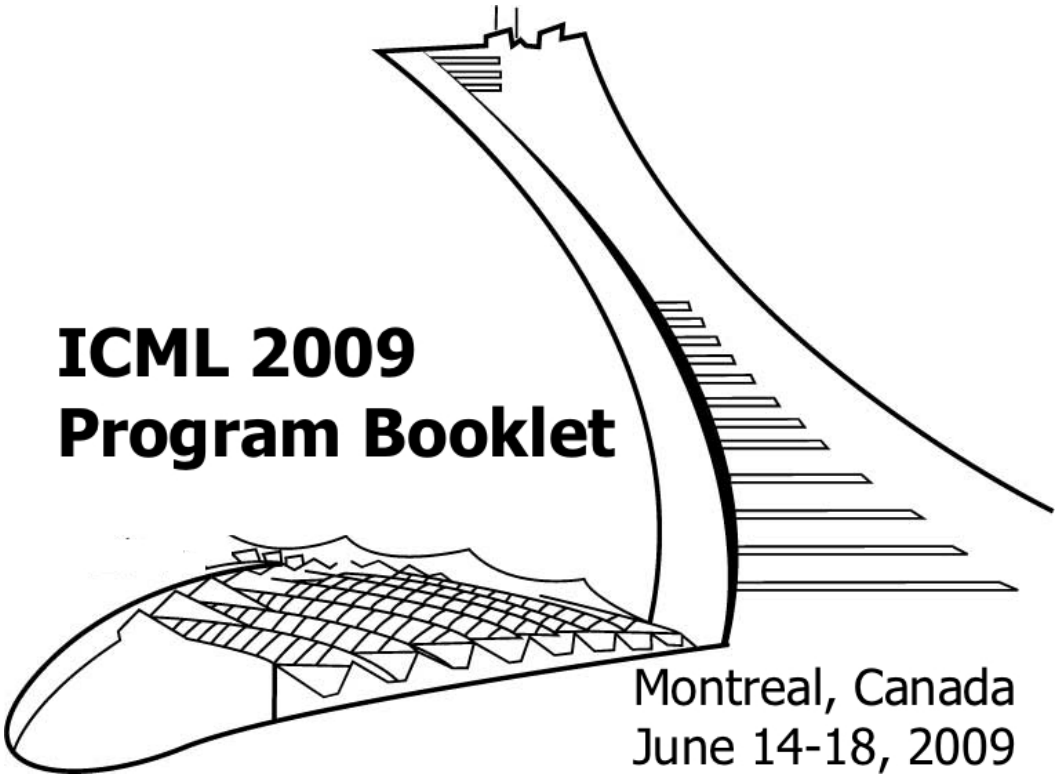


ICML 2009
Program Booklet

Montreal, Canada
June 14-18, 2009

Contents

Welcome to ICML 2009!

Information

Registration: The registration desk is open from 8 a.m. to 6 p.m., Sunday to Wednesday, in Leacock/Arts.

Conference Sessions are two hours long (either 10:20 a.m. to 12:20 p.m. or 2:00 p.m. to 4:00 p.m.) and each contain five talks. Each talk is allocated 20 minutes plus 4 minutes for questions and speaker transition. Each talk will also have an accompanying poster in one of the poster sessions to permit follow-up conversations.

Poster Sessions occur on Tuesday and Wednesday evening. Papers from sessions 1A to 3F have posters on Tuesday, and papers from sessions 4A to 6E have posters on Wednesday. The poster sessions will also include posters by student scholarship recipients. Sandwiches and a hot buffet will be served at both poster sessions.

Message Boards are located behind the registration desk and at the end of the hallway in Leacock/Arts.

Internet/E-mail Access: Free wireless access is available; login and password information are provided at registration. In addition, computers with Internet access are available upon request.

Food

Coffee Breaks are held just outside the conference rooms.

Lunches are on your own. See local information handout for dining suggestions.

The ICML Banquet is on Monday, June 15, from 6:45 to 11:00 p.m. at the Montréal Science Centre, 2 de La Commune Ouest, 2nd floor. See separate handout for information on how to get there.

Tutorials and Workshops

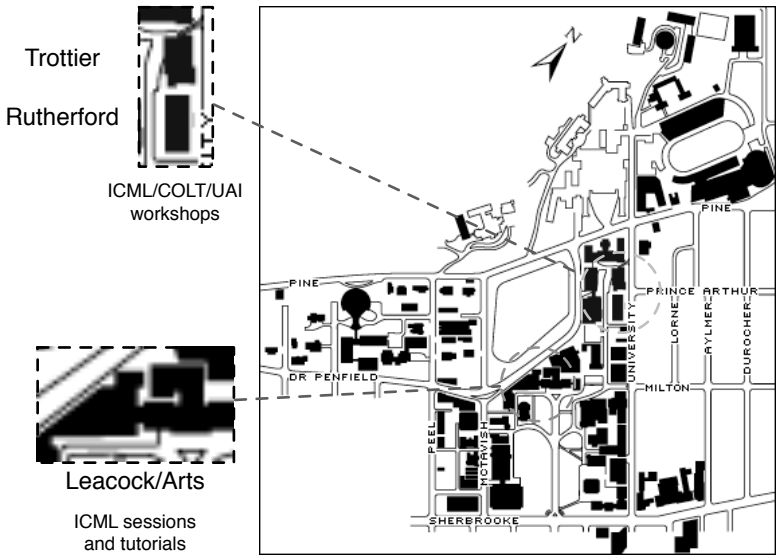
Tutorials take place on Sunday, June 14 in Leacock. They are 2.5 hours long (see p. 47 for schedule).

Workshops take place on Thursday, June 18, in the Trottier and Rutherford buildings. Please see the separate workshop handout for information.

Other Events

The ICML Business Meeting occurs on Wednesday, June 17, from 4:30 to 5:30 p.m. in Leacock 132. It is open to all attendees.

Exhibitors are located in the hallway along with registration, and their hours are 8 a.m. to 6 p.m. each day.



Conference Overview

Sunday, June 14		
9:00 – 18:30	Tutorials	Leacock

Monday, June 15		
9:30 – 9:50	Welcome	Leacock 132
9:50 – 10:20	*Coffee break*	
10:20 – 12:20	**Session 1**	Leacock/Arts
12:20 – 14:00	*Lunch break*	
14:00 – 16:00	**Session 2**	Leacock/Arts
16:00 – 16:30	*Coffee break*	
16:30 – 17:50	**Invited Talk: Yoav Freund** **Drifting games, boosting and online learning**	Leacock 132
18:45 – 23:00	Banquet	Montréal Science Centre

Tuesday, June 16		
8:30 – 9:50	**Invited Talk: Corinna Cortes** **Can learning kernels help performance?**	Leacock 132
9:50 – 10:20	*Coffee break*	
10:20 – 12:20	**Session 3**	Leacock/Arts
12:20 – 14:00	*Lunch break*	
14:00 – 16:00	**Session 4**	Leacock/Arts
16:00 – 16:30	*Coffee break*	
16:30 – 17:30	**Awards Session**	Leacock 132
18:45 – 22:30	Poster session: Papers from sessions 1A to 3F	Leacock/Arts

Wednesday, June 17		
8:30 – 9:50	**Invited Talk: Emmanuel Dupoux** **How do infants bootstrap into spoken language?: Models and challenges**	Leacock 132
9:50 – 10:20	*Coffee break*	
10:20 – 12:20	**Session 5**	Leacock/Arts
12:20 – 14:00	*Lunch break*	
14:00 – 16:00	**Session 6**	Leacock/Arts
16:00 – 16:30	*Coffee break*	
16:30 – 17:30	ICML business meeting	Leacock 132
18:45 – 22:30	Poster Session: Papers from sessions 4A to 6E	Leacock/Arts

