thus, we hold that flexible models such as the proposed convCP model that address the inevitable variability present in neuroimaging data are important for the identification of the consistent reproducible patterns in neuroimaging experiments. An additional benefit being that the multi-linear representation can alleviate model ambiguities encountered in traditional bilinear analysis.

Increasing Robustness of Sparse Regression with the Elastic Net

Melissa K. Carroll, PRINCETON UNIVERSITY

Guillermo Cecchi, IBM TJ WATSON RESEARCH CENTER

Irina Rish, IBM TJ WATSON RESEARCH CENTER

Rahul Garg, IBM TJ WATSON RESEARCH CENTER

A. Ravi Rao, IBM TJ WATSON RESEARCH CENTER

We explore to what extent the combination of predictive and interpretable modeling can provide new insights for functional brain imaging. For this, we apply a recently introduced regularized regression technique, the Elastic Net, to the analysis of the PBAIC 2007 competition data. Elastic Net regression controls via one parameter the number of voxels in the resulting model, and via another the degree to which correlated voxels are included. We find that this method produces highly predictive models of fMRI data that provide evidence for the distributed nature of neural function. We also use the flexibility of Elastic Net to demonstrate that model robustness can be improved without compromising predictability, in turn revealing the importance of localized clusters of activity. Our findings highlight the functional significance of patterns of distributed clusters of localized activity, and underscore the importance of models that are both predictive and interpretable.

Selecting and Identifying Regions of Identifying Regions of Interest using Groupwise Regularization

Marcel van Gerven, RADBOUD UNIVERSITY NJIMEGEN

Atusko Takashima, RADBOUD UNIVERSITY NJIMEGEN

Tom Heskes, RADBOUD UNIVERSITY NJIMEGEN

An important problem in the use of statistical learning methods for the analysis of functional magnetic resonance imaging (fMRI) data is how to link voxels that have been selected as features for classification to a structural interpretation. For example, it is well-known that ‘1 regularized classifiers give sparse solutions but the selected voxels are typically diffusely distributed throughout the measured volume and do not necessarily correspond to meaningful structures within the brain. In order to facilitate the mapping between functional and structural data we make use of groupwise regularization. In previous work, we have shown that groupwise regularization can be used to perform automated channel selection in EEG experiments and to perform transfer learning, where data for multiple subjects or sessions is combined in the learning process (which is relevant to fMRI analysis in its own right). Here, we demonstrate that groupwise regularization is also suitable for selecting and identifying regions of interest (ROIs) in fMRI experiments. Specifically, we show that groupwise regularization allows for (1) feature selection on the level of predefined ROIs and (2) identification of structure within those ROIs.

Analysis of Inter-Subject fMRI Data: Finding Adapted Spatial Representations for Group Inference

Bertrand Thirion, INRIA

Functional Holography (FH) Analysis Applied to fMRI Data

Yael Jacob, TEL AVIV UNIVERSITY

Amir Rapson, TEL AVIV UNIVERSITY

Michal Kafri, TEL AVIV UNIVERSITY

Itay Baruchi, TEL AVIV UNIVERSITY

Talma Hendler, TEL AVIV UNIVERSITY

Eshel Ben-Jacob, TEL AVIV UNIVERSITY

Here we present a hybrid approach (data driven voxels’ selection for paradigm driven measurements), that is aimed to extract hidden information about functional connectivity motifs in the network of voxel correlations. For that we adopted the functional holography (FH) analysis method that has been developed to analyze subdural EEG brain recordings. The idea is that a complex system behaves like a hologram. In a hologram we can take out a little piece and still see the “big picture”, only in less of a resolution. The approach is based on analysis of the matrices of voxel-voxel correlations. We calculate correlations between the voxels themselves, regardless of the voxel’s amplitude of the BOLD signal, to determine relevance to a functional task. We hypothesize that a large group of voxels showing similar behaviors in time (i.e. have higher temporal

correlations among themselves) are performing functional activity. This way we can acquire functional areas with no prior information. Then we look for clusters in the correlation matrix. Later, voxels are cut from the matrix to leave only the large functional clusters. The relevant voxels are selected by using dendrogram clustering algorithm combined with a special standard-deviation (STD) filtering in which voxels that show correlations with high STD are selected.

A Genetic Programming Approach to fMRI-Decoding

Rafael Ramirez, UNIVERSITAT POMPEU FABRA

Patricia Sanz, UNIVERSITAT POMPEU FABRA

We present a fMRI analysis technique based on genetic programming which combines the two goals of dimensionality reduction and classification into a single learning objective. We apply a multi-tree genetic programming algorithm in which each individual in a population of classifiers considers a different feature subset. The size of the feature subset is determined probabilistically by assigning higher probabilities to smaller sizes. The classifiers which are more accurate using a small number of features are given higher probability to evolve. We describe the results of applying this approach to a fMRI data set involving auditory stimuli.

Spatiotemporal Compression for fMRI Classification with K-Means

Vicente L. Malave, UCSD

We introduce a new way to incorporate temporal information without increasing the dimensionality of the signal. We first cluster response shapes (temporal pattern of image acquisition magnitudes across a trial) and then use cluster membership as a discrete 1-Dimensional signal per trial per voxel. The resulting compression and discretization combines the best attributes of all three approaches: low number of parameters, temporal information, and ability to compare different hemodynamic responses.

Cortical Microcircuits and their Computational Functions

<http://cnl.salk.edu/~terry/NIPS-Workshop/2008/>

Tomaso Poggio

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Terrence Sejnowski

THE SALK INSTITUTE

tp@ai.mit.edu

terry@salk.edu

Abstract

There are around 100,000 neurons under a mm^2 of cerebral cortex and about one billion synapses. Thalamic inputs to the cortex carry information that is transformed by local microcircuits of excitatory and inhibitory neurons. In recent years there has been an explosion of discoveries about the anatomical organization of the microcircuits and the physiological properties of the neurons and synapses that compose them. The goal of this workshop is to explore the functional implications of these new findings and in particular to attempt to characterize the elementary computational operations that are performed in different layers of cortex.

Some of the issues that speakers will address include:

- How is the input from the thalamus able to dominate the cortex when the vast majority of the synapses in cortex are from cortical neurons and the thalamic inputs constitute less than 5% of the synapses on the first layer of cells in layer 4.
- Is there a canonical microcircuit? How does it differ between sensory areas and motor areas, between the early and late stages in the cortical hierarchy, and in cortical areas that support working memory?
- How is the gain of the microcircuit affected by top down inputs from higher cortical areas through attentional control? How are microcircuit with positive feedback stabilized?
- What do the intrinsic properties of dendrites contribute to the computation performed by neurons?
- What is the relation between proposed operations for canonical microcircuits such as gain control, normalization, tuning, soft-max? Can one compare the ventral stream to the dorsal stream?
- What is the consequence of short-term synaptic plasticity on transient and tonic cortical processing?

7.30-10.30	Morning Session (chair — Terry Sejnowski)
7.30-8.10	Canonical cortical microcircuits RODNEY DOUGLAS
8:10-8.50	Microcircuits for perception THOMAS SERRE
8.50-9.00	coffee break
9.00-9.40	Normalization model of attention DAVID HEEGER
9.40-10.30	Microcircuits for gain control PAUL TIESINGA

16.00-19.00	Afternoon Session (chair — Rodney Douglas)
16:00-16.40	Dendritic computation ATTILA LOSONCZY
16:40-17.20	Local balance on dendritic branches TERRY SEJNOWSKI
17.20-17.30	coffee break
17.30-18.10	Progress and prospects for high-throughput reconstruction of neural circuitry VIREN JAIN
18.10-18.30	Mechanisms for cognitive memory BERNIE WIDROW

Canonical cortical microcircuit

Rodney Douglas, INI, ZURICH

Microcircuits for gain control

Paul Tiesinga, UNC, CHAPEL HILL

Dendritic computation

Attila Losonczy, JANELIA FARM

Thalamic inputs to cortex

Terry Sejnowski, SALK/UCSD

Functions of microcircuits

Tomaso Poggio, MIT

Mechanisms for cognitive memory

Bernie Widrow, STANFORD

Connectomics

Sebastian Seung, MIT

Microcircuits for normalization

David Heeger, NYU

Approximate inference - how far have we come?

<http://www.cs.huji.ac.il/~gamir/inference-workshop.html>

Amir Globerson

THE HEBREW UNIVERSITY OF JERUSALEM,

David Sontag

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

Tommi Jaakkola

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

gamir@cs.huji.ac.il

dsontag@csail.mit.edu

tommi@csail.mit.edu

Abstract

Graphical models have become a key tool in representing multi-variate distributions in many machine learning applications. They have been successfully used in diverse fields such as machine-vision, bioinformatics, natural language processing, reinforcement learning and many others.

Approximate inference in such models has attracted a great deal of interest in the learning community, and many algorithms have been introduced in recent years, with a specific emphasis on inference in discrete variable models. These new methods explore new and exciting links between inference, combinatorial optimization, and convex duality. They provide new avenues for designing and understanding message passing algorithms, and can give theoretical guarantees when used for learning graphical models.

The goal of this workshop is to assess the current state of the field and explore new directions. We shall specifically be interested in understanding the following issues:

1. State of the field: What are the existing methods, and how do they relate to each other? Which problems can be solved using existing algorithms, and which cannot?
2. Quality of approximations: What are the theoretical guarantees regarding the output of the approximate inference algorithms (e.g., upper or lower bounds on MAP or marginals, optimality within a given factor, certificates of optimality etc.). How do these depend on the complexity of the inference algorithms (i.e., what is the tradeoff between running time and accuracy)?
3. Efficiency issues: What type of convergence guarantees do different message passing algorithms have, and what can be proven about their running time? Do certain methods dominate others in terms of these guarantees? When are message passing algorithms better than other "out-of-the-box" convex optimization tools?
4. Scalability and applicability to real problems: How well do current methods scale to large-scale problems (e.g. in machine-vision and bioinformatics)? How hard is inference in typical real-world problems? Although inference is generally NP hard, this does not imply that a specific real problem cannot be solved exactly. The relative success of approximate inference methods on some real-world problems suggests that we are working in a regime of problems that are amenable to approximation. Can we characterize it?
5. Connections across fields: Approximate inference is closely linked to problems in combinatorial optimization (e.g., maximization of functions over graphs, or counting problems), and in convex optimization (e.g., dual ascent methods). What techniques can we "import" from these fields, and what from our advances in approximate inference can we "export" back?

6. Inference and learning: Learning the parameters of a graphical model often requires inference as a subroutine. How should approximate inference be embedded into learning? Once a model is learned, what inference method should be used, and how should it relate to the one used during learning? What efficient algorithms exist for joint learning and inference?
7. Continuous models: Many new approximate inference approaches have been developed in the context of discrete variable models. How can these be applied to continuous valued or hybrid models?

The workshop will bring together researchers from diverse communities: machine learning researchers working on approximate inference, practitioners of graphical models from applied communities such as machine-vision and bioinformatics, and researchers from the optimization community who are working on similar problems but in the optimization context.

7.30-8.00	Introduction, review and themes ORGANIZERS
8.00-8.40	TBA MARTIN WAINWRIGHT
8.40-9.10	Convergent Message-Passing algorithms for LP-relaxations and Convex Free Energy minimization using Fenchel Duality TAMIR HAZAN
9.10-9.20	Break
9.20-9.50	MAP Estimation: Setting the State of the Art with Duality Theory and Linear Programming NIKOS KOMODAKIS
9.50-10.30	Approximate inference methods for stochastic optimal control theory BERT KAPPEN
10.30-15.30	Break
15.30-16.10	Andrew Eats Crow: Why Reinforcement Learning Might be Useful After All; Massive-Scale, Relational MAP Inference by MCMC with Delayed Reward ANDREW MCCALLUM
16.10-16.50	Inference in Large Scale Nonparametric Bayesian Models MAX WELLING
16.50-17.00	Break
17.00-17.30	Learning Deep Boltzmann Machines RUSLAN SALAKHUTDINOV
17.30-18.10	TBA ADNAN DARWICHE
18.10-18.30	Discussion and Conclusions ORGANIZERS

Kernel Learning: Automatic Selection of Optimal Kernels

http://www.cs.nyu.edu/learning_kernels

Corinna Cortes

GOOGLE RESEARCH NEW YORK

corinna@google.com

Arthur Gretton

MAX PLANCK INSTITUTE FOR BIOLOGICAL CYBERNETICS

arthur@tuebingen.mpg.de

Gert Lanckriet

UNIVERSITY OF CALIFORNIA, SAN DIEGO

gert@ece.ucsd.edu

Mehryar Mohri

COURANT INSTITUTE OF MATHEMATICAL SCIENCES & GOOGLE RESEARCH

mohri@cims.nyu.edu

Afshin Rostamizadeh

COURANT INSTITUTE OF MATHEMATICAL SCIENCES

rostami@cs.nyu.edu

Abstract

Kernel methods are widely used to address a variety of learning tasks including classification, regression, ranking, clustering, and dimensionality reduction. The appropriate choice of a kernel is often left to the user. But, poor selections may lead to sub-optimal performance. Furthermore, searching for an appropriate kernel manually may be a time-consuming and imperfect art. Instead, the kernel selection process can be included as part of the overall learning problem. In this way, better performance guarantees can be given and the kernel selection process can be made automatic. In this workshop, we will be concerned with using sampled data to select or learn a kernel function or kernel matrix appropriate for the specific task at hand. We will discuss several scenarios, including classification, regression, and ranking, where the use of kernels is ubiquitous, and different settings including inductive, transductive, or semi-supervised learning. We also invite discussions on the closely related fields of features selection and extraction, and are interested in exploring further the connection with these topics. The goal is to cover all questions related to the problem of learning kernels: different problem formulations, the computational efficiency and accuracy of the algorithms that address these problems and their different strengths and weaknesses, and the theoretical guarantees provided. What is the computational complexity? Does it work in practice? The formulation of some other learning problems, e.g. multi-task learning problems, is often very similar. These problems and their solutions will also be discussed in this workshop.