Thursday, June 18		
Variable	Workshops	Trottier and Rutherford

Monday, June 15th

9:30 – 9:50	Welcome: Leacock 132	
9:50 – 10:20	*Coffee break*	
10:20 – 12:20	**1A: Clustering Geometry** (p. 14) **Leacock 132** Chair: Kiri Wagstaff	**1B: Applied Probabilistic Models** (p. 15) **Leacock 219** Chair: Florence d'Alche Buc
	Solution Stability in Linear Programming Relaxations: Graph Partitioning and Unsupervised Learning *Nowozin and Jegelka*	Unsupervised Hierarchical Modeling of Locomotion Styles *Pan and Torresani*
	♣ A Scalable Framework for Discovering Coherent Co-Clusters in Noisy Data *Deodhar, Ghosh, Gupta, Cho, and Dhillon*	Exploiting Sparse Markov *and* Covariance Structure in Multiresolution Models *Choi, Chandrasekaran, and Willsky*
	Multi-View Clustering via Canonical Correlation Analysis *Chaudhuri, Kakade, Livescu, and Sridharan*	A Bayesian Approach to Protein Model Quality Assessment *Kamisetty and Langmead*
	Spectral Clustering Based on the Graph p-Laplacian *Bühler and Hein*	Multi-Class Image Segmentation using Conditional Random Fields and Global Classification *Plath, Toussaint, and Nakajima*
	Nearest Neighbors in High-Dimensional Data: The Emergence and Influence of Hubs *Radovanović, Nanopoulos, and Ivanović*	GAODE and HAODE: Two Proposals Based on AODE to Deal with Continuous Variables *Flores, Gámez, Martínez, and Puerta*
12:20 – 14:00	*Lunch break*	
14:00 – 16:00	**2A: Graphs and Embeddings** (p. 19) **Leacock 219** Chair: Kilian Weinberger	**2B: Gaussian Processes** (p. 20) **Leacock 15** Chair: John Guiver
	Fitting a Graph to Vector Data *Daitch, Kelner, and Spielman*	Non-Linear Matrix Factorization with Gaussian Processes *Lawrence and Urtasun*
	⋆ ♣ Structure Preserving Embedding *Shaw and Jebara*	Analytic Moment-Based Gaussian Process Filtering *Deisenroth, Huber, and Hanebeck*
	Graph Construction and b-Matching for Semi-Supervised Learning *Jebara, Wang, and Chang*	Function Factorization using Warped Gaussian Processes *Schmidt*
	Partial Order Embedding with Multiple Kernels *McFee and Lanckriet*	♣ Tractable Nonparametric Bayesian Inference in Poisson Processes with Gaussian Process Intensities *Adams, Murray, and MacKay*
	Probabilistic Dyadic Data Analysis with Local and Global Consistency *Cai, Wang, and He*	Large-Scale Collaborative Prediction Using a Nonparametric Random Effects Model *Yu, Lafferty, and Zhu*
16:00 – 16:30	*Coffee break*	
16:30 – 17:50	**Invited Talk: Yoav Freund, Leacock 132** **Drifting games, boosting and online learning** (p. 12)	
18:45 – 23:00	Banquet at Montréal Science Centre	

⋆: Nominated for Best Paper
♠: Nominated for Best Application Paper
♣: Nominated for Best Student Paper

Monday, June 15th

Welcome: Leacock 132

Coffee break

1C: Exploration in RL (p. 16)
Leacock 232
Chair: Pascal Poupart

The Adaptive k-Meteorologists Problem and Its Application to Structure Learning and Feature Selection in Reinforcement Learning
Diuk, Li, and Leffler

Near-Bayesian Exploration in Polynomial Time
Kolter and Ng

Optimistic Initialization and Greediness Lead to Polynomial Time Learning in Factored MDPs
Szita and Lörincz

Dynamic Analysis of Multiagent Q-learning with ϵ-greedy Exploration
Gomes and Kowalczyk

Hoeffding and Bernstein Races for Selecting Policies in Evolutionary Direct Policy Search
Heidrich-Meisner and Igel

1D: Online Learning (p. 17)
Leacock 15
Chair: Claudio Gentile

A Simpler Unified Analysis of Budget Perceptrons
Sutskever

Efficient Learning Algorithms for Changing Environments
Hazan and Seshadhri

Online Learning by Ellipsoid Method
Yang, Jin, and Ye

Learning Prediction Suffix Trees with Winnow
Karampatziakis and Kozen

Identifying Suspicious URLs: An Application of Large-Scale Online Learning
Ma, Saul, Savage, and Voelker

1E: Ranking (p. 18)
Leacock 26
Chair: Marie desJardins

♣ BoltzRank: Learning to Maximize Expected Ranking Gain
Volkovs and Zemel

Decision Tree and Instance-Based Learning for Label Ranking
Cheng, Hühn, and Hüllermeier

Ranking with Ordered Weighted Pairwise Classification
Usunier, Buffoni, and Gallinari

Ranking Interesting Subgroups
Rueping

Generalization Analysis of Listwise Learning-to-Rank Algorithms
Lan, Liu, Ma, and Li

Lunch break

2C: Dynamical Systems (p. 21)
Leacock 232
Chair: Doina Precup

Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems
Song, Huang, Smola, and Fukumizu

Learning Nonlinear Dynamic Models
Langford, Salakhutdinov, and Zhang

Optimized Expected Information Gain for Nonlinear Dynamical Systems
Busetto, Ong, and Buhmann

Learning Linear Dynamical Systems without Sequence Information
Huang and Schneider

Dynamic Mixed Membership Block Model for Evolving Networks
Fu, Song, and Xing

2D: Kernels (p. 22)
Leacock 26
Chair: Dale Schuurmans

Route Kernels for Trees
Aiolli, Da San Martino, and Sperduti

The Graphlet Spectrum
Kondor, Shervashidze, and Borgwardt

Multi-Instance Learning by Treating Instances As Non-I.I.D. Samples
Zhou, Sun, and Li

Non-Monotonic Feature Selection
Xu, Jin, Ye, Lyu, and King

Regression by Dependence Minimization and its Application to Causal Inference
Mooij, Janzing, Peters, and Schölkopf

2E: Learning Codebooks and Dictionaries (p. 23)
Leacock 132
Chair: Samy Bengio

Gradient Descent with Sparsification: An Iterative Algorithm for Sparse Recovery with Restricted Isometry Property
Garg and Khandekar

Learning Dictionaries of Stable Autoregressive Models for Audio Scene Analysis, *Cho and Saul*

Online Dictionary Learning for Sparse Coding
Mairal, Bach, Ponce, and Sapiro

Learning Non-Redundant Codebooks for Classifying Complex Objects
Zhang, Surve, Fern, and Dietterich

Prototype Vector Machine for Large Scale Semi-supervised Learning
Zhang, Kwok, and Parvin

Coffee break

Invited Talk: Yoav Freund, Leacock 132
Drifting games, boosting and online learning (p. 12)

Banquet at Montréal Science Centre

Tuesday, June 16th

8:30 – 9:50	**Invited Talk: Corinna Cortes, Leacock 132** **Can learning kernels help performance?** (p. 12)		
9:50 – 10:20	*Coffee break*		
10:20 – 12:20	**3A: Clustering Algorithms** (p. 24) **Leacock 15** Chair: Tony Jebara	**3B: Inference in Probabilistic Models** (p. 25) **Leacock 219** Chair: Kevin Murphy	**3C: RL with Temporal Differences** (p. 26) **Leacock 232** Chair: Shimon Whiteson
	Multi-Assignment Clustering for Boolean Data *Streich, Frank, Basin, and Buhmann*	Orbit-Product Representation and Correction of Gaussian Belief Propagation *Johnson, Chernyak, and Chertkov*	Proto-Predictive Representation of States with Simple Recurrent Temporal-Difference Networks *Makino*
	K-means in Space: A Radiation Sensitivity Evaluation *Wagstaff and Bornstein*	Convex Variational Bayesian Inference for Large Scale Generalized Linear Models *Nickisch and Seeger*	Regularization and Feature Selection in Least Squares Temporal-Difference Learning *Kolter and Ng*
	Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary? *Nguyen, Epps, and Bailey*	★ ♣ Archipelago: Nonparametric Bayesian Semi-Supervised Learning *Adams and Ghahramani*	Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation *Sutton, Maei, Precup, Bhatnagar, Silver, Szepesvári, and Wiewiora*
	Fast Evolutionary Maximum Margin Clustering *Gieseke, Pahikkala, and Kramer*	The Bayesian Group-Lasso for Analyzing Contingency Tables *Raman, Fuchs, Wild, Dahl, and Roth*	Kernelized Value Function Approximation for Reinforcement Learning *Taylor and Parr*
	Discriminative k Metrics *Szlam and Sapiro*	Split Variational Inference *Bouchard and Zoeter*	Constraint Relaxation in Approximate Linear Programs *Petrik and Zilberstein*
12:20 – 14:00	*Lunch break*		
14:00 – 16:00	**4A: Weak Supervision** (p. 30) **Leacock 26** Chair: Fei Sha	**4B: Learning Structures** (p. 31) **Leacock 132** Chair: Pedro Domingos	**4C: Active Learning** (p. 32) **Leacock 219** Chair: Yoav Freund
	Semi-Supervised Learning Using Label Mean *Li, Kwok, and Zhou*	Structure Learning with Independent Non-Identically Distributed Data *Tillman*	Learning to Segment from a Few Well-Selected Training Images *Farhangfar, Greiner, and Szepesvári*
	Partially Supervised Feature Selection with Regularized Linear Models *Helleputte and Dupont*	Structure Learning of Bayesian Networks using Constraints *de Campos, Zeng, and Ji*	Importance Weighted Active Learning *Beygelzimer, Dasgupta, and Langford*
	★ Optimal Reverse Prediction: A Unified Perspective on Supervised, Unsupervised and Semi-supervised Learning, *Xu, White, and Schuurmans*	Learning Structurally Consistent Undirected Probabilistic Graphical Models *Roy, Lane, and Werner-Washburne*	Learning from Measurements in Exponential Families *Liang, Jordan, and Klein*
	Supervised Learning from Multiple Experts: Whom to Trust When Everyone Lies a Bit *Raykar, Yu, Zhao, Jerebko, Florin, Valadez, Bogoni, and Moy*	Sparse Gaussian Graphical Models with Unknown Block Structure *Marlin and Murphy*	Online Feature Elicitation in Interactive Optimization *Boutilier, Regan, and Viappiani*
	Good Learners for Evil Teachers *Dekel and Shamir*	Learning Markov Logic Network Structure via Hypergraph Lifting *Kok and Domingos*	Uncertainty Sampling and Transductive Experimental Design for Active Dual Supervision *Sindhwani, Melville, and Lawrence*
16:00 – 16:30	*Coffee break*		
16:30 – 17:30	Awards Session, Leacock 132		
18:45 – 23:00	Poster session: Papers from Sessions 1A to 3F, Leacock/Arts		