7:30-8:00	Invited Speaker: Shai Ben-David THE SAMPLE COMPLEXITY OF LEARNING THE KERNEL
8:00-8:20	Olivier Chapelle and Alain Rakotomamonjy SECOND ORDER OPTIMIZATION OF KERNEL PARAMETERS
8:20-8:50	Invited Speaker: William Stafford Noble MULTI-KERNEL LEARNING FOR BIOLOGY
8:50-9:20	Poster Session and Discussion
9:20-9:40	Corinna Cortes, Mehryar Mohri and Afshin Rostamizadeh LEARNING SEQUENCE KERNELS
9:40-10:00	Maria-Florina Balcan, Avrim Blum and Nathan Srebro LEARNING WITH MULTIPLE SIMILARITY FUNCTIONS

10:00-10:30	Invited Speaker: Andreas Argyriou MULTI-TASK LEARNING VIA MATRIX REGULARIZATION
10:30-15:30	Break until afternoon session.
15:30-16:00	Invited Speaker: Isabelle Guyon FEATURE SELECTION: FROM CORRELATION TO CAUSALITY
16:00-16:20	Nathan Srebro and Shai Ben-David LEARNING BOUNDS FOR SUPPORT VECTOR MACHINES WITH LEARNED KERNELS
16:20-16:50	Invited Speaker: Alex Smola MIXED NORM KERNELS, HYPERKERNELS AND OTHER VARIANTS
16:50-17:20	Poster Session and Discussion
17:20-17:40	Marius Kloft, Ulf Brefeld, Pavel Laskov and Sören Sonnenburg NON-SPARSE MULTIPLE KERNEL LEARNING
17:40-18:00	Peter Gehler INFINITE KERNEL LEARNING
18:00-18:30	Invited Speaker: John Shawe-Taylor KERNEL LEARNING FOR NOVELTY DETECTION
18:30	Closing Remarks

The Sample Complexity of Learning the Kernel

Shai Ben-David, UNIVERSITY OF WATERLOO

The success of kernel based learning algorithms depends upon the suitability of the kernel to the learning task. Ideally, the choice of a kernel should be based on prior information of the learner about the task at hand. However, in practice, kernel parameters are being tuned based on available training data. I will discuss the sample complexity overhead associated with such "learning the kernel" scenarios. I will address the setting in which the training data for the kernel selection is target labeled examples, as well as settings in which this training is based on different types of data, such as unlabeled examples and examples labeled by a different (but related) tasks. Part of this work is joint with Nati Srebro.

Second Order Optimization of Kernel Parameters

Olivier Chapelle et al., YAHOO! RESEARCH & UNIVERSITY ROUEN

We investigate the use of second order optimization approaches for solving the multiple kernel learning (MKL) problem. We show that the hessian of the MKL can be computed efficiently and this information can be used to compute a better descent direction than the gradient (used in the state-of-the-art SimpleMKL algorithm). We then empirically show that our new approaches outperforms SimpleMKL in terms of computational efficiency.

Multi-Kernel Learning for Biology

William Stafford Noble, UNIVERSITY OF WASHINGTON

One of the primary tasks facing biologists today is to integrate the different views of molecular biology that are provided by various types of experimental data. In yeast, for example, for a given gene we typically know the protein it encodes, that protein's similarity to other proteins, the mRNA expression levels associated with the given gene under hundreds of experimental conditions, the occurrences of known or inferred transcription factor binding sites in the upstream region of that gene, and the identities of many of the proteins that interact with the given gene's protein product. Each of these distinct data types provides one view of the molecular machinery of the cell.

Kernel methods allow us to represent these heterogeneous data types in a normal form, and to use kernel algebra to reason about more than one type of data simultaneously. Consequently, multi-kernel learning

methods have been applied to a variety of biology applications. In this talk, I will describe several of these applications, outline the lessons we have learned from applying multi-kernel learning methods to real data, and suggest several avenues for future research in this area.

Learning Sequence Kernels

Corinna Cortes et al., GOOGLE RESEARCH & COURANT INSTITUTE

Kernel methods are used to tackle a variety of learning tasks including classification, regression, ranking, clustering, and dimensionality reduction. The appropriate choice of a kernel is often left to the user. But, poor selections may lead to a sub-optimal performance. Instead, sample points can be used to learn a kernel function appropriate for the task by selecting one out of a family of kernels determined by the user. This paper considers the problem of *learning sequence kernel functions*, an important problem for applications in computational biology, natural language processing, document classification and other text processing areas. For most kernel-based learning techniques, the kernels selected must be positive definite symmetric, which, for sequence data, are found to be rational kernels. We give a general formulation of the problem of learning rational kernels and prove that a large family of rational kernels can be learned efficiently using a simple quadratic program both in the context of support vector machines and kernel ridge regression. This improves upon previous work that generally results in a more costly semi-definite or quadratically constrained quadratic program. Furthermore, in the specific case of kernel ridge regression, we give an alternative solution for the optimal *kernel matrix*, which in fact coincides with the objective prescribed by kernel alignment techniques.

Learning with Multiple Similarity Functions

Maria-Florina Balcan et al., MICROSOFT RESEARCH & CARNEGIE MELLON UNIVERSITY & TOYOTA TECHNOLOGICAL INSTITUTE

Kernel functions have become an extremely popular tool in machine learning, with many applications and an attractive theory. There has also been substantial work on learning kernel functions from data [LCBGJ04,SB06,AHMP08]. A sufficient condition for a kernel to allow for good generalization on a given learning problem is that it induce a large margin of separation between positive and negative classes in its implicit space. In recent work [BBS08,BBS07,BB06] we have developed a theory that more broadly holds for general similarity functions that are not necessarily legal kernel functions. In particular, we give sufficient conditions for a similarity function to be useful for learning that (a) are fairly natural and intuitive (do not require an implicit space and allow for functions that are not positive semi-definite) and (b) strictly generalize the notion of a large-margin kernel function in that any such kernel also satisfies these conditions, though not necessarily vice-versa. We also have partial progress on extending the theory of learning with *multiple* kernel functions to these more general conditions. In this talk we describe the main definitions and results of [BBS08], give our results on learning with multiple similarity functions, and present several open questions about learning good general similarity functions from data.

Multi-Task Learning via Matrix Regularization

Andreas Argyriou, UNIVERSITY COLLEGE LONDON

We present a method for learning representations shared across multiple tasks. The method consists in learning a low-dimensional subspace on which task regression vectors lie. Our formulation is a convex optimization problem, which we solve with an alternating minimization algorithm. This algorithm can be shown to always converge to an optimal solution. Our method can also be viewed as learning a linear kernel shared across the tasks and hence as an instance of kernel learning in which there are infinite kernels available. Moreover, the method can easily be extended in order to learn multiple tasks using nonlinear kernels. To justify this, we present general results characterizing representer theorems for matrix learning problems like the one above, as well as standard representer theorems. Finally, we briefly describe how our method connects to approaches exploiting sparsity such as group Lasso.

Feature Selection: From Correlation to Causality

Isabelle Guyon, CLOPINET, BERKELEY

Variable and feature selection have become the focus of much research in areas of application for which datasets with tens or hundreds of thousands of variables are available. These areas include text processing of

internet documents, gene expression array analysis, and combinatorial chemistry. The objective of variable selection is three-fold: improving the prediction performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. This tutorial will cover a wide range of aspects of such problems: providing a better definition of the objective function, feature construction, feature ranking, multivariate feature selection, efficient search methods, and feature validity assessment methods. Most feature selection methods do not attempt to uncover causal relationships between feature and target and focus instead on making best predictions. We will examine situations in which the knowledge of causal relationships benefits feature selection. Such benefits may include: explaining relevance in terms of causal mechanisms, distinguishing between actual features and experimental artifacts, predicting the consequences of actions performed by external agents, and making predictions in non-stationary environments.

Learning Bounds for Support Vector Machines with Learned Kernels

Nathan Srebro et al., TOYOTA TECHNOLOGICAL INSTITUTE & UNIVERSITY OF WATERLOO

Consider the problem of learning a kernel for use in SVM classification. We bound the estimation error of a large margin classifier when the kernel, relative to which this margin is defined, is chosen from a family of kernels based on the training sample. For a kernel family with pseudodimension d_ϕ , we present a bound of $\sqrt{\tilde{O}(d_\phi + 1/\gamma^2)}/n$ on the estimation error for SVMs with margin γ . This is the first bound in which the relation between the margin term and the family-of-kernels term is additive rather than multiplicative. The pseudodimension of families of linear combinations of base kernels is the number of base kernels. Unlike in previous (multiplicative) bounds, there is no non-negativity requirement on the coefficients of the linear combinations. We also give simple bounds on the pseudodimension for families of Gaussian kernels.

Non-sparse Multiple Kernel Learning

Marius Kloft et al., TU BERLIN & FRAUNHOFER INSTITUTE FIRST

Approaches to multiple kernel learning (MKL) employ ℓ_1 -norm constraints on the mixing coefficients to promote sparse kernel combinations. When features encode orthogonal characterizations of a problem, sparseness may lead to discarding useful information and may thus result in poor generalization performance. We study non-sparse multiple kernel learning by imposing an ℓ_2 -norm constraint on the mixing coefficients. Empirically, ℓ_2 -MKL proves robust against noisy and redundant feature sets and significantly improves the promoter detection rate compared to ℓ_1 -norm and canonical MKL on large scales.

Infinite Kernel Learning

Peter Gehler, MAX PLANCK INSTITUTE

In this paper we build upon the Multiple Kernel Learning (MKL) framework. We rewrite the problem in the standard MKL formulation which leads to a Semi-Infinite Program. We devise a new algorithm to solve it (Infinite Kernel Learning, IKL). The IKL algorithm is applicable to both the finite and infinite case and we find it to be faster and more stable than SimpleMKL. Furthermore we present the first large scale comparison of SVMs to MKL on a variety of benchmark datasets, also comparing IKL. The results show two things: a) for many datasets there is no benefit in using MKL/IKL instead of the SVM classifier, thus the flexibility of using more than one kernel seems to be of no use, b) on some datasets IKL yields massive increases in accuracy over SVM/MKL due to the possibility of using a largely increased kernel set. For those cases parameter selection through Cross-Validation or MKL is not applicable.

Kernel Learning for Novelty Detection

John Shawe-Taylor, UNIVERSITY COLLEGE LONDON

We consider kernel learning for one-class Support Vector Machines. We consider a mix of 2- and 1-norms of the individual weight vector norms allowing control of the sparsity of the resulting kernel combination. The resulting optimisation can be solved efficiently using a coordinate gradient method. We consider an application to automatically detecting the appropriate metric for a guided image search task.

POSTER SESSION

-**Ravi S. Ganti, Nikolaos Vasiloglou and Alexander Gray**: Hyperkernel Based Density Estimation

-**Andrew G. Howard and Tony Jebara**: Learning Large Margin Mappings

-**S. Mosci, M. Santoro, A. Verri, S. Villa and L. Rosasco**: A New Algorithm to Learn an Optimal Kernel Based on Fenchel Duality

-**Hua Ouyang and Alexander Gray**: Learning Nearest-Neighbor Classifiers with Hyperkernels

-**Nikolaos Vasiloglou, Alexander G. Gray and David V. Anderson**: Learning Isometric Separation Maps

All other submitted talks are also encouraged to give posters.

DECEMBER 13, 2008, 07:30–10:30 AND 3:30–6:30

HILTON: SUTCLIFFE A WS18

Parallel Implementations of Learning Algorithms: What have you done for me lately?

<http://www.cs.cmu.edu/~dst/NIPS/nips08-workshop>

Robert Thibadeau

SEAGATE RESEARCH

robert.thibadeau@seagate.com

Dan Hammerstrom

PORTLAND STATE UNIVERSITY

strom@cecs.pdx.edu

David Touretzky

CARNEGIE MELLON UNIVERSITY

dst@cs.cmu.edu

Tom Mitchell

CARNEGIE MELLON UNIVERSITY

tom.mitchell@cs.cmu.edu

Abstract

Interest in parallel hardware concepts, including multicore, specialized hardware, and multimachine, has recently increased as researchers have looked to scale up their concepts to large, complex models and large datasets. In this workshop, a panel of invited speakers will present results of investigations into hardware concepts for accelerating a number of different learning and simulation algorithms. Additional contributions will be presented in poster spotlights and a poster session at the end of the one-day workshop. Our intent is to provide a broad survey of the space of hardware approaches in order to capture the current state of activity in this venerable domain of study. Approaches to be covered include silicon, FPGA, and supercomputer architectures, for applications such as Bayesian network models of large and complex domains, simulations of cortex and other brain structures, and large-scale probabilistic algorithms.