Tuesday, June 16th

Invited Talk: Corinna Cortes, Leacock 132
Can learning kernels help performance? (p. 12)

Coffee break

3D: Structured Learning (p. 27)

Leacock 26

Chair: Lawrence Carin

♣ Large Margin Training for Hidden Markov Models with Partially Observed States
Do and Artières

Matrix Updates for Perceptron Training of Continuous Density Hidden Markov Models
Cheng, Sha, and Saul

Unsupervised Search-based Structured Prediction
Daume III

Sparse Higher Order Conditional Random Fields for improved sequence labeling
Qian, Jiang, Zhang, Huang, and Wu

Detecting the Direction of Causal Time Series
Peters, Janzing, Gretton, and Schölkopf

3E: Topic Models (p. 28)

Leacock 210

Chair: Tom Dietterich

Evaluation Methods for Topic Models
Wallach, Murray, Salakhutdinov, and Mimno

Accounting for Burstiness in Topic Models
Doyle and Elkan

Topic-Link LDA: Joint Models of Topic and Author Community
Liu, Niculescu-Mizil, and Gryc

MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification
Zhu, Ahmed, and Xing

Independent Factor Topic Models
Putthividhya, Attias, and Nagarajan

3F: Transfer and MultiTask Learning (p. 29)

Leacock 132

Chair: Kai Yu

Deep Transfer via Second-Order Markov Logic
Davis and Domingos

Feature Hashing for Large Scale Multitask Learning
Weinberger, Dasgupta, Attenberg, Langford, and Smola

A Convex Formulation for Learning Shared Structures from Multiple Tasks
Chen, Tang, Liu, and Ye

EigenTransfer: A Unified Framework for Transfer Learning
Dai, Jin, Xue, Yang, and Yu

Domain Adaptation from Multiple Sources via Auxiliary Classifiers
Duan, Tsang, Xu, and Chua

Lunch break

4D: Lassos and Other L1s (p. 33)

Leacock 232

Chair: Volker Roth

Stochastic Methods for ℓ_1 Regularized Loss Minimization
Shalev-Shwartz and Tewari

★ ♣ Blockwise Coordinate Descent Procedures for the Multi-Task Lasso, with Applications to Neural Semantic Basis Discovery
Liu, Palatucci, and Zhang

An Efficient Projection for $l_{1,\infty}$ Regularization
Quattoni, Carreras, Collins, and Darrell

An Accelerated Gradient Method for Trace Norm Minimization
Ji and Ye

Group Lasso with Overlaps and Graph Lasso
Jacob, Obozinski, and Vert

4E: Document Collections (p. 34)

Leacock 210

Chair: Johannes Fuernkranz

Bayesian Clustering for Email Campaign Detection
Haider and Scheffer

A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining
Jin and Ho

Learning Spectral Graph Transformations for Link Prediction
Kunegis and Lommatzsch

Interactively Optimizing Information Retrieval Systems as a Dueling Bandits Problem, *Yue and Joachims*

Transfer Learning for Collaborative Filtering via a Rating-Matrix Generative Model
Li, Yang, and Xue

4F: Food for Thought (p. 35)

Leacock 15

Chair: Léon Bottou and Michael Littman

Curriculum Learning
Bengio, Louradour, Collobert, and Weston

Herding Dynamical Weights to Learn
Welling

Sequential Bayesian Prediction in the Presence of Changepoints
Garnett, Osborne, and Roberts

Model-Free Reinforcement Learning as Mixture Learning
Vlassis and Toussaint

Active Learning for Directed Exploration of Complex Systems
Burl and Wang

Coffee break

Awards Session, Leacock 132

Poster session: Papers from Sessions 1A to 3F, Leacock/Arts

Wednesday, June 17th		
8:30 – 9:50	**Invited Talk: Emmanuel Dupoux, Leacock 132** **How do infants bootstrap into spoken language?: Models and challenges** (p. 12)	
9:50 – 10:20	*Coffee break*	
10:20 – 12:20	**5A: Algorithms** (p. 36) **Leacock 26** Chair: Alex Smola	**5B: Priors in Probabilistic Models** (p. 37) **Leacock 132** Chair: Max Welling
	Proximal Regularization for Online and Batch Learning *Do, Le, and Foo*	Bayesian Inference for Plackett-Luce Ranking Models *Guiver and Snelson*
	A Majorization-Minimization Algorithm for (Multiple) Hyperparameter Learning *Foo, Do, and Ng*	Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors *Andrzejewski, Zhu, and Craven*
	A Least Squares Formulation for a Class of Generalized Eigenvalue Problems in Machine Learning *Sun, Ji, and Ye*	Nonparametric Factor Analysis with Beta Process Priors *Paisley and Carin*
	On Sampling-based Approximate Spectral Decomposition *Kumar, Mohri, and Talwalkar*	Accelerated Gibbs Sampling for the Indian Buffet Process *Doshi-Velez and Ghahramani*
	Efficient Euclidean Projections in Linear Time *Liu and Ye*	A Stochastic Memoizer for Sequence Data *Wood, Archambeau, Gasthaus, James, and Teh*
12:20 – 14:00	*Lunch break*	
14:00 – 16:00	**6A: Structured Learning: Metric Learning** (p. 41) **Leacock 219** Chair: Sofus Attila Macskassy	**6B: Deep Architectures** (p. 42) **Leacock 132** Chair: Yann LeCun
	Polyhedral Outer Approximations with Application to Natural Language Parsing *Martins, Smith, and Xing*	♠ Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations *Lee, Grosse, Ranganath, and Ng*
	On Primal and Dual Sparsity of Markov Networks *Zhu and Xing*	Using Fast Weights to Improve Persistent Contrastive Divergence *Tieleman and Hinton*
	Learning Structural SVMs with Latent Variables *Yu and Joachims*	Large-scale Deep Unsupervised Learning using Graphics Processors *Raina, Madhavan, and Ng*
	An Efficient Sparse Metric Learning in High-Dimensional Space via ℓ_1-Penalized Log-Determinant Regularization *Qi, Tang, Chua, and Zhang*	Factored Conditional Restricted Boltzmann Machines for Modeling Motion Style *Taylor and Hinton*
	Learning Instance Specific Distances Using Metric Propagation *Zhan, Li, Li, and Zhou*	Deep Learning from Temporal Coherence in Video *Mobahi, Collobert, and Weston*
16:00 – 16:30	*Coffee break*	
16:30 – 17:50	ICML Business Meeting, Leacock 132	
18:45 – 23:00	Poster session: Papers from Sessions 4A to 6E, Leacock/Arts	

Wednesday, June 17th

Invited Talk: Emmanuel Dupoux, Leacock 132
How do infants bootstrap into spoken language?: Models and challenges (p. 12)

Coffee break

5C: RL in High Order Environments (p. 38)	**5D: Learning Theory** (p. 39)	**5E: Paul Utgoff Memorial Session: Learning for Discrete Problems** (p. 40)
Leacock 219	**Leacock 15**	**Leacock 232**
Chair: Kurt Driessens	Chair: Sham Kakade	Chair: Carla Brodley
Binary Action Search for Learning Continuous-Action Control Policies *Pazis and Lagoudakis*	Nonparametric Estimation of the Precision-Recall Curve *Clémençon and Vayatis*	Bandit-Based Optimization on Graphs with Application to Library Performance Tuning *de Mesmay, Rimmel, Voronenko, and Püschel*
Predictive Representations for Policy Gradient in POMDPs *Boularias and Chaib-draa*	Surrogate Regret Bounds for Proper Losses *Reid and Williamson*	Robust Feature Extraction via Information Theoretic Learning *Yuan and Hu*
Stochastic Search using the Natural Gradient *Sun, Wierstra, Schaul, and Schmidhuber*	Robust Bounds for Classification via Selective Sampling *Cesa-Bianchi, Gentile and Orabona*	Block-Wise Construction of Acyclic Relational Features with Monotone Irreducibility and Relevancy Properties *Kuželka and Železný*
Approximate Inference for Planning in Stochastic Relational Worlds *Lang and Toussaint*	PAC-Bayesian Learning of Linear Classifiers *Germain, Lacasse, Laviolette, and Marchand*	Rule Learning with Monotonicity Constraints *Kotlowski and Slowiński*
Discovering Options from Example Trajectories *Zang, Zhou, Minnen, and Isbell*	Piecewise-Stationary Bandit Problems with Side Observations *Yu and Mannor*	Grammatical Inference as a Principal Component Analysis Problem *Bailly, Denis, and Ralaivola*

Lunch break

6C: Learning Actions and Sequences (p. 43)	**6D: Learning Kernels** (p. 44)	**6E: Boosting** (p. 45)
Leacock 232	**Leacock 15**	**Leacock 26**
Chair: Nikos Vlassis	Chair: Francis Bach	Chair: Shai Shalev-Schwartz
Robot Trajectory Optimization using Approximate Inference *Toussaint*	More Generality in Efficient Multiple Kernel Learning *Varma and Babu*	Boosting Products of Base Classifiers *Kégl and Busa-Fekete*
Trajectory Prediction: Learning to Map Situations to Robot Trajectories *Jetchev and Toussaint*	Multiple Indefinite Kernel Learning with Mixed Norm Regularization *Kowalski, Szafranski, and Ralaivola*	ABC-Boost: Adaptive Base Class Boost for Multi-class Classification *Li*
Learning Complex Motions by Sequencing Simpler Motion Templates *Neumann, Maass, and Peters*	Learning Kernels from Indefinite Similarities *Chen, Gupta, and Recht*	Compositional Noisy-Logical Learning *Yuille and Zheng*
Learning When to Stop Thinking and Do Something! *Póczos, Abbasi-Yadkori, Szepesvári, Russell Greiner, and Nathan Sturtevant*	SimpleNPKL: Simple NonParametric Kernel Learning *Zhuang, Tsang, and Hoi*	Boosting with Structural Sparsity *Duchi and Singer*
Monte-Carlo Simulation Balancing *Silver and Tesauro*	Geometry-Aware Metric Learning *Lu, Jain, and Dhillon*	Learning with Structured Sparsity *Huang, Zhang, and Metaxas*

Coffee break

ICML Business Meeting, Leacock 132

Poster session: Papers from Sessions 4A to 6E, Leacock/Arts

Invited Talks

Drifting games, boosting and online learning

Yoav Freund, University of California, San Diego, U.S.A.

Monday, June 15, 4:30–5:50 p.m. (Leacock 132)

Drifting games is a mathematical framework for modeling learning problems. In this talk I will present the framework and show how it is used to derive a new boosting algorithm called Robustboost and a new online prediction algorithm called NormalHedge. I will present two sets of experiments using these algorithms on synthetic and real world data. The first experiments demonstrate that Robustboost outperforms Adaboost and Logitboost when there are many outliers in the training data. The second set of experiments demonstrate that a tracking algorithm based on NormalHedge is more robust against noise than particle filters.

Can learning kernels help performance?

Corinna Cortes, Google Research, U.S.A.

Tuesday, June 16, 8:30–9:50 a.m. (Leacock 132)

Kernel methods combined with large-margin learning algorithms such as SVMs have been used successfully to tackle a variety of learning tasks since their introduction in the early 90s. However, in the standard framework of these methods, the choice of an appropriate kernel is left to the user and a poor selection may lead to sub-optimal performance. Instead, sample points can be used to select a kernel function suitable for the task out of a family of kernels fixed by the user. While this is an appealing idea supported by some recent theoretical guarantees, in experiments, it has proven surprisingly difficult to consistently and significantly outperform simple fixed combination schemes of kernels. This talk will survey different methods and algorithms for learning kernels and will present novel results that tend to suggest that significant performance improvements can be obtained with a large number of kernels.

(Includes joint work with Mehryar Mohri and Afshin Rostamizadeh.)

How do infants bootstrap into spoken language?: Models and challenges

Emmanuel Dupoux, Ecole Normale Superieure, Ecole des Hautes Etudes en Sciences Sociales, Centre National de la Recherche Scientifique, France

Wednesday, June 17, 8:30–9:50 a.m. (Leacock 132)

Human infants learn spontaneously and effortlessly the language(s) spoken in their environments, despite the extraordinary complexity of the task. Here, I will present an overview of the early phases of language acquisition and focus on one area where a modeling approach is currently being conducted using tools of signal processing and automatic speech recognition: the unsupervized acquisition of phonetic categories. During their first year of life, infants construct a detailed representation of the phonemes of their native language and lose the ability to distinguish nonnative phonemic contrasts. Unsupervised statistical clustering is not sufficient; it does not converge on the inventory of phonemes, but rather on contextual allophonic units or subunits. I present an information-theoretic algorithm that groups together allophonic variants based on three sources of information that Can be acquired independently: the statistical distribution of their contexts, the phonetic plausibility of the grouping, and the existence of lexical minimal pairs. This algorithm is tested on several natural speech corpora. We find that these three sources of information are probably not language specific. What is presumably unique to language is the way in which they are combined to optimize the emergence of linguistic categories.

Awards Session Tuesday, 16:30 – 17:30, Leacock 132

The ICML 2009 Awards Committee is pleased to announce the nominees for this year's best paper awards. All best paper award winners will receive a certificate and a check for $800. All nominated papers that do not receive top honors will receive checks for $200. Awards will be presented in a special session at the conference. The best 10-Year Paper winner will give a 20-minute "retrospective" talk during the special session.

Best Paper (⋆)

A paper the committee feels contains innovative and creative results and best conveys deeper insights that are of interest to researchers both inside and outside of machine learning.

- Archipelago: Nonparametric Bayesian Semi-Supervised Learning
  Ryan Adams and Zoubin Ghahramani (* Also a best student paper nominee)
- Blockwise Coordinate Descent Procedures for the Multi-task Lasso, with Applications to Neural Semantic Basis Discovery
  Han Liu, Mark Palatucci, and Jian Zhang (* Also a best student paper nominee)
- Structure Preserving Embedding
  Blake Shaw and Tony Jebara (* Also a best student paper nominee)
- Optimal Reverse Prediction: A Unified Perspective on Supervised, Unsupervised and Semi-supervised Learning
  Linli Xu, Martha White, and Dale Schuurmans

Best Application Paper (♠)

A paper the committee feels best demonstrates the promise of machine learning's impact on applications.

- Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations
  Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng (* Also a best student paper nominee)

Best Student Paper (♣)

Papers the committee judges to be a top-notch machine-learning paper and whose first author is a graduate student. We are grateful to a generous contribution from Springer, the publishers of *Machine Learning*, for helping to fund this award. In addition to the three papers mentioned above, the nominees include:

- Tractable Nonparametric Bayesian Inference in Poisson Processes with Gaussian Process Intensities
  Ryan Adams, Iain Murray, and David MacKay
- A Scalable Framework for Discovering Coherent Co-clusters in Noisy Data
  Meghana Deodhar, Joydeep Ghosh, Gunjan Gupta, Hyuk Cho, and Inderjit Dhillon
- Large Margin Training for Hidden Markov Models with Partially Observed States
  Trinh-Minh-Tri Do and Thierry Artières
- BoltzRank: Learning to Maximize Expected Ranking Gain
  Maksims Volkovs and Richard Zemel

Best 10-Year Paper

A paper published in ICML 1999 the committee feels has had the most significant and lasting impact.