7.30-7.40	Introduction and Overview
7.40-8.10	When (And Why) Storage Devices Become Computers ROBERT THIBADEAU
8.40-9.10	A Neocortex-Inspired Cognitive Model on the Cray XD1 KENNETH RICE
9.10-9.30	coffee break
9.30-10.00	Nanoelectronics: The Original Positronic Matrix? DAN HAMMERSTROM
10.00-10:30	CNP: An FPGA-based Processor for Convolutional Networks CLEMENT FARABET, CYRIL POULET, AND YANN LECUN
15.30-16.00	Using a Fast Array of Wimpy Nodes DAVID ANDERSEN
16.00-16.30	Learning Large Deep Belief Networks using Graphics Processors RAJAT RAINA
16.30-17.00	A Bird's-Eye View of PetaVision, the World's First Petaflop/s Neural Simulation DANIEL COATES
17.00-17.20	coffee break

- 17.20** Poster Spotlights
- 17.20-17.24** **Reinforcement Learning Recordbook <RL@Home>**
BRIAN TANNER
- 17.24-17.28** **Efficient, Scalable, and Parallel Event-Drive Simulation Techniques for Complex Spiking Neuron Models**
MICHIEL D’HAENE, BENJAMIN SCHRAUWEN, AND DIRK STROOBANDT
- 17.28-17.32** **FPGA-based Accelerators for “Learning to Rank” in Web Search Engines**
NING-YI XU, JING YAN, RUI GAO, XIONGFEI CAI, ZENGLIN XIA, AND FENG-HSIUNG HSU
- 17.32-17.36** **An FPGA-based Massively Parallel hardware Accelerator for SVM and CN**
HANS PETER GRAF, SRIHARI CADAMBI, IGOR DURDANOVIC, VENKATA JAKKULA, MURUGAN SANKARDADASS, ERIC COSATTO, AND SRIMAT CHAKRADHAR
- 17.40-18.00** General Discussion
- 18.00-18.30** Poster Session

DECEMBER 13, 2008, 07:30–10:30 AND 15:30–18:30

WESTIN: CALLAGHAN WS19

Model Uncertainty and Risk in Reinforcement Learning

<http://www.cs.uwaterloo.ca/~ppoupart/nips08-workshop.html>

Yaakov Engel

yakiengel@gmail.com

Mohammad Ghavamzadeh
INRIA LILLE - TEAM SEQUEL

mgh@cs.ualberta.ca

Shie Mannor
MCGILL UNIVERSITY

shie.mannor@mcgill.ca

Pascal Poupart
UNIVERSITY OF WATERLOO

ppoupart@cs.uwaterloo.ca

Abstract

Reinforcement Learning (RL) problems are typically formulated in terms of Stochastic Decision Processes (SDPs), or a specialization thereof, Markovian Decision Processes (MDPs), with the goal of identifying an optimal control policy. In contrast to planning problems, RL problems are characterized by the lack of complete information concerning the transition and reward models of the SDP. Hence, algorithms for solving RL problems need to estimate properties of the system from finite data. Naturally, any such estimated quantity has inherent uncertainty. One of the interesting and challenging aspects of RL is that the algorithms have partial control over the data sample they observe, allowing them to actively control the amount of this uncertainty, and potentially trade it off against performance. Reinforcement Learning as a field of research, has over the past few years seen renewed interest in methods that explicitly consider the uncertainties inherent to the learning process. Indeed, interest in data-driven models that take uncertainties into account, goes beyond RL to the fields of Control Theory, Operations Research and Statistics. Within the RL community, relevant lines of research include Bayesian RL, risk sensitive and robust dynamic decision making, RL with confidence intervals and applications of risk-aware and uncertainty-aware decision-making. The goal of the workshop is to bring together researchers in RL and related fields that work on issues related to risk and model uncertainty, stimulate interactions and discuss directions for future work.

- | | |
|--------------------|---|
| 07.30-07.40 | Welcome |
| 07.40-08.20 | Invited Talk
MICHAEL LITTMAN |
| 08.20-08.40 | Bias Correction and Confidence Intervals for Fitted Q-iteration
BIBHAS CHAKRABORTY, VICTOR STRECHER, SUSAN MURPHY |
| 08.40-09.00 | Risk-Aware Decision Making and Dynamic Programming
BORIS DEFURNY, DAMIEN ERNST, LOUIS WEHENKEL |
| 09.00-09.15 | Break |
| 09.15-09.40 | Poster Spotlights |
- Paper Posters
 1. **Kalman Temporal Differences: Uncertainty and Value Function Approximation**
MATTHIEU GEIST, GABRIEL FRICOUT, OLIVIER PIETQUIN

2. **Missing Data and Uncertainty in Batch Reinforcement Learning**
DANIEL J. LIZOTTE, LACEY GUNTER, ERIC LABER, SUSAN A. MURPHY
 3. **Error Reducing Sampling in Reinforcement Learning**
BRUNO SCHERRER AND SHIE MANNOR
 4. **Towards Global Reinforcement Learning**
MILEN PAVLOV AND PASCAL POUPART
 5. **Incorporating External Evidence in Reinforcement Learning via Power Prior Bayesian Analysis**
FUNLADE T. SUNMOLA AND JEREMY L. WYATT
 6. **Uncertainty Handling in Evolutionary Direct Policy Search**
VERENA HEIDRICH-MEISNER, CHRISTIAN IGEL
 7. **The Optimal Unbiased Value Estimator and its Relation to LSTD**
STEFFEN GRÜNEWÄLDER, KLAUS OBERMAYER
- Extended Abstract Posters
 1. **Model-based Bayesian Reinforcement Learning with Tree-based State Aggregation**
COSMIN PADURARU, DOINA PRECUP, STEPHANE ROSS, JOELLE PINEAU
 2. **Near-Bayesian Exploration in Polynomial Time**
J. ZICO KOLTER, ANDREW Y. NG
 3. **How Close is Close Enough? Finding Optimal Policies in PAC-style Reinforcement Learning**
EMMA BRUNSKILL
 4. **Bayesian Exploration using Gaussian Processes: Fast Convergence via Generalization**
ERICK CHASTAIN AND RAJESH P. N. RAO
- 09.40-10.30** Panel Discussion (Benchmarks and Challenges) and Poster Session
- 15.30-16.10** **Invited talk**
CSABA SZEPESVARI
- 16.10-16.30** **Risk Sensitive Control: Mean-Variance Tradeoffs**
STEFFEN GRÜNEWÄLDER, AKI NAITO AND KLAUS OBERMAYER
- 16.30-16.50** **PAC-MDP Reinforcement Learning with Bayesian Priors in Deterministic MDPs**
ALI NOURI, MICHAEL LITTMAN AND LIHONG LI
- 16.50-17.05** Break
- 17.05-17.25** **Regret-based Reward Elicitation for Markov Decision Processes**
KEVIN REGAN AND CRAIG BOUTILIER
- 17.25-17.45** **Data Biased Robust Counter Strategies**
MICHAEL JOHANSON, MICHAEL BOWLING
- 17.45-18.30** Panel Discussion (Models that Work and Don't Work) and Poster Session

DECEMBER 12, 2008, 07:30–10:30 AND 15:30–18:30

HILTON: MT. CURRIE S WS20

Principled Theoretical Frameworks for the Perception-Action Cycle

http://homepages.feis.herts.ac.uk/~comqdp1/NIPS_2008/NIPS_Symposium_Workshop.html

Daniel Polani

UNIVERSITY OF HERTFORDSHIRE

d.polani@herts.ac.uk

Naftali Tishby

THE HEBREW UNIVERSITY

tishby@cs.huji.ac.il

Abstract

A significant emphasis in trying to achieve adaptation and learning in the perception-action cycle of agents lies in the development of suitable algorithms. While partly these algorithms result from mathematical constructions, in modern research much attention is given to methods that mimic biological processes.

However, mimicking the apparent features of what appears to be a biologically relevant mechanism makes it difficult to separate the essentials of adaptation and learning from accidents of evolution. This is a challenge both for the understanding of biological systems as well as for the design of artificial ones. Therefore, recent work is increasingly concentrating on identifying general principles rather than individual mechanisms for biologically relevant information processing.

One advantage is that a small selection of principles can give rise to a variety of — effectively equivalent — mechanisms. The ultimate goal is to attain a more transparent and unified view on the phenomena in question. Possible candidates for such principles governing the dynamics of the perception-action cycle include but are not limited to information theory, Bayesian models, energy-based concepts or principles emerging from neuroscience

7.30-8.30	Information Theory and the Perception-Action Loop NAFTALI TISHBY, HEBREW UNIVERSITY
8.30-9.15	Information Bottleneck Optimization with Spiking Neurons with Application to Predictive Coding LARS BÜSING
9.15-9.45	Discussion and Coffee
9.45-10.30	TBA NIHAT AY, MAX-PLANCK INSTITUTE FOR MATHEMATICS IN THE SCIENCES, LEIPZIG
10.30	Noon Break
15.30-16.30	Information Theory and the Perception-Action Loop II DANIEL POLANI, UNIVERSITY OF HERTFORDSHIRE
16.30-17.00	Bayesian Modelling of a Sensorimotor Loop: Application to Handwriting ESTELLE GILET, CNRS — INRIA RHÔNE-ALPES
17.00-17.30	Discussion and Coffee
17.30-18.00	Fundamental Dynamic Properties of Coupled Systems STEFAN WINTER, UNIVERSITY OF MAINZ

18.00-18.30 Empowerment: The External Channel Capacity of a Sensorimotor Loop and a Method for its Estimation

TOBIAS JUNG, UNIVERSITY OF TEXAS AT AUSTIN

Information Theory and the Perception-Action Loop

Naftali Tishby, HEBREW UNIVERSITY

Information Bottleneck Optimization with Spiking Neurons with Application to Predictive Coding

Lars Büsing, TU GRAZ

Wolfgang Maass, TU GRAZ

We apply the online learning algorithm for IB optimization to the predictive coding task outlined in (Bialek et al., 2007) on the closed loop.

TBA

Nihat Ay, MAX PLANCK INSTITUTE, LEIPZIG

Information Theory and the Perception-Action Loop II

Daniel Polani, UNIVERSITY OF HERTFORDSHIRE

We outline recent approaches to characterize the sensorimotor loop of agents from an informational perspective. This view allows the principled characterization of agent behaviours in an environment and of possible paths towards minimalistic AI.

Bayesian Modelling of a Sensorimotor Loop: Application to Handwriting

Estelle Gilet, CNRS — INRIA RHÔNE-ALPES

Julien Diard, CNRS — UNIVERSITÉ PIERRE MENDÈS

Pierre Bessière, CNRS — UNIVERSITÉ PIERRE MENDÈS

This paper concerns the Bayesian modelling of a sensorimotor loop. We present a preliminary model of handwriting, that provides both production of letters and their recognition. It is structured around an abstract internal representation of letters, which acts as a pivot between motor and sensors models. The representation of letters is independent of the effector usually used to perform the movement. We show how our model allows to solve a variety of tasks, like letter reading, recognizing the writer, and letter writing (with different effectors). We show how the joint modelling of the sensory and motor systems allows to solve reading tasks in the case of noisy inputs by internal simulation of movements.

Fundamental Dynamic Properties of Coupled Systems

Stefan Winter, UNIVERSITY OF MAINZ

Empowerment: The External Channel Capacity of a Sensorimotor Loop and a Method for its Estimation

Tobias Jung, UNIVERSITY OF TEXAS AT AUSTIN

Empowerment, the external channel capacity of a sensorimotor loop has been introduced recently as a quantity, similar to predictive information to characterize the sensorimotor efficiency and evaluate the compatibility of the niche of an agent with its sensorimotor loop, and its quality. While computable in simple scenarios, its evaluation in continuous or higher-dimensional situations is still difficult. Here, we present a computational approach to that purpose and demonstrate an instructive application.

DECEMBER 13, 2008, 07:30–10:30 AND 15:30–18:30

WESTIN: ALPINE CD WS21

Probabilistic Programming: Universal Languages and Inference; Systems; and Applications

<http://probabilistic-programming.org>

Daniel Roy	droy@mit.edu
MASSACHUSETTS INSTITUTE OF TECHNOLOGY	
Vikash Mansinghka	vkm@mit.edu
MASSACHUSETTS INSTITUTE OF TECHNOLOGY	
John Winn	jwinn@microsoft.com
MICROSOFT RESEARCH CAMBRIDGE	
David McAllester	mcallester@tti-c.org
TOYOTA TECHNICAL INSTITUTE AT CHICAGO	
Joshua Tenenbaum	jbt@mit.edu
MASSACHUSETTS INSTITUTE OF TECHNOLOGY	

Abstract

Probabilistic graphical models provide a formal lingua franca for modeling and a common target for efficient inference algorithms. Their introduction gave rise to an extensive body of work in machine learning, statistics, robotics, vision, biology, neuroscience, AI and cognitive science. However, many of the most innovative and exciting probabilistic models published by the NIPS community far outstrip the representational capacity of graphical models and are instead communicated using a mix of natural language, pseudo code, and mathematical formulae and solved using special purpose, one-off inference methods. Very often, graphical models are used only to describe the coarse, high-level structure rather than the precise specification necessary for automated inference. Probabilistic programming languages aim to close this representational gap; literally, users specify a probabilistic model in its entirety (e.g., by writing code that generates a sample from the joint distribution) and inference follows automatically given the specification. Several existing systems already satisfy this specification to varying degrees of expressiveness, compositionality, universality, and efficiency. We believe that the probabilistic programming language approach, which has been emerging over the last 10 years from a range of diverse fields including machine learning, computational statistics, systems biology, probabilistic AI, mathematical logic, theoretical computer science and programming language theory, has the potential to fundamentally change the way we understand, design, build, test and deploy probabilistic systems. The NIPS workshop will be a unique opportunity for this diverse community to meet, share ideas, collaborate, and help plot the course of this exciting research area.