- Transductive Inference for Text Classification using Support Vector Machines
  Thorsten Joachims

Honorable mentions:

- Least-Squares Temporal Difference Learning
  Justin A. Boyan
- Policy Invariance Under Reward Transformations: Theory and Application to Reward Shaping
  Andrew Y. Ng, Daishi Harada, and Stuart J. Russell

ICML 2009 Awards Committee: Francis Bach, Leon Bottou, Andrea Danyluk, Yann LeCun, Michael Littman, Michele Sebag, and Stefan Wrobel

Session 1A: Clustering Geometry — Monday, 10:20 – 12:20, Leacock 132

Chair: Kiri Wagstaff

Solution Stability in Linear Programming Relaxations: Graph Partitioning and Unsupervised Learning

Sebastian Nowozin and Stefanie Jegelka

We propose a new method to quantify the solution stability of a large class of combinatorial optimization problems arising in machine learning. As practical example we apply the method to correlation clustering, clustering aggregation, modularity clustering, and relative performance significance clustering. Our method is extensively motivated by the idea of linear programming relaxations. We prove that when a relaxation is used to solve the original clustering problem, then the solution stability calculated by our method is conservative, that is, it never overestimates the solution stability of the true, unrelaxed problem. We also demonstrate how our method can be used to compute the entire path of optimal solutions as the optimization problem is increasingly perturbed. Experimentally, our method is shown to perform well on a number of benchmark problems.

A Scalable Framework for Discovering Coherent Co-Clusters in Noisy Data

Meghana Deodhar, Joydeep Ghosh, Gunjan Gupta, Hyuk Cho, and Inderjit Dhillon

Clustering problems often involve datasets where only a part of the data is relevant to the problem, e.g., in microarray data analysis only a subset of the genes show cohesive expressions within a subset of the conditions/features. The existence of a large number of non-informative data points and features makes it challenging to hunt for coherent and meaningful clusters from such datasets. Additionally, since clusters could exist in different subspaces of the feature space, a co-clustering algorithm that simultaneously clusters objects and features is often more suitable as compared to one that is restricted to traditional "one-sided" clustering. We propose Robust Overlapping Co-clustering (ROCC), a scalable and very versatile framework that addresses the problem of efficiently mining dense, arbitrarily positioned, possibly overlapping co-clusters from large, noisy datasets. ROCC has several desirable properties that make it extremely well suited to a number of real life applications.

Multi-View Clustering via Canonical Correlation Analysis

Kamalika Chaudhuri, Sham Kakade, Karen Livescu, and Karthik Sridharan

Clustering data in high dimensions is believed to be a hard problem in general. 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. Here, we consider constructing such projections using multiple views of the data, via Canonical Correlation Analysis (CCA).

Under the assumption that the views are uncorrelated given the cluster label, we show that the separation conditions required for the algorithm to be successful are significantly weaker than prior results in the literature. We provide results for mixtures of Gaussians and mixtures of log concave distributions. We also provide empirical support from audio-visual speaker clustering (where we desire the clusters to correspond to speaker ID) and from hierarchical Wikipedia document clustering (where one view is the words in the document and other is the link structure).

Spectral Clustering Based on the Graph p-Laplacian

Thomas Bühler and Matthias Hein

We present a generalized version of spectral clustering using the graph p-Laplacian, a non-linear generalization of the standard graph Laplacian. We show that the second eigenvector of the graph p-Laplacian interpolates between a relexation of the normalized and the Cheeger cut. Moreover, we provide an efficient numerical scheme to compute the second eigenvector of the graph p-Laplacian and give theoretical and experimental evidence that in the limit as p approaches 1 the cut found by thresholding the second eigenvector of the graph p-Laplacian is closely related to the optimal Cheeger cut. Moreover, the experiments show show that the clustering found by p-spectral clustering is at least as good as normal spectral clustering, but often leads to significantly better results.

Nearest Neighbors in High-Dimensional Data: The Emergence and Influence of Hubs

Miloš Radovanović, Alexandros Nanopoulos, and Mirjana Ivanović

High dimensionality can pose severe difficulties, widely recognized as different aspects of the curse of dimensionality. In this paper we study a new aspect of the curse pertaining to the distribution of k-occurrences, i.e., the number of times a point appears among the k nearest neighbors of other points in a data set. We show that, as dimensionality increases, this distribution becomes considerably skewed and hub points emerge (points with very high k-occurrences). We examine the origin of this phenomenon, showing that it is an inherent property of high-dimensional vector space, and explore its influence on applications based on measuring distances in vector spaces, notably classification, clustering, and information retrieval.

Session 1B: Applied Probabilistic Models — Monday, 10:20 – 12:20, Leacock 219

Chair: Florence d'Alche Buc

Unsupervised Hierarchical Modeling of Locomotion Styles

Wei Pan and Lorenzo Torresani

This paper describes an unsupervised learning technique for modeling human locomotion styles, such as distinct related activities (e.g., running and striding) or variations of the same motion performed by different subjects. Modeling motion styles requires identifying the common structure in the motions and detecting style-specific characteristics. We propose an algorithm that learns a hierarchical model of styles from unlabeled motion capture data by exploiting the cyclic property of human locomotion. We assume that sequences with the same style contain locomotion cycles generated by noisy, temporally warped versions of a single latent cycle. We model these style-specific latent cycles as random variables drawn from a common "parent" cycle distribution, representing the structure shared by all motions. Given these hierarchical priors, the algorithm learns, in a completely unsupervised fashion, temporally aligned latent cycle distributions, each modeling a specific locomotion style, and computes for each example the style label posterior distribution, the segmentation into cycles, and the temporal warping with respect to the latent cycles. We demonstrate the flexibility of the model by illustrating its performance on several application problems such as style clustering, animation, style blending, and filling in of missing data.

Exploiting Sparse Markov *and* Covariance Structure in Multiresolution Models

Myung Jin Choi, Venkat Chandrasekaran, and Alan Willsky

We consider Gaussian multiresolution (MR) models in which coarser, hidden variables serve to capture statistical dependencies among the finest scale variables. Tree-structured MR models have limited modeling capabilities, as variables at one scale are forced to be independent of each other conditioned on other scales. We propose a new class of Gaussian MR models that capture the residual correlations within each scale using sparse covariance structure. Our goal is to learn a tree-structured graphical model connecting variables across different scales, while at the same time learning sparse structure for the conditional covariance within each scale conditioned on other scales. This model leads to an efficient, new inference algorithm that is similar to multipole methods in computational physics.

A Bayesian Approach to Protein Model Quality Assessment

Hetunandan Kamisetty and Christopher James Langmead

Given multiple possible models for a protein structure, a common sub-task in in-silico Protein Structure Prediction is ranking these models according to their quality.

Extant approaches use MLE estimates of parameters to obtain point estimates of the Model Quality. We describe a Bayesian alternative to assessing the quality of these models that builds an MRF over the parameters of each model and performs approximate inference to integrate over them. Hyper-parameters are learnt by optimizing a list-wise loss function over training data.

Our results indicate that our Bayesian approach can significantly outperform MLE estimates and that optimizing the hyper-parameters can further improve results.

Multi-Class Image Segmentation using Conditional Random Fields and Global Classification

Nils Plath, Marc Toussaint, and Shinichi Nakajima

A key aspect of semantic image segmentation is to integrate local and global features for the prediction of local segment labels. We present an approach to multi-class segmentation which combines two methods for this integration: a Conditional Random Field (CRF) which couples to local image features and an image classification method which considers global features. The CRF follows the approach of Reynolds & Murphy (2007) and is based on an unsupervised multi scale pre-segmentation of the image into patches, where patch labels correspond to the random variables of the CRF. The output of the classifier is used to constraint this CRF. We demonstrate and compare the approach on a standard semantic segmentation data set.

GAODE and HAODE: Two Proposals Based on AODE to Deal with Continuous Variables

M. Julia Flores, José A. Gámez, Ana M. Martínez, and José M. Puerta

AODE (Aggregating One-Dependence Estimators) is considered one of the most interesting representatives of the Bayesian classifiers, taking into account not only the low error rate it provides but also its efficiency. Until now, all the attributes in a dataset have had to be nominal to build an AODE classifier or they have had to be previously discretized. In this paper, we propose two different approaches in order to deal directly with numeric attributes. One of them uses conditional Gaussian networks to model a dataset exclusively with numeric attributes; and the other one keeps the superparent on each model discrete and uses univariate Gaussians to estimate the probabilities for the numeric attributes and multinomial distributions for the categorical ones, it also being able to model hybrid datasets. Both of them obtain competitive results compared to AODE, the latter in particular being a very attractive alternative to AODE in numeric datasets.