7.30-7.45	Opening Address DANIEL ROY, MIT
7.45-8.15	Invited talk: The open universe STUART RUSSELL, UC BERKELEY
8:15-8:45	What We Can Do with Markov Logic Today PEDRO DOMINGOS, UNIVERSITY OF WASHINGTON
8:45-9.00	Coffee
9.00-9.30	Church: a universal language for generative models VIKASH MANSINGHKA, MIT

- 9:30-10:00** **Automatic Differentiation for Probabilistic Programming**
JEFFREY SISKIND, PURDUE UNIV.
- 10:00-10:30** Poster Session
- 10:30-11:30** Extended Poster Session and Demo Session
- 12:00-2:00** Group Lunch
- 3:30-4:00** **A Probabilistic Language for Biological Modeling**
ANDREW PHILLIPS, MICROSOFT RESEARCH
- 4:00-4:30** **Dyna: A Non-Probabilistic Programming Language for Probabilistic AI**
JASON EISNER, JOHN HOPKINS UNIV.
- 4:30-5:00** **PMTK: probabilistic modeling toolkit**
KEVIN MURPHY, UNIVERSITY OF BRITISH COLUMBIA
- 5:00-5:30** **FACTORIE: Efficient Probabilistic Programming for Relational Factor
Graphs via Imperative Declarations of Structure, Inference and Learning**
ANDREW MCCALLUM, UNIVERSITY OF MASSACHUSETTS AMHEARST
- 5:30-6:00** **Infer.NET and the CSOFT language**
JOHN WINN, MICROSOFT RESEARCH
- 6:00-6:30** **Software or language? Panel Discussion**

Sunday, December 14th

- 8:00-8:30** **Constraint-based Probabilistic Modeling**
TAISUKE SATO, TOKYO INSTITUTE OF TECHNOLOGY
- 8:30-9:00** **Probabilistic Programming for Cognitive Science**
NOAH GOODMAN, MIT
- 9:00-9:30** **Syntactic Independence in Stochastic Functional Languages**
DAVID MCALLESTER, TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO
- 9:30-10:00** **Towards Digesting the Alphabet-Soup of Statistical Relational Learning**
LUC DE RAEDT, KATHOLIEKE UNIVERSITEIT LEUVEN
- 10:00-10:30** Brainstorming Session

The open universe

Stuart Russell, UNIVERSITY OF CALIFORNIA, BERKELEY

Recent advances in knowledge representation for probability models have allowed for uncertainty about the properties of objects and the relations that might hold among them. Such models, however, typically assume exact knowledge of which objects exist and of which object is which—that is, they assume *domain closure* and *unique names*. These assumptions simplify the sample space for probability models, but are inappropriate for many real-world situations in which the universe of discourse is *open*. These include most instances of perception and language understanding. This talk describes the evolution of open-universe probability models, our current efforts associated with the BLOG language, and some future directions. *Joint work with Brian Milch, Bhaskara Marthi, Rodrigo Braz, Hanna Pasula, David Sontag, Andrey Kolobov, and Daniel Ong.*

What We Can Do with Markov Logic Today

Pedro Domingos, UNIVERSITY OF WASHINGTON

A program in Markov logic is a set of weighted formulas in first-order logic, interpreted as templates for features of a Markov random field. Most widely used graphical models can be specified very compactly in Markov logic. In turn, this makes it easy to define large, complex models involving combinations of these components. For example, a complete information extraction system, performing joint segmentation and entity resolution, can be specified with just seven formulas. Current implementations of Markov logic enable routine learning and inference on graphical models with millions of variables and billions of features. State-of-the-art solutions to a wide variety of problems have been developed using Markov logic, including information extraction, robot mapping, unsupervised coreference resolution, semantic role labeling, prediction of protein beta-partners, parts of the CALO personal assistant, probabilistic extensions of the Cyc knowledge base, and others. This talk will survey the state of the art in Markov logic: language features, inference and learning algorithms, applications to date and research challenges.

Church: a universal language for generative models

Vikash Mansinghka, MIT

To engineer robust, adaptive, autonomous systems and explain human intelligence in computational terms, we must develop formal languages for the structured representation of uncertain knowledge and machines capable of efficiently solving the resulting inference problems. In this talk, I will present the first of these two layers, Church, a probabilistic programming language which provides a unified procedural notation for stochastic generative processes and uncertain beliefs. Church recovers the pure subset of Scheme in its deterministic limit, recovers McCarthy's amb in its deductive limit, and generalizes Lisp evaluation to conditional stochastic simulation for universal Bayesian inference. I will show how to use Church to compactly specify a range of problems, including nonparametric Bayesian models, planning as inference, and Bayesian learning of programs from data. *Joint work with Noah Goodman, Daniel Roy, and Joshua Tenenbaum.*

Automatic Differentiation for Probabilistic Programming

Jeffrey Mark Siskind, SCHOOL OF ELECTRICAL AND COMPUTER ENGINEERING, PURDUE UNIVERSITY

We have developed powerful and efficient techniques for taking gradients of functional programs. Our methods extends the efficiency of prior approaches that are limited to imperative numeric programs to the realm of functional programs that mix symbolic and numeric computation. We demonstrate the relevance and significance of this effort to the enterprise of probabilistic programming by constructing evaluators for two different probabilistic programming languages and using our methods to take the gradients of such evaluators in order to perform gradient-based maximum-likelihood parameter estimation.

Poster: Distributional Logic Programming: a brief overview

Nikolaos Angelopoulos, EDINBURGH UNIVERSITY

Poster: Observational Languages and Ontologies

David Poole, UNIVERSITY OF BRITISH COLUMBIA

Clinton Smyth, UNIVERSITY OF BRITISH COLUMBIA

Poster: Multi-agent Markov Logic

Miroslav Dudík, CARNEGIE MELLON UNIVERSITY

Geoff Gordon, CARNEGIE MELLON UNIVERSITY

Austin McDonald, CARNEGIE MELLON UNIVERSITY

Poster: Program inference by importance sampling

Georges Harik, HS LABS

Noam Shazeer, HS LABS

Poster: Open-Universe State Estimation with DBLOG

Rodrigo de Salvo Braz, UC BERKELEY

Nimar Arora, UC BERKELEY
Erik Sudderth, UC BERKELEY
Stuart Russell, UC BERKELEY

Poster: Automatic Inference in PyBLOG

Nimar Arora, UC BERKELEY
Rodrigo de Salvo Braz, UC BERKELEY
Erik Sudderth, UC BERKELEY
Stuart Russell, UC BERKELEY

Poster: Dependency Diagrams for Probabilistic Models and their Sampling and Inference Algorithms

Todd Johnson, UC IRVINE
Eric Mjolsness, UC IRVINE

Poster: CP-Logic Theory Inference with Contextual Variable Elimination

Wannes Meert, KATHOLIEKE UNIVERSITEIT LEUVEN
Jan Struyf, KATHOLIEKE UNIVERSITEIT LEUVEN
Hendrik Blockeel, KATHOLIEKE UNIVERSITEIT LEUVEN

Poster: Bach: Probabilistic Declarative Programming

John Lloyd, AUSTRALIAN NATIONAL UNIVERSITY
Kee Siong Ng, AUSTRALIAN NATIONAL UNIVERSITY
Will Uther, AUSTRALIAN NATIONAL UNIVERSITY

Poster: Embedded Probabilistic Programming

Chung-chieh Shan, RUTGERS UNIVERSITY
Oleg Kiselyov, FNMOC

Poster: Bayesian Programming formalism and ProBT API

Pierre Bessiere, CNRS - INRIA
J-M Ahuactzin, PROBAYES, INC.
E Mazer, PROBAYES, INC.
K Mekhnacha, PROBAYES, INC.

A Probabilistic Language for Biological Processes

Andrew Phillips, MICROSOFT RESEARCH CAMBRIDGE

This talk presents a programming language for designing and simulating computer models of biological processes. The language is based on a mathematical formalism known as the pi-calculus, and the simulation algorithm is based on standard kinetic theory of physical chemistry. The language is first presented using a simple graphical notation, which is subsequently used to model and simulate an immune system pathway relating to the detection of pathogens inside a cell. One of the benefits of the language is its ability to model large systems incrementally, by directly composing simpler models of subsystems.

Dyna: A Non-Probabilistic Programming Language for Probabilistic AI

Jason Eisner, JOHN HOPKINS UNIVERSITY

The Dyna programming language is intended to provide an declarative abstraction layer for building systems in ML and AI. It extends logic programming with weights in a way that resembles functional programming. The weights are often probabilities. Yet Dyna does not enforce a probabilistic semantics, since many AI and

ML methods work with inexact probabilities (e.g., bounds) and other numeric and non-numeric quantities. Instead Dyna aims to provide a flexible abstraction layer that is “one level lower,” and whose efficient implementation will be able to serve as infrastructure for building a variety of toolkits, languages, and specific systems.

PMTK: probabilistic modeling toolkit for Matlab

Matt Dunham, UNIVERSITY OF BRITISH COLUMBIA

Kevin Murphy, UNIVERSITY OF BRITISH COLUMBIA

PMTK is a new open-source Matlab toolkit for probabilistic data modeling. It emphasizes simplicity, uniformity, and efficiency. Simplicity and uniformity are achieved by posing all learning and prediction problems as Bayesian inference. Efficiency is obtained by allowing the user to specify different kinds of posterior approximations, such as point estimates, as well as supporting many different algorithms for computing these approximate posteriors, such as fast convex optimization methods. The basic data type is “probability distribution” which can be combined together in various ways. Graphical models are treated like any other (multivariate) distribution. Non-parametric (conditional) distributions, such as Gaussian processes, are also supported.

FACTORIE: Efficient Probabilistic Programming for Relational Factor Graphs via Imperative Declarations of Structure, Inference and Learning

Andrew McCallum, UNIVERSITY MASSACHUSETTS AMHEARTS

Discriminatively trained undirected graphical models, or conditional random fields, have had wide empirical success, and there has been increasing interest in toolkits that ease their application to complex relational data. Although there has been much historic interest in the combination of logic and probability, we argue that in this mixture ‘logic’ is largely a red herring. The power in relational models is in their repeated structure and tied parameters; and logic is not necessarily the best way to define these structures. Rather than using a declarative language, such as SQL or first-order logic, we advocate using an object-oriented imperative language to express various aspects of model structure, inference and learning. By combining the traditional, declarative, statistical semantics of factor graphs with imperative definitions of their construction and operation, we allow the user to mix declarative and procedural domain knowledge, and also gain significant efficiencies. We have implemented our ideas in a system we call FACTORIE, a software library for an object-oriented, strongly-typed, functional JVM language named Scala. *Joint work with Khashayar Rohanemaneh, Michael Wick, Karl Schultz, Sameer Singh.*

Infer.NET and CSOFT

John Winn, MICROSOFT RESEARCH CAMBRIDGE

Infer.NET is a framework for performing Bayesian inference by message passing in graphical models. The first version of Infer.NET constructed in-memory factor graphs and performed message passing by traversing these graphs. This commonly-used architecture was found to have two major drawbacks: that it added significant overhead to the inference procedure and that the message passing code rapidly became complex and unmaintainable as additional algorithm features were added. In this talk, I will describe how we overcame these problems in the second version of Infer.NET by using a programming language called Csoft as our internal representation and architecting the framework as a Csoft compiler. I will show how the new architecture has allowed us to provide a rich set of inference features whilst still being efficient, flexible and easy to maintain.

Constraint-based Probabilistic Modeling

Taisuke Sato, TOKYO INSTITUTE OF TECHNOLOGY

We propose constraint-based probabilistic models defining a joint distribution $P(x \mid \text{KB})$ over possible worlds x consisting of random boolean variables which are independent but satisfy the knowledge-base KB, a set of clauses, as logical constraints. They cover (discrete) log-linear models and Bayesian networks. We also propose an EM algorithm for the parameter learning of constraint-based probabilistic models.

Probabilistic Programming for Cognitive Science

Noah Goodman, MIT

Syntactic Independence in Stochastic Functional Languages**David McCallester**, TOYOTA TECHNOLOGICAL INSTITUTE AT CHICAGO

This talk considers stochastic functional languages in general as probabilistic modeling languages. For models defined by stochastic functional programs, independence plays a central role in both exact inference and in MCMC sampling. This talk presents a nontrivial syntactic independence criterion and gives examples where the criterion greatly improves the efficiency of both exact inference and MCMC sampling.

Towards Digesting the Alphabet-Soup of Statistical Relational Learning**Luc De Raedt**, KATHOLIEKE UNIVERSITEIT LEUVEN

I will report on our work towards the development of a probabilistic logic programming environment intended as a target language in which other probabilistic languages can be compiled, thereby contributing to the digestion of the “alphabet soup” of statistical relational learning. The language combines principles of the probabilistic Prolog environments, ProbLog and PRISM, and is fully integrated in YAP-Prolog. Providing such a low-level efficient probabilistic programming language should facilitate making comparisons and experimental evaluations of different statistical relational learning approaches. *Joint work with Bart Demoen; Daan Fierens; Bernd Gutmann; Gerda Janssens; Angelika Kimmig; Niels Landwehr; Theofrastos Mantadelis; Wannes Meert; Ricardo Rocha; Vitor Santos Costa; Ingo Thon; Joost Vennekens.*

DECEMBER 13, 2008, 07:30–11:00 AND 15:30–18:45

HILTON: BLACK TUSK **WS22**

Stochastic Models of Behaviour

<http://www.eng.cam.ac.uk/~aaf23/NIPS2008workshop.html>

Aldo Faisal

UNIVERSITY OF CAMBRIDGE

Marta Gonzalez

NORTHEASTERN UNIVERSITY

aaf23@cam.ac.uk

marta.gonzalez.v@gmail.com

Abstract

The ultimate performance measure of an animal's neuronal information processing system, is whether it can produce adequate behaviour to increase its species fitness. Advances in experimental methods have vastly increased the availability, amount and quality of behavioural data for both humans and animals. Thus, a major challenge in analyzing behavior is to discover some underlying simplicity in a complex stream of behavioral actions. The gain of such an analysis is that the underlying simplicity is often a reflection of the mechanism driving behavior. Yet most behavioural studies lack adequate quantitative methods to model behaviour and its variability in a natural manner. These approaches make use of simple experiments with straightforward interpretation and subjectively defined behavioural performance indicators — often averaging out meaningful variability.

Thus, two major questions emerge.

1. How can we describe and quantify behavior such that its variability is reflected in a tractable manner?
2. What can we infer from such models of behaviour about the underlying mechanisms?

Why are Stochastic Models of Behavior going to be important? The relevance of behavioral models of humans that account for variability has straightforward application in society. From a neuroscience perspective the quantitative study of behaviour in non-humans will become of increasing relevance, because in the post-genomic age the development of genetic mutants through knock-out or knock-in of each individual gene or targeted sets of neurons will allow us to explore the link between genes, neural circuits and behavior. In fact, efforts comparable to the human genome project are already ramping up involving large scale automated behavioral measurements in both invertebrate and vertebrates genetic model organisms.

This workshop aims at engaging all participants by framing the round-table spirit of discussions with cutting-edge research talks. We much encourage participants to be active contributors to this cross-disciplinary workshop and get us together thinking about a Bioinformatics of Behavior.

7.30-11.00	Morning session
7.30-7.40	Introductions & Wake-up coffee ALDO FAISAL & MARTA GONZALES
7.40-8.30	More bits for behavior: from <i>C elegans</i> movement toward the principles of animal action GREG STEPHENS
8.30-9.20	Markov models of decision making and tool making in insects & humans ALDO FAISAL, UNIVERSITY OF CAMBRIDGE
9.20-9.30	Coffee break (Posters should be up by now)

9.30-10.30	Spotlight talks (10 minutes)
10.30-11.00	Discussion: Linking behavioral to circuit & genetic models
11.00-15.30	Break-away discussions and lunch break
15.30-18.45	Afternoon session
15.30-16.20	Understanding individual human mobility patterns MARTA GONZALES
16.20-17.10	Cross-Cultural Inference from Massive Behavioral Datasets NATHAN EAGLE, SANTA FE INSTITUTE
17.10-17.20	Coffee break
17.20-18.10	Embedding, Clustering and Matching with Graphs of GPS Behaviour TONY JEBARA
18.10-18.40	Discussion: Towards a Bioinformatics of Behavior
18.40-18.45	Closing remarks

Spotlight talks

1. “Distinct motion processing pathways inform separate aspects of locomotion in drosophila”; A. Katsov and T. Clandinin (Stanford University School of Medicine)
2. “A hidden Markov model for the directed outgrowth and turning motions of axons”; N. E. Sanjana and H. S. Seung (MIT)
3. “Classification of Animal Behavior Using Dynamic Models of Movement”; M.A. Dewar, T.C. Lukins, J.A. Heward, and J.D. Armstrong (University of Edinburgh)
4. “Modeling Social Diffusion Phenomena using Reality Mining”; Anmol Madan and Alex Pentland, (MIT)
5. “Interacting with an artificial partner: modeling the role of emotional aspects”; I. Cattinelli, M. Goldwurm, N.A. Borghese (Universita degli Studi di Milano)
6. “Learning Animal Movement Models and Location Estimates using HMMs” B. Kapicioglu, R. E. Schapire, M. Wikelski, Tamara Broderick (Princeton University, Max Planck Institute for Ornithology, University of Cambridge)
7. “Human Activity Patterns” S. Lehmann (Northeastern University)

More bits for behavior: from *C. elegans* movement toward the principles of animal action

Greg Stephens, PRINCETON UNIVERSITY

A major challenge in analyzing animal behavior is to discover some underlying simplicity in complex motor actions. Here we show that the space of shapes adopted by the nematode *C. elegans* is surprisingly low dimensional, with just four dimensions accounting for 95% of the shape variance. These “eigenworms” provide a complete, quantitative description of worm behavior, and we partially reconstruct equations of motion for the dynamics in this space. The reconstructed dynamics contain multiple attractors, revealing novel pause states and we find that the worm visits these in a rapid and almost completely deterministic response to weak thermal stimuli. Stimulus dependent correlations among the different modes suggest that one can generate more reliable behaviors by synchronizing stimuli to the state of the worm in shape space. We confirm this prediction, effectively “steering” the worm in real time.

Discrete state approaches in decision-making and tool-making behavior

A. Aldo Faisal, UNIVERSITY OF CAMBRIDGE

A hallmark animal and human behavior is the highly variable, interwoven stream of events which characterizes purposeful actions. Behavioral biologists and anthropologists typically have to rely on qualitative (and antropomorphic) descriptions obtained using un-calibrated video data. We show how such video data can be quantitatively analyzed by assuming that behavior breaks down into discrete observable states that form a symbolic sequence and where we can ignore time as an explicit factor. These behavioral sequences can then be straightforward analyzed using methods inspired from sequence analysis in bioinformatics. (Hidden) Markov Models offer a generative description of these behavioral sequences that lead to a natural visualization of behavior as behavioral networks We demonstrate this approach in two distinct applications: First, analysis of decision making in freely moving invertebrates facing a decision making task, showing that behaviors prior to the decision are direct reflections of the internal (neuronal) processing. Second, we quantify the evolution of human flint-stone tools by analyzing the 'grammatical' complexity required to reproduce tools of the past 2,000,000 years.

Understanding individual human mobility patterns

Marta Gonzales, NORTHEASTERN UNIVERSITY

Despite their importance for urban planning, traffic forecasting and the spread of biological and mobile viruses, our understanding of the basic laws governing human motion remains limited owing to the lack of tools to monitor the time-resolved location of individuals. Here we study the trajectory of 100,000 anonymized mobile phone users whose position is tracked for a six-month period. We find that, in contrast with the random trajectories predicted by the prevailing Levy flight and random walk models, human trajectories show a high degree of temporal and spatial regularity, each individual being characterized by a time independent characteristic travel distance and a significant probability to return to a few highly frequented locations. After correcting for differences in travel distances and the inherent anisotropy of each trajectory, the individual travel patterns collapse into a single spatial probability distribution, indicating that, despite the diversity of their travel history, humans follow simple reproducible patterns. This inherent similarity in travel patterns could impact all phenomena driven by human mobility, from epidemic prevention to emergency response, urban planning and agent-based modeling.

Cross-Cultural Inference from Massive Behavioral Datasets

Nathan Eagle, SANTA FE INSTITUTE

Today, behavioral data about communication, movement, and purchasing patterns is continuously being collected from the 4 billion mobile phone users around the world. The study of these large-scale behavioral data sets presents an interesting set of challenges and opportunities for the social science community. I will present telecommunications data involving almost one billion individuals worldwide, provide a broad overview of the types of research questions these data allow us to address, and discuss the tools we are developing to deal with the challenges of scale they present. Beginning with a study of 15,000 randomly sampled subjects from the US, I will discuss some preliminary results from a collaboration with John Quinn (Makerere University, Uganda) on how deviations from routines can be quantified using a standard dynamic Bayesian network technique. We incorporate a latent variable, the X-factor, switches the model between abnormal and normal modes to detect outlying behavior due to externalities such as traffic jams or natural disasters. In addition, I will present a custom software package that uses parallel binary search trees to efficiently traverse communication networks of hundreds of millions of individuals from Europe, South America, and Africa. Finally, by aggregating longitudinal data from global telecommunication companies, I will show how meta-population models may be used to establish causal relationships between mobile phone data (communication and mobility) and outcomes of interest (socio-economic status, crime, and disease prevalence), with the goal of better informing the decisions of public policy makers within both the developed and developing worlds.

Embedding, Clustering and Matching with Graphs of GPS Behavior

Tony Jebara, COLUMBIA UNIVERSITY

Many machine learning tasks can naturally be framed as problems on graphs. I first describe how to convert data into graphs using kernels and stochastic models of objects. Linking the data is then done by

using generalized matching and connectivity algorithms. Once in graph form, many learning tasks are then straightforward including dimensionality reduction, clustering and classification. I will show results applying minimum volume embedding algorithms that recover low dimensional visualizations from graphs and new kernel-clustering algorithms that partition graphs into pieces. In particular, I will describe how to build graphs from spatio-temporal location data from many GPS equipped phones and devices. One example is a graph or network of places in the city that shows similarity between different locations and how active they are right now. Another graph is the network of users showing how similar person X is to person Y by comparing their movement trails or histories. Embedding and clustering these graphs reveal interesting trends in behavior and tribes of people that are far more detailed than traditional census demographics. With machine learning algorithms applied to these human activity graphs, it becomes possible to make predictions for advertising, marketing and collaborative recommendation.

Learning from Multiple Sources

<http://web.mac.com/davidrh/LMSworkshop08/>

David Hardoon

UNIVERSITY COLLEGE LONDON

Gayle Leen

HELSINKI UNIVERSITY OF TECHNOLOGY

Samuel Kaski

HELSINKI UNIVERSITY OF TECHNOLOGY

John Shawe-Taylor

UNIVERSITY COLLEGE LONDON

D.Hardoon@cs.ucl.ac.uk

gleen@cis.hut.fi

samuel.kaski@tkk.fi

jst@cs.ucl.ac.uk

Abstract

While the machine learning community has primarily focused on analysing the output of a single data source, there has been relatively few attempts to develop a general framework, or heuristics, for analysing several data sources in terms of a shared dependency structure. Learning from multiple data sources (or alternatively, the data fusion problem) is a timely research area. Due to the increasing availability and sophistication of data recording techniques and advances in data analysis algorithms, there exists many scenarios in which it is necessary to model multiple, related data sources, i.e. in fields such as bioinformatics, multi-modal signal processing, information retrieval, sensor networks etc. The open question is to find approaches to analyse data which consists of more than one set of observations (or view) of the same phenomenon. In general, existing methods use a discriminative approach, where a set of features for each data set is found in order to explicitly optimise some dependency criterion. However, a discriminative approach may result in an *ad hoc* algorithm, require regularisation to ensure erroneous shared features are not discovered, and it is difficult to incorporate prior knowledge about the shared information. A possible solution is to overcome these problems is a generative probabilistic approach, which models each data stream as a sum of a shared component and a private component that models the within-set variation. In practice, related data sources may exhibit complex co-variation (for instance, audio and visual streams related to the same video) and therefore it is necessary to develop models that impose structured variation within and between data sources, rather than assuming a so-called ‘flat’ data structure. Additional methodological challenges include determining what is the ‘useful’ information to extract from the multiple data sources, and building models for predicting one data source given the others. Finally, as well as learning from multiple data sources in an unsupervised manner, there is the closely related problem of multitask learning, or transfer learning where a task is learned from other related tasks.

7.30 - 10.30	Morning Session
7.30-8.25	INVITED TALK: Challenges of Supervised Learning from Multiple Sources FRANCIS BACH
8.25-8.45	Multiview Clustering via Canonical Correlation Analysis K. LIVESCU, K. SRIDHARAN, S. KAKADE, AND K. CHAUDHURI
8.45-9.05	Semi-supervised Dimensionality Reduction via Canonical Correlation Analysis S. KAKADE AND D. FOSTER
9.05-9.20	Coffee (Posters should be put up at this time)

- 9.20-9.40** **The Double-Barrelled LASSO (Sparse Canonical Correlation Analysis)**
D. R. HARDOON AND J. SHAWE-TAYLOR
- 9.40-9.50** Poster Spotlights
- Learning Shared and Separate Features of Two Related Data Sets using GPLVMs**
G. LEEN AND C. FYFE
- Multiview Learning with Labels**
T. DIETHE, D. R. HARDOON AND J. SHAWE-TAYLOR
- Selective Multitask Learning by Coupling Common and Private Representations**
J. MADRID-SÁNCHEZ, E. PARRADO-HERNÁNDEZ, AND A. FIGUEIRAS-VIDAL
- Regression Canonical Correlation Analysis**
J. RUPNIK AND B. FORTUNA
- 9.50-10.30** Poster Session (can continue into lunch)
- On Asymptotic Generalization Error of Asymmetric Multitask Learning**
K. YAMAZAKI AND S. KASKI
- A Maximal Eigenvalue Method for Detecting Process Representative Genes by Integrating Data from Multiple Sources**
H. YANG, P. BHAT, H. SHANAHAN AND A. PACCANARO
- Clustering by Heterogeneous Data Fusion: Framework and Applications**
S. YU, B. DE MOOR AND Y. MOREAU
- Online Learning of Multiple Cues**
L. JIE, F. ORABONA AND B. CAPUTO
- Variational Bayes Learning from Relevant Tasks Only**
J. PELTONEN, Y. YASLAN AND S. KASKI
- Active Learning with Extremely Sparse Labeled Examples**
S. SUN
- Semantic Dimensionality Reduction for the Classification of EEG According to Musical Tonality**
T. DIETHE, S. DURRANT, J. SHAWE-TAYLOR AND H. NEUBAUER
- KCCA Based Audio-Visual Speech Recognition**
B. HALL, J. SHAWE-TAYLOR AND A. JOHNSTON
- 15.30 - 18.30** Afternoon Session
- 15.30-16.25** **INVITED TALK: Learning from Multiple Sources by Matching Their Distributions**
T. SCHEFFER
- 16.25-16.45** **GP-LVM for Data Consolidation**
C. H. EK, P. TORR, AND N. D. LAWRENCE
- 16.45-17.05** **Two-level Infinite Mixture for Multi-Domain Data**
S. ROGERS, J. SINKKONEN, A. KLAMI, M. GIROLAMI, AND S. KASKI
- 17.05-17.20** Coffee

- 17.20-17.40** **Probabilistic Models for Data Combination in Recommender Systems**
S. WILLIAMSON AND Z. GHAHRAMANI
- 17.40-18.00** **Classification from Disparate Multiple Streaming Data Sources**
A. TALUKDER AND S. HO
- 18.00-18.30** Discussion and Future Directions

INVITED TALK: Challenges of Supervised Learning from Multiple Sources

Francis Bach, DÉPARTEMENT D'INFORMATIQUE, ÉCOLE NORMALE SUPÉRIEURE

In this talk, I will consider the problem of learning a predictor from multiple sources of information, a situation common in many domains such as computer vision or bioinformatics. I will focus primarily on the multiple kernel learning framework, which amounts to consider one positive definite kernel for each source of information. Natural unanswered questions arise in this context, namely: Can one learn from infinitely many sources? Should one prefer closely related sources, or very different sources? Is it worth considering a large kernel-induced feature space as multiple sources?

Multiview Clustering via Canonical Correlation Analysis

Karen Livescu, TOYOTA TECHNOLOGICAL INSTITUTE, CHICAGO

Karthik Sridharan, TOYOTA TECHNOLOGICAL INSTITUTE, CHICAGO

Sham Kakade, TOYOTA TECHNOLOGICAL INSTITUTE, CHICAGO

Kamalika Chaudhuri, UNIVERSITY OF CALIFORNIA, SAN DIEGO

Clustering algorithms such as K-means perform poorly when the data is high-dimensional. A number of efficient clustering algorithms developed in recent years address this problem by projecting the data into a lower-dimensional subspace, e.g. via principal components analysis (PCA) or random projections, before clustering. Such techniques typically require stringent requirements on the separation between the cluster means (greater than the statistical limit). Here we present ongoing work on projection-based clustering that addresses this using multiple views of the data. We use canonical correlation analysis (CCA) to project the data in each view to a lower-dimensional subspace, under the assumption that the correlated dimensions capture the information about the cluster identities. We describe experiments on two domains, (a) speech audio and images of the speakers' faces, and (b) text and links in Wikipedia articles. We discuss several issues that arise when clustering in these domains, in particular the existence of multiple possible 'cluster variables' and of a hierarchical cluster structure.