Session 1C: Exploration in Reinforcement Learning Monday, 10:20 – 12:20, Leacock 232

Chair: Pascal Poupart

The Adaptive k-Meteorologists Problem and Its Application to Structure Learning and Feature Selection in Reinforcement Learning

Carlos Diuk, Lihong Li, and Bethany Leffler

The purpose of this paper is three-fold. First, we formalize and study a problem of learning probabilistic concepts in the recently proposed KWIK framework. We give details of an algorithm, known as the Adaptive k-Meteorologists Algorithm, analyze its sample complexity upper bound, and give a matching lower bound. Second, this algorithm is used to create a new reinforcement learning algorithm for factoredstate problems that enjoys significant improvement over the previous state-of-the-art algorithm. Finally, we apply the Adaptive k-Meteorologists Algorithm to remove a limiting assumption in an existing reinforcement-learning algorithm. The effectiveness of our approaches are demonstrated empirically in a couple benchmark domains as well as a robotics navigation problem.

Near-Bayesian Exploration in Polynomial Time

J. Zico Kolter and Andrew Ng

We consider the exploration/exploitation problem in reinforcement learning (RL). The Bayesian approach to model-based RL offers an elegant solution to this problem, by considering a distribution over possible models and acting to maximize expected reward; unfortunately, the Bayesian solution is intractable for all but very restricted cases. In this paper we present a simple algorithm, and prove that with high probability it is able to perform epsilon-close to the true (intractable) optimal Bayesian policy after some small (polynomial in quantities describing the system) number of time steps. The algorithm and analysis are motivated by the so-called PAC-MDP approach, and extend such results into the setting of Bayesian RL. In this setting, we show that we are able to achieve lower sample complexity bounds than existing PAC-MDP algorithms, while using exploration strategies that are much greedier than the (extremely cautious) exploration strategies used by these existing algorithms.

Optimistic Initialization and Greediness Lead to Polynomial Time Learning in Factored MDPs

István Szita and András Lörincz

In this paper we propose an algorithm for polynomial-time reinforcement learning in factored Markov decision processes (FMDPs). The factored optimistic initial model (FOIM) algorithm, maintains an empirical model of the FMDP in a conventional way, and always follows a greedy policy with respect to its model. The only trick of the algorithm is that the model is initialized optimistically. We prove that with suitable initialization (i) FOIM converges to the fixed point of approximate value iteration (AVI); (ii) the number of steps when the agent makes non-near-optimal decisions (with respect to the solution of AVI) is polynomial in all relevant quantities; (iii) the per-step costs of the algorithm are also polynomial. To our best knowledge, FOIM is the first algorithm with these properties.

Dynamic Analysis of Multiagent Q-learning with ϵ-greedy Exploration

Eduardo Rodrigues Gomes and Ryszard Kowalczyk

The development of mechanisms to understand and model the expected behaviour of multiagent learners is becoming increasingly important as the area rapidly find application in a variety of domains. In this paper we present a framework to model the behaviour of Q-learning agents using the e-greedy exploration mechanism. For this, we analyse a continuous-time version of the Q-learning update rule and study how the presence of other agents and the e-greedy mechanism affect it. We then model the problem as a system of difference equations which is used to theoretically analyse the expected behaviour of the agents. The applicability of the framework is tested through experiments in typical games selected from the literature.

Hoeffding and Bernstein Races for Selecting Policies in Evolutionary Direct Policy Search

Verena Heidrich-Meisner and Christian Igel

Uncertainty arises in reinforcement learning from various sources, and therefore it is necessary to consider statistics based on several roll-outs for evaluating behavioral policies. We add an adaptive uncertainty handling based on Hoeffding and empirical Bernstein races to the CMA-ES, a variable metric evolution strategy proposed for direct policy search. The uncertainty handling adjusts individually the number of episodes considered for the evaluation of a policy. The performance estimation is kept just accurate enough for a sufficiently good ranking of candidate policies, which is in turn sufficient for the CMA-ES to find better solutions. This increases the learning speed as well as the robustness of the algorithm.

Session 1D: Online Learning — Monday, 10:20 – 12:20, Leacock 15

Chair: Claudio Gentile

A Simpler Unified Analysis of Budget Perceptrons

Ilya Sutskever

The kernel Perceptron is an appealing online learning algorithm that has a drawback: whenever it makes an error it must increase its support set, which slows training and testing if the number of errors is large. The Forgetron and the Randomized Budget Perceptron algorithms overcome this problem by restricting the number of support vectors the Perceptron is allowed to have. These algorithms have regret bounds whose proofs are dissimilar. In this paper we propose a unified analysis of both of these algorithms by observing that the way in which they remove support vectors can be seen as types of L_2-regularization. By casting these algorithms as instances of online convex optimization problems and applying a variant of Zinkevich's theorem for noisy and incorrect gradient, we can bound the regret of these algorithms more easily than before. Our bounds are similar to the existing ones, but the proofs are less technical.

Efficient Learning Algorithms for Changing Environments

Elad Hazan and C. Seshadhri

We study online learning in an oblivious changing environment. The standard measure of regret bounds the difference between the cost of the online learner and the best decision in hindsight. Hence, regret minimizing algorithms tend to converge to the static best optimum, clearly a suboptimal behavior in changing environments. On the other hand, various metrics proposed to strengthen regret and allow for more dynamic algorithms produce inefficient algorithms.

We propose a different performance metric which strengthens the standard metric of regret and measures performance with respect to a changing comparator. We then describe a series of data-streaming-based reductions which transform algorithms for minimizing (standard) regret into adaptive algorithms albeit incurring only poly-logarithmic computational overhead.

Using this reduction, we obtain efficient low adaptive-regret algorithms for the problem of online convex optimization. This can be applied to various learning scenarios, i.e. online portfolio selection, for which we describe experimental results showing the advantage of adaptivity.

Online Learning by Ellipsoid Method

Liu Yang, Rong Jin, and Jieping Ye

In this work, we extend the ellipsoid method, which was originally designed for convex optimization, for online learning. The key idea is to approximate by an ellipsoid the classification hypotheses that are consistent with all the training examples received so far. This is in contrast to most online learning algorithms where only a single classifier is maintained at each iteration. Efficient algorithms are presented for updating both the centroid and the positive definite matrix of ellipsoid given a misclassified example. In addition to the classical ellipsoid method, an improved version for online learning is also presented. Mistake bounds for both ellipsoid methods are derived. Evaluation with the USPS dataset and three UCI data-sets shows encouraging results when comparing the proposed online learning algorithm to two state-of-the-art online learners.

Learning Prediction Suffix Trees with Winnow

Nikos Karampatziakis and Dexter Kozen

Prediction suffix trees (PSTs) are a popular tool for modeling sequences and have been successfully applied in many domains such as compression and language modeling. In this work we adapt the well studied Winnow algorithm to the task of learning PSTs. The proposed algorithm automatically grows the tree, so that it provably remains competitive with any fixed PST determined in hindsight. At the same time we prove that the depth of the tree grows only logarithmically with the number of mistakes made by the algorithm. Finally, we empirically demonstrate its effectiveness in two different tasks.

Identifying Suspicious URLs: An Application of Large-Scale Online Learning

Justin Ma, Lawrence K. Saul, Stefan Savage, and Geoffrey M. Voelker

This paper explores online learning approaches for detecting malicious Web sites (those involved in criminal scams) using lexical and host-based features of the associated URLs. We show that this application is particularly appropriate for online algorithms as the size of the training data is larger than can be efficiently processed in batch and because the distribution of features that typify malicious URLs is changing continuously. Using a real-time system we developed for gathering URL features, combined with a real-time source of labeled URLs from a large Web mail provider, we demonstrate that recently-developed online algorithms can be as accurate as batch techniques, achieving classification accuracies up to 99% over a balanced data set.

Chair: Marie desJardins

BoltzRank: Learning to Maximize Expected Ranking Gain

Maksims Volkovs and Richard Zemel

Ranking a set of retrieved documents according to their relevance to a query is a popular problem in information retrieval. Methods that learn ranking functions are difficult to optimize, as ranking performance is typically judged by metrics that are not smooth. In this paper we propose a new listwise approach to learning to rank. Our method creates a conditional probability distribution over rankings assigned to documents for a given query, which allows for gradient ascent optimization of the expected value of some performance measure. The rank probabilities take the form of a Boltzmann distribution, based on an energy function that depends on a scoring function composed of individual and pairwise potentials. Including pairwise potentials is a novel contribution, allowing the model to encode regularities in the relative scores of documents; existing models assign scores at test time based only on individual documents, with no pairwise constraints between documents. Experimental results on the LETOR3.0 data sets show that our method out-performs existing learning approaches to ranking.

Decision Tree and Instance-Based Learning for Label Ranking

Weiwei Cheng, Jens Hühn, and Eyke Hüllermeier

The label ranking problem consists of learning a model that maps instances to total orders over a finite set of predefined labels. This paper introduces new methods for label ranking that complement and improve upon existing approaches. More specifically, we propose extensions of two methods that have been used extensively for classification and regression so far, namely instance-based learning and decision tree induction. The unifying element of the two methods is a procedure for locally estimating predictive probability models for label rankings.