Semi-supervised Dimensionality Reduction via Canonical Correlation Analysis

Sham Kakade, TOYOTA TECHNOLOGICAL INSTITUTE, CHICAGO

Dean Foster, UNIVERSITY OF PENNSYLVANIA

We analyze the multi-view regression problem where we have two views (X_1, X_2) of the input data and a real target variable Y of interest. In a semi-supervised learning setting, we consider two separate assumptions (one based on redundancy and the other based on (de)correlation) and show how, under either assumption alone, dimensionality reduction (based on CCA) could reduce the labeled sample complexity. The basic semi-supervised algorithm is as follows: with unlabeled data, perform CCA; with the labeled data, project the inputs onto a certain CCA subspace (i.e. perform dimensionality reduction) and then do least squares regression in this lower dimensional space. We show how, under either assumption, the number of labeled samples could be significantly reduced (in comparison to the single view setting) - in particular, we show how this dimensionality reduction only introduces little bias but could drastically reduce the variance. Under the redundancy assumption, we have that the best predictor from each view is roughly as good as the best predictor using both views. Under the uncorrelated assumption, we have that conditioned on Y the views X_1 and X_2 are uncorrelated. We show that under either of these assumptions, CCA is appropriate as a dimensionality reduction technique. We are also in the process of large scale experiments on word disambiguation (using Wikipedia, with the disambiguation pages as helping to provide labels).

The Double-Barrelled LASSO

David Hardoon, UNIVERSITY COLLEGE LONDON

John Shawe-Taylor, UNIVERSITY COLLEGE LONDON

We present a new method which solves a double-barelled LASSO in a convex least squares approach. In the presented method we focus on the scenario where one is interested in (or limited to) a primal (feature) representation for the first view while having a dual (kernel) representation for the second view. DB-LASSO minimises the number of features used in both the primal and dual projections while minimising the error (maximising the correlation) between the two views.

Learning Shared and Separate Features of Two Related Data Sets using GPLVM's

Gayle Leen, HELSINKI UNIVERSITY OF TECHNOLOGY

Colin Fyfe, UNIVERSITY OF THE WEST OF SCOTLAND

Dual source learning problems can be formulated as learning a joint representation of the data sources, where the shared information is represented in terms of a shared underlying process. However, there may be situations in which the shared information is not the only useful information, and interesting aspects of the data are not common to both data sets. Some useful features within one data set may not be present in the other and vice versa; this complementary property motivates the use of multiple data sources over single data sources which capture only one type of useful information. In this work, we present a probabilistic generative framework for analysing two sets of data, where the structure of each data set is represented in terms of a shared and private latent space. Explicitly modeling a private component for each data set avoids an oversimplified representation of the within-set variation such that the between-set variation can be modeled more accurately, as well as giving insight into potentially interesting features particular to a data set. Since two data sets may have a complex (possibly nonlinear) relationship, we use nonparametric Bayesian techniques - we define Gaussian process priors over the functions from latent to data spaces, such that each data set is modelled as a Gaussian Process Latent Variable Model (GPLVM) where the dependency structure is captured in terms of shared and private kernels.

Multiview Learning with Labels

Tom Diethe, UNIVERSITY COLLEGE LONDON

David Hardoon, UNIVERSITY COLLEGE LONDON

John Shawe-Taylor, UNIVERSITY COLLEGE LONDON

CCA can be seen as a multiview extension of PCA, in which information from two sources is used for learning by finding a subspace in which the two views are most correlated. However PCA, and by extension CCA, does not use label information. Fisher Linear Discriminant Analysis uses label information to find informative projections, which results in. We show that LDA and its dual can both be formulated as generalized eigenproblems, enabling a kernel formulation. We derive a regularised two-view equivalent of Fisher Linear Discriminant (LDA-2) and its corresponding dual (LDA-2K), both of which can also be formulated as generalized eigenproblems.

Selective Multitask Learning by Coupling Common and Private Representations

Jaisiel Madrid-Sánchez, UNIVERSIDAD CARLOS III DE MADRID

Emilio Parrado-Hernández, UNIVERSIDAD CARLOS III DE MADRID

Aníbal Figueiras-Vidal, UNIVERSIDAD CARLOS III DE MADRID

In this contribution we address the problem of selective transfer of knowledge in multitask learning for classification. In the multitask learning paradigm, we have to find a mapping and a threshold for each task in a way that all tasks interact. We assume that the relationship among tasks is expressed in the following way: all the mappings are formed by the combination of a common part, that is shared among all the tasks, and a private part, exclusive of each task. All these mappings are determined from the optimisation of the same joint functional. Selective transfer is a way of expressing the degree of task relatedness in the joint functional. If tasks are truly closely related we would like the optimisation to yield mappings with a strong shared component and small private parts. However, if the tasks are weakly related, we would expect that the private components of the mappings assume a dominant role over the shared component. In this work we discuss the introduction of these two behaviors in the joint functional through coupling parameters that affect the regularization term of the functional. The job of these parameters is to trade-off between

regularizing the common part or the private parts. We borrow this framework from a recent viewpoint of the cognitive human brain behavior.

Regression Canonical Correlation Analysis

Jan Rupnik, JOZEF STEFAN INSTITUTE

Blaz Fortuna, JOZEF STEFAN INSTITUTE

In this paper we present Regression Canonical Correlation Analysis, an extension of Canonical Correlation Analysis, where one of the dimensions is fixed and demonstrate how it can be solved efficiently. We applied the extension to the task of query translation in the context of Cross-Lingual Information Retrieval.

On Asymptotic Generalization Error of Asymmetric Multitask Learning

Keisuke Yamazaki, TOKYO INSTITUTE OF TECHNOLOGY

Samuel Kaski, HELSINKI UNIVERSITY OF TECHNOLOGY

A recent variant of multi-task learning uses the other tasks to help in learning a task-of-interest, for which there is too little training data. The task can be classification, prediction, or density estimation. The problem is that only some of the data of the other tasks are relevant or representative for the task-of-interest. It has been experimentally demonstrated that a generative model works well in this *relevant subtask learning task*. In this paper we analyze the generalization error of the model, to show that it is smaller than in standard alternatives, and to point out connections to semi-supervised learning, multi-task learning, and active learning or covariate shift.

A Maximal Eigenvalue Method for Detecting Process Representative Genes by Integrating Data from Multiple Sources

Haixuan Yang, ROYAL HOLLOWAY UNIVERSITY OF LONDON

Prajwal Bhat, ROYAL HOLLOWAY UNIVERSITY OF LONDON

Hugh Shanahan, ROYAL HOLLOWAY UNIVERSITY OF LONDON

Alberto Paccanaro, ROYAL HOLLOWAY UNIVERSITY OF LONDON

An important problem in computational biology is the identification of candidate genes which can be considered as representative of the different cellular processes taking place in the cell as it evolves through time. Multiple and very noisy data sources contain information about such processes and should therefore be integrated in order to obtain a reliable identification of such candidate genes. In this paper, we present a novel ranking algorithm which determines process representative genes by integrating a set of noisy binary relations between genes. We present some preliminary results on two artificial toy datasets and one real biological problem. In the biological problem, we use this method to identify representative genes of some of the fundamental biological mechanisms taking place during cellular growth in *A. thaliana* by integrating gene expression data and information from the gene GO annotation.

Clustering by Heterogeneous Data Fusion: Framework and Applications

Shi Yu, KATHOLIEKE UNIVERSITEIT LEUVEN

Bart De Moor, KATHOLIEKE UNIVERSITEIT LEUVEN

Yves Moreau, KATHOLIEKE UNIVERSITEIT LEUVEN

In this paper, we present a unified framework to obtain partitions from heterogeneous sources. When clustering by data fusion, the determination about the ‘relevance’ or ‘usefulness’ of the source is a vital issue to statistically guarantee a lower bound of the performance. In other words, if the clustering algorithm is able to detect the most ‘relevant’ data source, we can expect that the fusion approach works at least as good as the best individual data. In order to achieve the above objective, two different strategies are applied in the framework. The effectiveness of the clustering performance is evaluated on several experiments and applications.

Online Learning of Multiple Cues

Luo Jie, SWISS FEDERAL INSTITUTE OF TECHNOLOGY IN LAUSANNE

Francesco Orabona, SWISS FEDERAL INSTITUTE OF TECHNOLOGY IN LAUSANNE

Barbara Caputo, SWISS FEDERAL INSTITUTE OF TECHNOLOGY IN LAUSANNE

Online learning is the process by which a system learns continuously from experience, updating and enriching

its internal models. This process is one of the main reasons why cognitive agents show a robust, yet flexible capability to react to novel stimuli. Current research on artificial cognitive systems faces several issues, of which robustness and adaptability are the most challenging. Multiple-cues/sources inputs guarantee diverse and information-rich sensory data. It makes it possible to achieve higher and robust performance in varied, unconstrained settings. However, when using multiple inputs, the expansion of the input space and memory requirements are linearly proportional to the number of inputs, as well as the computational time in both training and test phase. In this work, we propose a different approach: we first sacrifice performance in favor of bounded memory growth and fast update of the solution for each separate cue. We then recover back performance by using multiple cues in the online setting. We also focus on how to learn an optimal/suboptimal combination of multiple sources in an online approach.

Variational Bayes Learning from Relevant Tasks only

Jaakko Peltonen, HELSINKI UNIVERSITY OF TECHNOLOGY

Yusuf Yaslan, ISTANBUL TECHNICAL UNIVERSITY

Samuel Kaski, HELSINKI UNIVERSITY OF TECHNOLOGY

We extend our recent work on *relevant subtask learning*, a new variant of multitask learning where the goal is to learn a good classifier for a task-of-interest with too few training samples, by exploiting ‘supplementary data’ from several other tasks. It is crucial to model the uncertainty about which of the supplementary data samples are relevant for the task-of-interest, that is, which samples are classified in the same way as in the task-of-interest. We have shown that the problem can be solved by careful *mixture modeling*: all tasks are modeled as mixtures of relevant and irrelevant samples, and the model for irrelevant samples is flexible enough so that the relevant model only needs to explain the relevant data. Previously we used simple maximum likelihood learning; now we extend the method to variational Bayes inference more suitable for high-dimensional data. We compare the method experimentally to a recent multi-task learning method and two naive methods.

Active Learning with Extremely Sparse Labeled Examples

Shiliang Sun, EAST CHINA NORMAL UNIVERSITY

An active learner usually assumes there are some labeled data available based on which a moderate classifier is learned and then examines unlabeled data to manually label the most informative examples. However, for some application domains there are only extremely sparse labeled examples, such as one labeled example per category, attainable. In this case, existing active learning methods can not successfully apply, or the inefficient way of random selection for labeling will be first implemented. In this paper, a method seeking more high-informative examples for labeling based on very limited labeled data is proposed. By investigating the correlation between different views through canonical correlation analysis, our method can launch active learning using only one labeled example from each class. Promising experimental results are presented on several applications.

Semantic Dimensionality Reduction for the Classification of EEG According to Musical Tonality

Tom Diethe, UNIVERSITY COLLEGE LONDON

Simon Durrant, UNIVERSITY OF MANCHESTER

John Shawe-Taylor, UNIVERSITY COLLEGE LONDON

Heinrich Neubauer, LEIBNIZ INSTITUTE FOR NEUROBIOLOGY, MAGDEBURG

A common structural element of Western tonal music is the change of key within a melodic sequence. The present paper examines data from a set of experiments that were conducted to analyse human perception of different modulations of key. EEG recordings were taken of participants who were given melodic sequences containing changes in key of varying distances, as well as atonal sequences, with a behavioural task of identifying the change in key. Analysis of EEG involved derivation of 122120 separate dependent variables (features), including measures such as inter-electrode spectral power, coherence, and phase. The paper presents a novel method of performing semantic dimensionality reduction based on KCCA that produces a representation enabling high accuracy identification of out-of-subject tonal versus atonal sequences.

KCCA Based Audio-Visual Speech Recognition

Benjamin Hall, UNIVERSITY COLLEGE LONDON

John Shawe-Taylor, UNIVERSITY COLLEGE LONDON

Alan Johnston, UNIVERSITY COLLEGE LONDON

We present a different approach to audio visual speech recognition, using kernel Canonical Correlation Analysis (kCCA) to correlate Mel Frequency Cepstral Coefficients (MFCC) and the processed parameters from a biological inspired optical flow algorithm: Multi Channel Gradient Model (MCGM). Utilizing these methods on a small vocabulary of spoken numbers we have found it possible to significantly decrease the word error rate of a Support Vector Machine (SVM) classifier in comparison to a classifier trained solely on MFCCs.

INVITED TALK: Learning from Multiple Sources by Matching Their Distributions

Tobias Scheffer, UNIVERSITÄT POTSDAM

My talk will address problems of transfer learning in which auxiliary data sources produce observations that are potentially helpful, but are not governed by the target distribution of the task at hand. The target distribution may only be reflected in unlabeled test data, or in a limited supply of (possibly even biased) labeled data. Such problems arise from many inspiring applications: In pattern recognition, auxiliary data may have been obtained in different geographic regions or under laboratory conditions. In therapy screening, outcomes of similar - but not identical - treatments might facilitate the learning process even though they reflect different target functions. I will discuss how the technique of distribution matching solves these problems by matching the auxiliary data to the target distribution. The transformation which performs this match is derived from a discriminative model that characterizes the discrepancy between target and auxiliary data. I will discuss recent findings on transfer learning and distribution matching, and present case studies from a number of application areas.

GP-LVM for Data Consolidation

Carl Ek, OXFORD BROOKES UNIVERSITY

Philip Torr, OXFORD BROOKES UNIVERSITY

Neil Lawrence, UNIVERSITY OF MANCHESTER

Many machine learning tasks are involved with the transfer of information from one representation to a corresponding representation or tasks where several different observations represent the same underlying phenomenon. A classical algorithm for feature selection using information from multiple sources or representations is Canonical Correlation Analysis (CCA). In CCA the objective is to select features in each observation space that are maximally correlated compared to dimensionality reduction where the objective is to re-represent the data in a more efficient form. We suggest a dimensionality reduction technique that builds on CCA. By extending the latent space with two additional spaces, each specific to a partition of the data, the model is capable of representing the full variance of the data. In this paper we suggest a generative model for shared dimensionality reduction analogous to that of CCA.

Two-level Infinite Mixture for Multi-Domain Data

Simon Rogers, UNIVERSITY OF GLASGOW

Janne Sinkkonen, HELSINKI UNIVERSITY OF TECHNOLOGY

Arto Klami, HELSINKI UNIVERSITY OF TECHNOLOGY

Mark Girolami, UNIVERSITY OF GLASGOW

Samuel Kaski, HELSINKI UNIVERSITY OF TECHNOLOGY

The combined, unsupervised analysis of coupled data sources is an open problem in machine learning. A particular important example from the biological domain is the analysis of mRNA and protein profiles derived from the same set of genes (either over time or under different conditions). Such analysis has the potential to provide a far more comprehensive picture of the mechanisms of transcription and translation than the individual analysis of the separate data sets. In this work, we present a nonparametric model for coupled data that provides an interpretable description of the shared variability in the data (as well as that that isn't shared) whilst being free of restrictive assumptions such as those found in CCA. The hierarchical model is built from two marginal mixtures (one for each representation - generalisation to three or more is straightforward). Each object will be assigned to one component in each marginal and the contingency

table describing these joint assignments is assumed to have been generated by a mixture of tables with independent margins. This top-level mixture captures the shared variability whilst the marginal models are free to capture variation specific to the respective data sources. The number of components in all three mixtures is inferred from the data using a novel Dirichlet Process (DP) formulation.

Probabilistic Models for Data Combination in Recommender Systems

Sinead Williamson, UNIVERSITY OF CAMBRIDGE

Zoubin Ghahramani, UNIVERSITY OF CAMBRIDGE

We propose a method for jointly learning multiple related matrices, and show that, by sharing information between two matrices, such an approach allows us to improve predictive performances for items where one of the matrices contains very sparse, or no, information. While the above justification has focused on recommender systems, the approach described is applicable to any two datasets that relate to a common set of items and can be represented in matrix form. Examples of such problems could include image data where each image is associated with a set of words (for example captioned or tagged images); sets of scientific papers that can be represented either using a bag-of-words representation or in terms of citation links to and from other papers; corpora of documents that exist in two languages.

Classification from Disparate Multiple Streaming Data Sources

Ashit Talukder, CALIFORNIA INSTITUTE OF TECHNOLOGY

Shen Shyang Ho, CALIFORNIA INSTITUTE OF TECHNOLOGY

We discuss a new multisource classification solution using a generative model that reduces the multiple measurement spaces into a common feature space and maintains the unique feature space for each measurement space. A single classifier is used in temporal data labeling to track correspondence between consecutive measurements from different sources in the common feature space. In addition, an auxiliary data-specific classifier is used for each data source. A transfer learning solution is then used to transfer knowledge between the common classifier (applied to the common feature spaces) and the data-specific classifier (applied to the unique feature spaces) to ensure robust classification labeling even during instances when only measurements from a weak data source is used. Experimental results on the cyclone detection and tracking problem are used to show the usefulness of our proposed approach.

Index

- Haertel, Robbie, 82
- Agarwal, Deepak, 28
 Ahuactzin, J-M, 108
 Airoldi, Edoardo, 51, 58
 Alaiz-Rodríguez, Rocío, 82
 Amari, Shun-ichi, 45
 Angelopoulos, Nikolaos, 107
 Argyriou, Andreas, 96
 Arora, Nimar, 108
 Asadi, Narges, 56
 Aukia, Janne, 53
 Auvray, Vincent, 48
 Ay, Nihat, 104
- Bach, Francis, 53, 117
 Bair, Wyeth, 41
 Balcan et al., Maria-Florina, 96
 Balcan, Doru, 49
 Balcan, Maria-Florina, 84
 Baruchi, Itay, 88
 Barzilay, Regina, 36
 Ben-David, Shai, 84, 95
 Ben-Jacob, Eshel, 88
 Bengio, Yoshua, 62
 Berger, Denise, 38, 43
 Berkes, Pietro, 39
 Bertsekas, Dimitri, 73
 Bessiere, Pierre, 108
 Bessi re, Pierre, 104
 Bethge, Matthias, 44
 Bhat, Prajwal, 119
 Bilmes, Jeff, 33
 Birlutiu, Adriana, 81
 Blaschko, Matthew, 76
 Blei, David, 36, 37, 51, 54
 Blockeel, Hendrik, 108
 Bloodgood, Michael, 82
 Blum, Avrim, 84
 Borchert, Thomas, 26
 Borgelt, Christian, 43
 Borgwardt, Karsten, 74
 Bottou, Leon, 73
 Botvinick, Matthew, 64
 Bouchard-Cote, Alexandre, 37
 Boyd-Graber, Jordan, 36, 54
 Braun, Mikio, 67
 Brenner, Steven, 23
 Briggs, Peter, 29
- Brown, Emery, 45
 Burges, Chris, 28
 B sling, Lars, 104
- Caetano, Tiberio, 50
 Caputo, Barbara, 119
 Carin, Lawrence, 79, 80
 Carroll, James, 82
 Carroll, James L., 82
 Carroll, Melissa K., 86, 87
 Cecchi, Guillermo, 86–88
 Chang, Edward, 28
 Chang, Jonathan, 54
 Chapelle et al., Olivier, 95
 Chapelle, Olivier, 26
 Chaudhuri, Kamalika, 117
 Chechik, Gal, 22, 59
 Churchland, Mark, 41
 Cid-Sueiro, Jes s, 82
 Cohen, Shay, 36
 Cohen, William, 52
 Collins, Michael, 36
 Collobert, Ronan, 68
 Cortes et al., Corinna, 96
 Cortes, Corinna, 94
 Craven, Mark, 81
 Cunningham, John, 41
- D. Piatko, Christine, 82
 Dallmeier, Valentin, 55
 Das, Sammay, 29
 Daw, Nathaniel, 61
 De Moor, Bart, 119
 De Raedt, Luc, 110
 de Salvo Braz, Rodrigo, 107, 108
 DeNero, John, 37
 Deng, Li, 31
 Diard, Julien, 104
 Diesmann, Markus, 43
 Diethe, Tom, 118, 120
 Dietz, Laura, 55
 Dobra, Adrian, 48
 Domingos, Pedro, 29, 33, 107
 Douglas, Rodney, 91
 Dragoi, Valentin, 38, 42
 Dudi k, Miroslav, 107
 Dunham, Matt, 109
 Durrant, Simon, 120
- Eagle, Nathan, 113

- Eaton, John W., 68
 Eisenstein, Jacob, 36
 Eisner, Jason, 82, 108
 Ek, Carl, 121
 Engel, Yaakov, 101
 Evans, Michael, 56

 Faisal, A. Aldo, 113
 Faisal, Aldo, 111
 Fienberg, Stephen, 52, 56
 Fienberg, Stephen E., 48
 Figueiras-Vidal, Aníbal, 118
 Fortuna, Blaz, 119
 Foster, Dean, 117
 Francisco, Alexandre, 54
 Freeman, Bill, 21
 Freeman, William, 19
 Friedland, Lewis, 81
 Fyfe, Colin, 118

 Gaertner, Thomas, 75
 Garg, Rahul, 88
 Gehler, Peter, 97
 Gerstein, George, 43
 Ghahramani, Zoubin, 122
 Ghavamzadeh, Mohammad, 101
 Ghosh, Joydeeo, 81
 Gilet, Estelle, 104
 Girolami, Mark, 121
 Glasmachers, Tobias, 68
 Globerson, Amir, 92
 Goldwater, Sharon, 36
 Gomes, Carla, 76
 Gonzales, Marta, 113
 Gonzalez, Marta, 111
 Goodman, Noah, 62, 109
 Gordon, Geoff, 107
 Graf, Arnulf, 42
 Gretton, Arthur, 94
 Griffiths, Tom, 61
 Gross, Jonathan, 49
 Gruber, Amit, 57
 Grün, Sonja, 43, 45
 Grünewälder, Steffen, 38, 41
 Guerrero-Currieses, Alicia, 82
 Guestrin, Carlos, 50
 Guibas, Leonidas J., 49, 50
 Guinney, Justin, 56
 Guo, Yunsong, 76
 Gureckis, Todd, 65
 Guyon, Isabelle, 77, 96

 Haertel, Robbie, 81
 Haghghi, Aria, 35
 Hall, Benjamin, 121
 Hammerstrom, Dan, 99
 Hansen, Lars Kai, 87
 Hardoon, David, 115, 118
 Harik, Georges, 107
 He, Xiaodong, 31
 Heeger, David, 91
 Hendler, Talma, 88
 Heskes, Tom, 81, 88
 Hill, Jeremy, 70
 Hinton, Geoffrey, 32
 Hofman, Jake, 51, 58
 Hsu, Chun-Nan, 56
 Huang, Bert, 55
 Huang, Jonathan, 50
 Huang, Yi, 75
 Huang, Yi-Hung, 55
 Huh, Seungil, 56
 Hunter, John D., 69

 Jaakkola, Tommi, 92
 Jacob, Yael, 88
 Jang, Gun Ho, 56
 Janzing, Dominik, 77
 Jebara, Tony, 51, 55, 57, 113
 Jenison, Rick, 40
 Jiang, Xiaoye, 49, 50
 Jie, Luo, 119
 Joachims, Thorsten, 30, 76
 Johnson, Todd, 108
 Johnston, Alan, 121
 Jun, Goo, 81
 Jung, Tobias, 104
 Jurman, Giuseppe, 69

 Kafri, Michal, 88
 Kakade, Sham, 117
 Kanani, Pallika, 81
 Karatzoglou, Alexandros, 69
 Karlis, Dimitris, 45
 Kaski, Samuel, 53, 115, 119–121
 Keating, Peter, 41
 Kersting, Kristian, 54
 Kiselyov, Oleg, 108
 Klami, Arto, 121
 Klein, Dan, 35
 Kleinfeld, David, 25
 Kloft et al., Marius, 97
 Kohn, Adam, 44
 Kolar, Mladen, 57

- Kondor, Risi, 17, 47
 Kriegel, Hans-Peter, 75
 Krishnapuram, Balaji, 79
 Krivitsky, Pavel, 53
 Ku, Tien-Chuan, 55

 Lampert, Christoph, 76
 Lanckriet, Gert, 94
 Lawrence, Neil, 121
 Lebanon, Guy, 17, 47
 Leen, Gayle, 115, 118
 Lengauer, Thomas, 23
 Lengyel, Mate, 63
 Leskovec, Jure, 52
 Leslie, Christina, 22, 59
 Liang, Percy, 35
 Lim, Lek-Heng, 49
 Lin, Chung-Chi, 55
 Lin, Yu-Shi, 55
 Lippert, Christoph, 75
 Littman, Michael, 64
 Liu, Alexander, 81
 Liu, Yanxi, 18
 Livescu, Karen, 117
 Lloyd, John, 108
 Losonczy, Attila, 91
 Lunagomez, Simon, 56

 Müller, Klaus-Robert, 48
 Maass, Wolfgang, 104
 Macke, Jakob, 44
 Madrid-Sánchez, Jaisiel, 118
 Madsen, Kristoffer Hougaard, 87
 Maes, Francis, 69
 Magdon-Ismael, Malik, 29
 Maggioni, Mauro, 56
 Malave, Vicente L., 89
 Mannor, Shie, 101
 Mansinghka, Vikash, 105, 107
 Marbukh, Vladimir, 57
 Mariadassou, Mahendra, 57
 Masuda, Naoki, 40
 Matusik, Wojciech, 20
 Mazer, E, 108
 McAllester, David, 105
 McAuley, Julian, 50
 McCallester, David, 110
 McCallum, Andrew, 36, 54, 109
 McDonald, Austin, 107
 Meert, Wannes, 108
 Meila, Marina, 18
 Meinecke, Frank C., 48

 Mekhnacha, K, 108
 Melville, Prem, 81
 Mihalkova, Lilyana, 29
 Mimno, David, 54
 Mitchell, Tom, 99
 Mjolsness, Eric, 108
 Mohri, Mehryar, 94
 Mooij, Joris, 70
 Mooney, Raymond, 29
 Moore, Robert, 34
 Moreau, Yves, 119
 Morris, Quaid, 22, 55, 59
 Morton, Jason, 17, 18, 47
 Morup, Morten, 87
 Mostafavi, Sara, 55
 Mozer, Michael, 65
 Mukherjee, Sayan, 56
 Murphy, Kevin, 109

 Najm, Tarek, 30
 Neubauer, Heinrich, 120
 Ng, Kee Siong, 108
 Niv, Yael, 25, 64
 Noble, William, 22, 59
 Nowozin, Sebastian, 72

 Obermayer, Klaus, 38, 42
 Ong, Cheng Soon, 67
 Onken, Arno, 38, 41
 Opper, Manfred, 44
 Orabona, Francesco, 119
 Ostendorf, Mari, 33

 Püschel, Markus, 49
 Paccanaro, Alberto, 119
 Pan, Junfeng, 26
 Parkkinen, Juuso, 53
 Parrado-Hernández, Emilio, 118
 Pelckmans, Kristiaan, 84
 Peltonen, Jaakko, 120
 Pereira, Francisco, 86, 87
 Petrov, Slav, 35
 Petrović, Sonja, 48
 Petterson, James, 50
 Phillips, Andrew, 108
 Poggio, Tomaso, 90, 91
 Polani, Daniel, 24, 103, 104
 Poole, David, 107
 Poupart, Pascal, 101

 Quiñonero Candela, Joaquin, 26

 Rättsch, Gunnar, 22

- Raeder, Troy, 70
 Raetsch, Gunnar, 59
 Ramirez, Rafael, 89
 Rao, A. Ravi, 88
 Rao, Bharat, 79
 Rapson, Amir, 88
 Raskar, Ramesh, 20
 Reckow, Stefan, 75
 Regev, Aviv, 22
 Reutemann, Peter, 71
 Rinaldo, Alessandro, 48
 Ringger, Eric K., 81
 Rish, Irina, 56, 86, 88
 Robin, Stephane, 57
 Rogers, Simon, 121
 Rosen-Zvi, Michal, 57
 Rostamizadeh, Afshin, 94
 Roy, Daniel, 105
 Rupnik, Jan, 119
 Russell, Stuart, 106, 108
 Ryu, Stephen, 41