Ranking with Ordered Weighted Pairwise Classification

Nicolas Usunier, David Buffoni, and Patrick Gallinari

In ranking with the pairwise classification approach, the loss associated to a predicted ranked list is the mean of the pairwise classification losses. This loss is inadequate for tasks such as information retrieval where we prefer ranked lists with high precision on the top of the list. We propose to optimize a larger class of loss functions for ranking, based on an ordered weighted average (OWA) (Yager, 88) of the classification losses. Convex OWA aggregation operators range from the max to the mean depending on their weights, and can be used to focus on the top ranked elements as they give more weight to the largest losses. When aggregating hinge losses, the optimization problem is similar to the SVM for interdependent output spaces. Moreover, we show that an OWA aggregation of margin-based classification losses has good generalization properties. Experiments on the Letor 3.0 benchmark dataset for information retrieval validate our approach.

Ranking Interesting Subgroups

Stefan Rueping

Subgroup discovery is the task of identifying the top k patterns in a database with most significant deviation in the distribution of a target attribute Y . Subgroup discovery is a popular approach for identifying interesting patterns in data, because it effectively combines statistical significance with an understandable representation of patterns as a logical formula. However, it is often a problem that some subgroups, even if they are statistically highly significant, are not interesting to the user for some reason. In this paper, we present an approach based on the work on ranking Support Vector Machines that ranks subgroups with respect to the user's concept of interestingness, and finds subgroups that are interesting to the user. It will be shown that this approach can significantly increase the quality of the subgroups.

Generalization Analysis of Listwise Learning-to-Rank Algorithms

Yanyan Lan, Tie-Yan Liu, Zhiming Ma, and Hang Li

This paper presents a theoretical framework for ranking, and demonstrates how to perform generalization analysis of listwise ranking algorithms using the framework. Many learning-to-rank algorithms have been proposed in recent years. Among them, the listwise approach has shown higher empirical ranking performance when compared to the other approaches. However, there is no theoretical study on the listwise approach as far as we know. In this paper, we propose a theoretical framework for ranking, which can naturally describe various listwise learning- to-rank algorithms. With this framework, we prove a theorem which gives a generalization bound of a listwise ranking algorithm, on the basis of Rademacher Average of the class of compound functions. The compound functions take listwise loss functions as outer functions and ranking models as inner functions. We then compute the Rademacher Averages for existing listwise algorithms of ListMLE, ListNet, and RankCosine. We also discuss the tightness of the bounds in different situations with regard to the list length and transformation function.

Session 2A: Graphs and Embeddings — Monday, 14:00 – 16:00, Leacock 219

Chair: Kilian Weinberger

Fitting a Graph to Vector Data

Samuel Daitch, Jonathan Kelner, and Daniel Spielman

We introduce a measure of how well a combinatorial graph fits a collection of vectors. The optimal graphs under this measure may be computed by solving convex quadratic programs, and have many interesting properties. For vectors in d-dimensional space, the graphs always have average degree at most 2(d+1); and for vectors in 2 dimensions they are always planar. We compute these graphs for many standard data sets and show that they can be used to obtain good solutions to classification, regression and clustering problems.

Structure Preserving Embedding

Blake Shaw and Tony Jebara

Structure Preserving Embedding (SPE) is an algorithm for embedding graphs in Euclidean space such that the embedding is low-dimensional and preserves the global topological properties of the input graph. Topology is preserved if a connectivity algorithm, such as k-nearest neighbors, can easily recover the edges of the input graph from only the coordinates of the nodes after embedding. SPE is formulated as a semidefinite program that learns a low-rank kernel matrix constrained by a set of linear inequalities which captures the connectivity structure of the input graph. Traditional graph embedding algorithms do not preserve structure according to our definition, and thus the resulting visualizations can be misleading or less informative. SPE provides significant improvements in terms of visualization and lossless compression of graphs, outperforming popular methods such as spectral embedding and Laplacian eigenmaps. We find that many classical graphs and networks can be properly embedded using only a few dimensions. Furthermore, introducing structure preserving constraints into dimensionality reduction algorithms produces more accurate representations of high-dimensional data.

Graph Construction and b-Matching for Semi-Supervised Learning

Tony Jebara, Jun Wang, and Shih-Fu Chang

Graph based semi-supervised learning (SSL) methods play an increasingly important role in practical machine learning systems. A crucial step in graph based SSL methods is the conversion of data into a weighted graph. However, most of the SSL literature focuses on developing label inference algorithms without extensively studying the graph building method and its effect on performance. This article provides an empirical study of leading semi-supervised methods under a wide range of graph construction algorithms. These SSL inference algorithms include the Local and Global Consistency (LGC) method, the Gaussian Random Field (GRF) method, the Graph Transduction via Alternating Minimization (GTAM) method as well as other techniques. Several approaches for graph construction, sparsification and weighting are explored including the popular k-nearest neighbors method (kNN) and the b-matching method. As opposed to the greedily constructed kNN graph, the b-matched graph ensures each node in the graph has the same number of edges and produces a balanced or regular graph. Experimental results on both artificial data and real benchmark datasets indicate that b-matching produces more robust graphs and therefore provides significantly better prediction accuracy without any significant change in computation time.

Partial Order Embedding with Multiple Kernels

Brian McFee and Gert Lanckriet

We consider the problem of embedding arbitrary objects (e.g., images, audio, documents) into Euclidean space subject to a partial order over pairwise distances. Partial order constraints arise naturally when modeling human perception of similarity. Our partial order framework enables the use of graph-theoretic tools to more efficiently produce the embedding, and exploit global structure within the constraint set.

We present an embedding algorithm based on semidefinite programming, which can be parameterized by multiple kernels to yielding a unified space from heterogeneous features.

Probabilistic Dyadic Data Analysis with Local and Global Consistency

Deng Cai, Xuanhui Wang, and Xiaofei He

Dyadic data arises in many real world applications such as social network analysis and information retrieval. In order to discover the underlying or hidden structure in the dyadic data, many topic modeling techniques were proposed. The typical algorithms include Probabilistic Latent Semantic Analysis (PLSA) and Latent Dirichlet Allocation (LDA). The probability density functions obtained by both of these two algorithms are supported on the Euclidean space. However, many previous studies have shown naturally occurring data may reside on or close to an underlying submanifold. We introduce a probabilistic framework for modeling both the topical and geometrical structure of the dyadic data that explicitly takes into account the local manifold structure. Specifically, the local manifold structure is modeled by a graph. The graph Laplacian, analogous to the Laplace-Beltrami operator on manifolds, is applied to smooth the probability density functions. As a result, the obtained probabilistic distributions are concentrated around the data manifold. Experimental results on real data sets demonstrate the effectiveness of the proposed approach.

Session 2B: Gaussian Processes — Monday, 14:00 – 16:00, Leacock 15

Chair: John Guiver

Non-Linear Matrix Factorization with Gaussian Processes

Neil D. Lawrence and Raquel Urtasun

A popular approach to collaborative filtering is matrix factorization. In this paper we consider the "probabilistic matrix factorization" and by taking a latent variable model perspective we show its equivalence to Bayesian PCA. This inspires us to consider probabilistic PCA and its non-linear extension, the Gaussian process latent variable model (GP-LVM) as an approach for probabilistic non-linear matrix factorization. We apply approach to benchmark movie recommender data sets. The results show better than previous state-of-the-art performance.

Analytic Moment-Based Gaussian Process Filtering

Marc Peter Deisenroth, Marco F. Huber, and Uwe D. Hanebeck

We propose an analytic moment-based filter for nonlinear stochastic dynamical systems modeled by Gaussian processes. Exact expressions for the expected value and the covariance matrix are provided for both the prediction and the filter step, where an additional Gaussian assumption is exploited in the latter case. The new filter does not require further approximations. In particular, it avoids sample approximations. We compare the filter to a variety of available Gaussian filters, such as the EKF, the UKF, and the GP-UKF recently proposed by Ko et al. (2007).

Function Factorization using Warped Gaussian Processes

Mikkel N. Schmidt

We introduce a new approach to non-linear regression called function factorization, that is suitable for problems where an output variable can reasonably be modeled by a number of multiplicative interaction terms between non-linear functions of the inputs. The idea is to approximate a complicated function on a high-dimensional space by the sum of products of simpler functions on lower-dimensional subspaces. Function factorization can be seen as a generalization of matrix and tensor factorization methods, in which the data are approximated by the sum of outer products of vectors. We present a non-parametric Bayesian approach to function factorization where the priors over the factorizing functions are warped Gaussian processes, and we do inference using Hamiltonian Markov chain Monte Carlo. We demonstrate the superior predictive performance of the method on a food science data set compared to Gaussian process regression and tensor factorization using PARAFAC and GEMANOVA models.