 Sahani, Maneesh, 41
 Sanborn, Adam, 63
 Sandryhaila, Aliaksei, 49
 Santhanam, Gopal, 41
 Sanz, Patricia, 89
 Sato, Taisuke, 109
 Schölkopf, Bernhard, 19, 77
 Scheffer, Tobias, 121
 Scheinberg, Katya, 56
 Schlüter, Ralf, 32
 Schubert, Matthias, 75
 Schwaighofer, Anton, 26
 Schwarz, Michael, 29
 Seitz, Steve, 20
 Sejnowski, Terrence, 90
 Sejnowski, Terry, 91
 Seppi, Kevin, 82
 Seppi, Kevin D., 81
 Settles, Burr, 81
 Seung, Sebastian, 91
 Shan, Chung-chieh, 108
 Shanahan, Hugh, 119
 Shaw, Blake, 57
 Shawe-Taylor, John, 80, 84, 97, 115, 118, 120, 121
 Shazeer, Noam, 107
 Sheldon, Daniel, 75
 Shenoy, Krishna, 41
 Shervashidze, Nino, 75
 Shimazaki, Hideaki, 45
 Shyang Ho, Shen, 122

 Sinkkonen, Janne, 53
 Sinkkonen, Janne, 121
 Siskind, Jeffrey Mark, 107
 Smith, Noah, 36
 Smyth, Barry, 29
 Smyth, Clinton, 107
 Sonnenburg, Soeren, 67
 Sontag, David, 92
 Sra, Suvrit, 72
 Srebro et al., Nathan, 97
 Sridharan, Karthik, 117
 Stafford Noble, William, 95
 Stephens, Greg, 112
 Steyvers, Mark, 63
 Stolcke, Andreas, 34
 Struyf, Jan, 108
 Sudderth, Erik, 108
 Sun, Shiliang, 120
 Surendran, AC, 30

 Takashima, Atusko, 88
 Talukder, Ashit, 122
 Tanner, Brian, 70
 Tenenbaum, Josh, 61, 63
 Tenenbaum, Joshua, 105
 Thibadeau, Robert, 99
 Thirion, Bertrand, 88
 Thrun, Sebastian, 25
 Tiesinga, Paul, 91
 Tishby, Naftali, 24, 103, 104
 Toronto, Neil, 82
 Torr, Philip, 121
 Touretzky, David, 99
 Tresp, Volker, 54, 75, 80
 Tsai, Yuh-Show, 55
 Tsuda, Koji, 74, 75
 Tuulos, Ville, 69

 Uther, Will, 108

 Vacher, Corrine, 57
 Vaddadi, Phani, 30
 van Gerven, Marcel, 88
 Vanschoren, Joaquin, 71
 Vembu, Shankar, 75
 Vert, Jean-Philippe, 53
 Vijay-Shanker, K., 82
 Vishwanathan, S V N, 72, 74
 von Bünau, Paul, 48

 Wallach, Hanna, 54
 Wang, Chong, 36

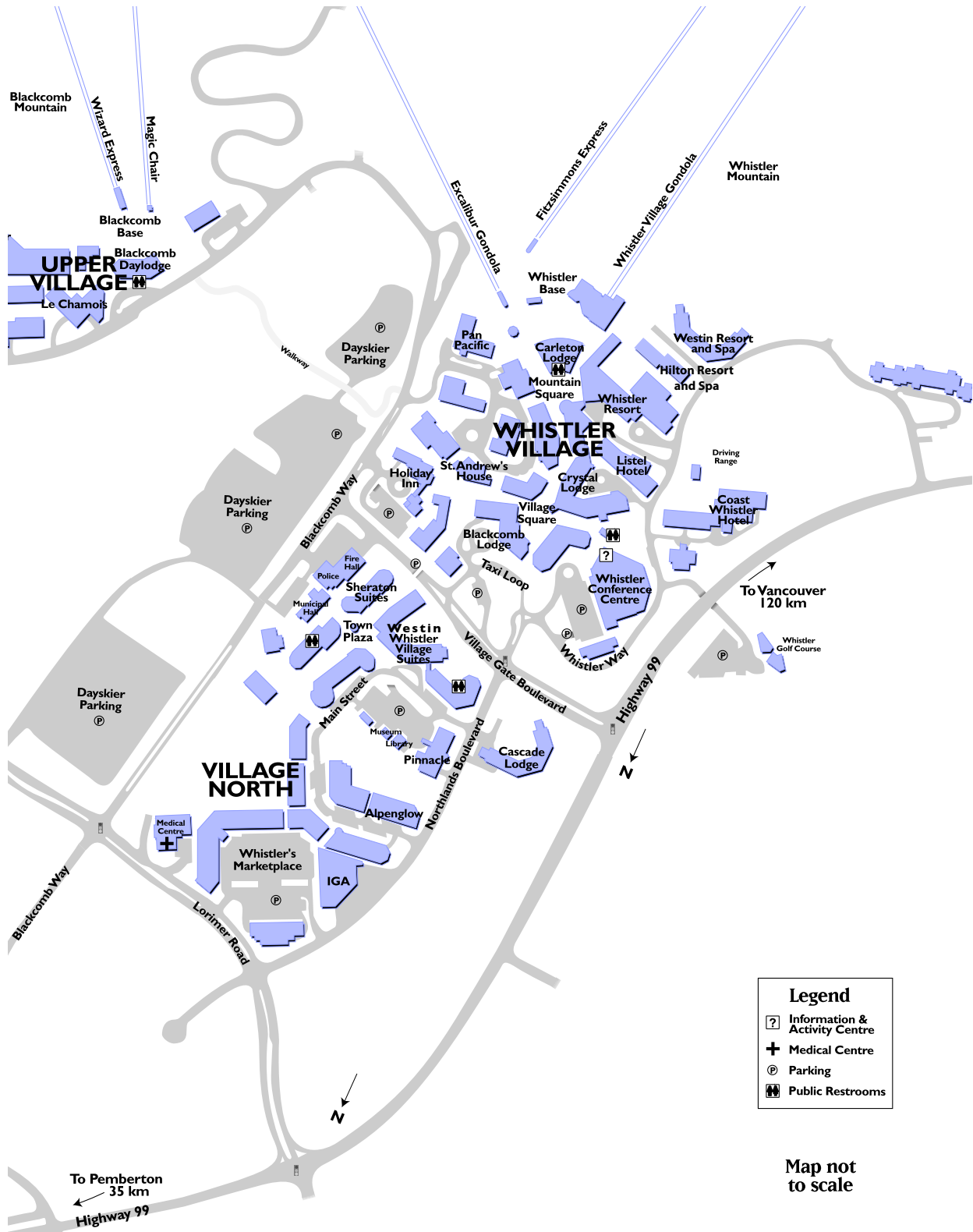
Watanabe, Sumio, 48
Weber, Stefan-Hagen, 75
Wehenkel, Louis, 48
Weiss, Yair, 19, 57
Widrow, Bernie, 91
Williamson, Sinead, 122
Winn, John, 105, 109
Winter, Stefan, 104
Wolfe, Patrick, 52
Wright, Stephen, 73

Xing, Eric, 51, 57
Xu, Zhao, 54

Yakhnenko, Oksana, 79
Yamazaki, Keisuke, 48, 119
Yan, Xifeng, 74
Yang, Haixuan, 119
Yaslan, Yusuf, 120
Yu, Byron, 41
Yu, Chun-Na, 76
Yu, Kai, 54
Yu, Shi, 119
Yu, Shipeng, 79
Yue, Yisong, 30
Yuille, Alan, 63

Zaidan, Omar, 82
Zaslaviskiy, Mikhail, 53
Zettlemoyer, Luke, 36
Zhang, Jun, 65
Zhao, Yunxin, 33
Zhu, Shenghuo, 55
Zhu, Xiaojin, 61, 64
Zito, Tiziano, 69
Zoeter, Onno, 30

Whistler Map

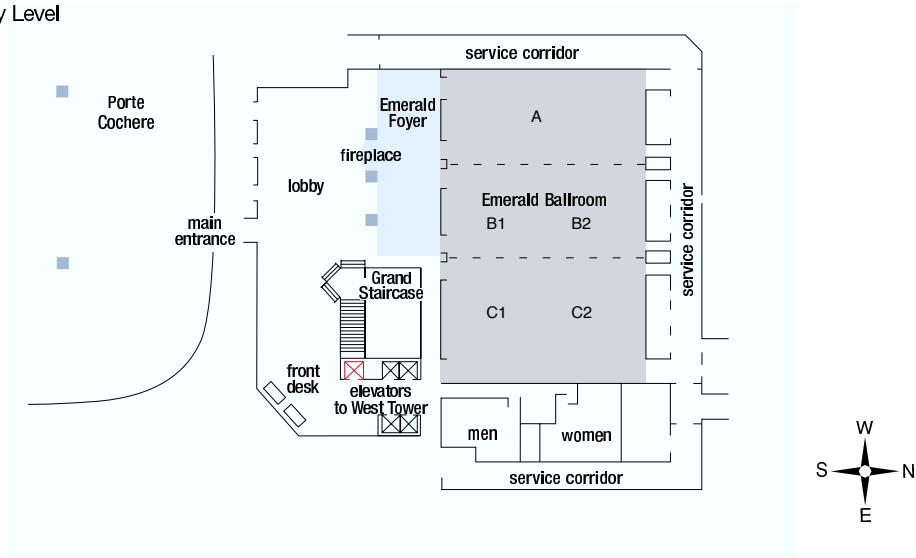


Legend	
	Information & Activity Centre
	Medical Centre
	Parking
	Public Restrooms

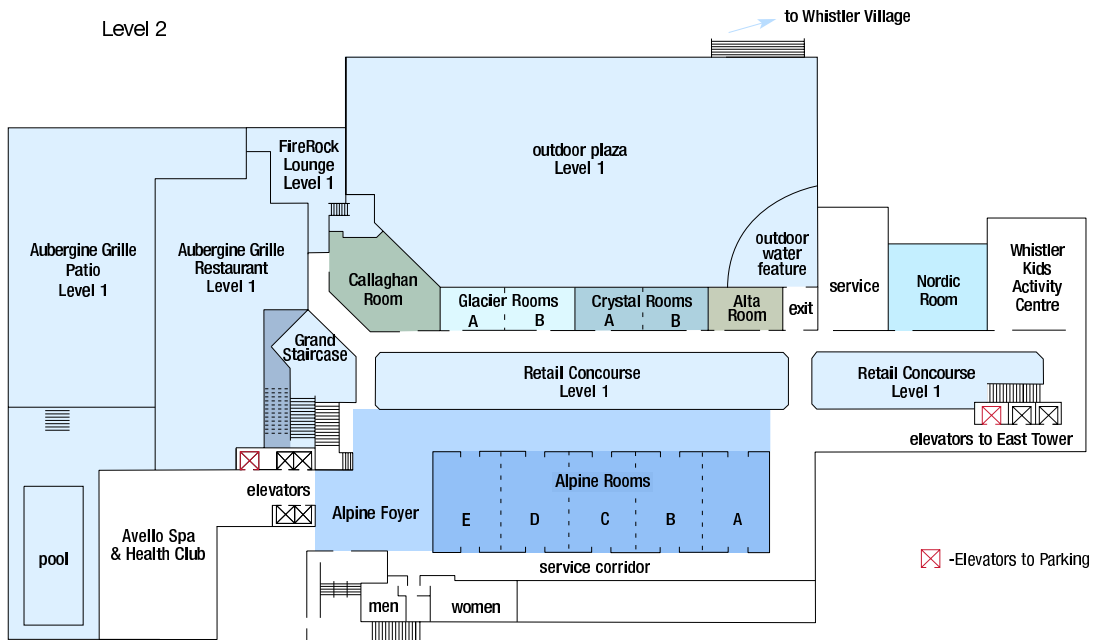
Map not to scale

Westin Resort Workshop Rooms

Lobby Level

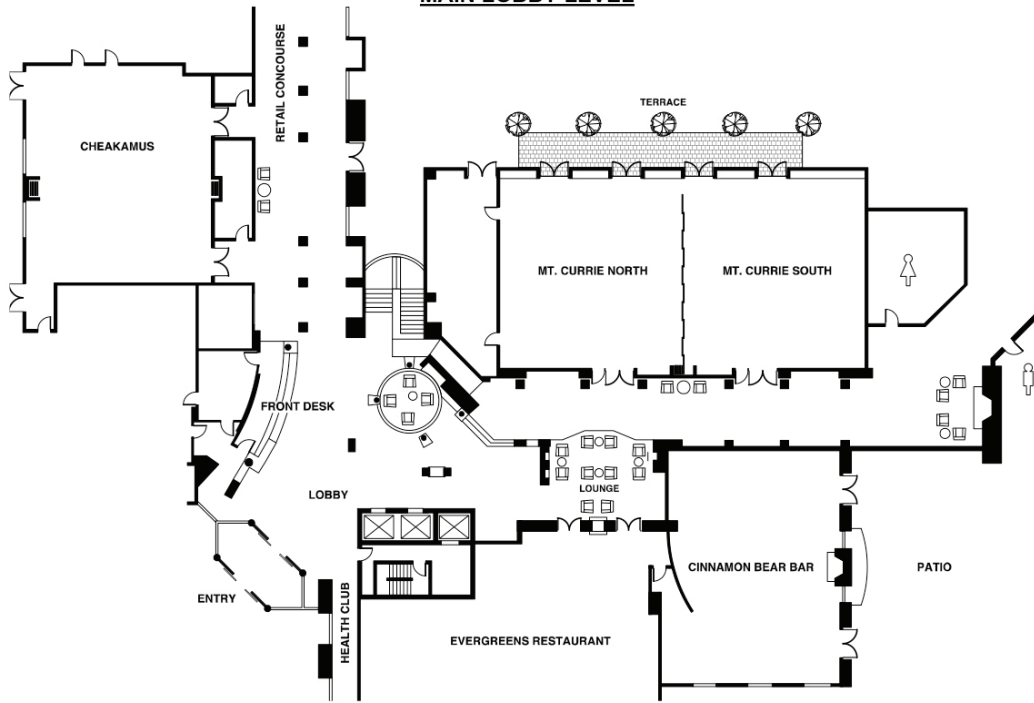


Level 2



Hilton Resort Workshop Rooms

Hilton Whistler Resort MAIN LOBBY LEVEL



LOWER LEVEL