Tractable Nonparametric Bayesian Inference in Poisson Processes with Gaussian Process Intensities

Ryan Adams, Iain Murray, and David MacKay

The inhomogeneous Poisson process is a point process that has varying intensity across its domain (usually time or space). For nonparametric Bayesian modeling, the Gaussian process is a useful way to place a prior distribution on this intensity. The combination of an Poisson process and GP is known as a Gaussian Cox process, or doubly-stochastic Poisson process. Likelihood-based inference in these models requires an intractable integral over an infinite-dimensional random function. In this paper we present the first approach to Gaussian Cox processes in which it is possible to perform inference without introducing approximations or finite-dimensional proxy distributions. We call our method the Sigmoidal Gaussian Cox Process, which uses a generative model for Poisson data to enable tractable inference via Markov chain Monte Carlo. We compare our methods to competing methods on synthetic data and also apply it to several real-world data sets.

Large-Scale Collaborative Prediction Using a Nonparametric Random Effects Model

Kai Yu, John Lafferty, and Shenghuo Zhu

A nonparametric model is introduced that allows multiple related regression tasks to take inputs from a common data space. Traditional transfer learning models can be inappropriate if the dependence among the outputs cannot be fully resolved by known input-specific and task-specific predictors. The proposed model treats such output responses as conditionally independent, given known predictors and appropriate unobserved random effects. The model is nonparametric in the sense that the dimensionality of random effects is not specified a priori but is instead determined from data. An approach to estimating the model is presented uses an EM algorithm that is efficient on a very large scale collaborative prediction problem. The obtained prediction accuracy is competitive with state-of-the-art results.

Session 2C: Dynamical Systems — Monday, 14:00 – 16:00, Leacock 232

Chair: Doina Precup

Hilbert Space Embeddings of Conditional Distributions with Applications to Dynamical Systems

Le Song, Jonathan Huang, Alex Smola, and Kenji Fukumizu

In this paper, we extend the Hilbert space embedding approach to handle conditional distributions. This leads us to a nonparametric method for modeling dynamical systems, and allows us to update the belief state of a dynamical system by maintaining a conditional embedding. Our method is very general in terms of both the domains and the types of distributions that it can handle, and we demonstrate the effectiveness of our method in various dynamical systems. We expect that Hilbert space embedding of *conditional* distributions will have wide applications beyond modeling dynamical systems.

Learning Nonlinear Dynamic Models

John Langford, Ruslan Salakhutdinov, and Tong Zhang

We present a novel approach for learning nonlinear dynamic models, which leads to a new set of tools capable of solving problems that are otherwise difficult. We provide theory showing this new approach is consistent for models with long range structure, and apply the approach to motion capture and high-dimensional video data, yielding results superior to standard alternatives.

Optimized Expected Information Gain for Nonlinear Dynamical Systems

Alberto Giovanni Busetto, Cheng Soon Ong, and Joachim M. Buhmann

This paper addresses the problem of active model selection for nonlinear dynamical systems. We propose a novel learning approach that selects the most informative subset of time-dependent variables for the purpose of Bayesian model inference. The model selection criterion maximizes the expected Kullback-Leibler divergence between the prior and the posterior probabilities over the models. The proposed strategy generalizes the standard D-optimal design, which is obtained from a uniform prior with Gaussian noise. In addition, our approach allows us to determine an information halting criterion for model identification. We illustrate the benefits of our approach by differentiating between 18 published biochemical models of the TOR signaling pathway, a model selection problem in systems biology. By generating pivotal selection experiments, our strategy outperforms the standard A-optimal, D-optimal and E-optimal sequential design techniques.

Learning Linear Dynamical Systems without Sequence Information

Tzu-Kuo Huang and Jeff Schneider

Virtually all methods of learning dynamic systems from data start from the same basic assumption: that the learning algorithm will be provided with a sequence, or trajectory, of data generated from the dynamic system. In this paper we consider the case where the data is not sequenced. The learning algorithm is presented a set of data points from the system's operation but with no temporal ordering. The data are simply drawn as individual disconnected points.

While making this assumption may seem absurd at first glance, we observe that many scientific modeling tasks have exactly this property. In this paper we restrict our attention to learning linear, discrete time models. We propose several algorithms for learning these models based on optimizing approximate likelihood functions and test the methods on several synthetic data sets.

Dynamic Mixed Membership Block Model for Evolving Networks

Wenjie Fu, Le Song, and Eric Xing

In a dynamic social or biological environment, the interactions between the underlying actors can undergo large and systematic changes. Each actor in the networks can assume multiple related roles and their affiliation to each role as determined by the dynamic links will also exhibit rich temporal phenomenon. We propose a state space mixed membership stochastic blockmodel which captures the dependency between these multiple correlated roles, and enables us to track the mixed membership of each actor in the latent space across time. We derived efficient approximate learning and inference algorithms for our model, and applied the learned models to analyze an email network in Enron Corp., and a rewiring gene interaction network of yeast collected during its full cell cycle. In both cases, our model reveals interesting patterns of the dynamic roles of the actors.

Session 2D: Kernels — Monday, 14:00 – 16:00, Leacock 26

Chair: Dale Schuurmans

Route Kernels for Trees

Fabio Aiolli, Giovanni Da San Martino, and Alessandro Sperduti

Almost all tree kernels proposed in the literature match substructures without taking into account their relative positioning with respect to one another. In this paper, we propose a novel family of kernels which explicitly focus on this type of information. Specifically, after defining a family of tree kernels based on routes between nodes, we present an efficient implementation for a member of this family. Experimental results on four different datasets show that our method is able to reach state of the art performances, obtaining in some cases performances better than computationally more demanding tree kernels.

The Graphlet Spectrum

Risi Kondor, Nino Shervashidze, and Karsten Borgwardt

Current graph kernels suffer from two limitations: graph kernels based on counting particular types of subgraphs ignore the relative position of these subgraphs to each other, while graph kernels based on feature extraction by algebraic methods are limited to graphs without node labels. In this paper we present the graphlet spectrum, a system of graph invariants derived by means of group representation theory, that captures information about the number as well as the position of labeled subgraphs in a given graph. In our experimental evaluation the graphlet spectrum outperforms state-of-the-art graph kernels.

Multi-Instance Learning by Treating Instances As Non-I.I.D. Samples

Zhi-Hua Zhou, Yu-Yin Sun, and Yu-Feng Li

Previous studies on multi-instance learning typically treated instances in the *bags* as independently and identically distributed. The instances in a bag, however, are rarely independent in real tasks, and a better performance can be expected if the instances are treated in an non-i.i.d. way that exploits relations among instances. In this paper, we propose two simple yet effective methods. In the first method, we explicitly map every bag to an undirected graph and design a graph kernel for distinguishing the positive and negative bags. In the second method, we implicitly construct graphs by deriving affinity matrices and propose an efficient graph kernel considering the clique information. The effectiveness of the proposed methods are validated by experiments.

Non-Monotonic Feature Selection

Zenglin Xu, Rong Jin, Jieping Ye, Michael R. Lyu, and Irwin King

We consider the problem of selecting a subset of m most informative features where m is the number of required features. This feature selection problem is essentially a combinatorial optimization problem, and is usually solved by an approximation. Conventional feature selection methods address the computational challenge in two steps: (a) ranking all the features by certain scores that are usually computed independently from the number of specified features m, and (b) selecting the top m ranked features. One major shortcoming of these approaches is that if a feature f is chosen when the number of specified features is m, it will always be chosen when the number of specified features is larger than m. We refer to this property as the "*monotonic*" property of feature selection. In this work, we argue that it is important to develop efficient algorithms for non-monotonic feature selection. To this end, we develop an algorithm for non-monotonic feature selection that approximates the related combinatorial optimization problem by a Multiple Kernel Learning (MKL) problem. We also present a strategy that derives a discrete solution from the approximate solution of MKL, and show the performance guarantee for the derived discrete solution when compared to the global optimal solution for the related combinatorial optimization problem. An empirical study with a number of benchmark data sets indicates the promising performance of the proposed framework compared with several state-of-the-art approaches for feature selection.

Regression by Dependence Minimization and its Application to Causal Inference

Joris Mooij, Dominik Janzing, Jonas Peters, and Bernhard Schölkopf

Motivated by causal inference problems, we propose a novel method for regression that minimizes the statistical dependence between regressors and residuals. The key advantage of this approach to regression is that it does not assume a particular distribution of the noise, i.e., it is non-parametric with respect to the noise distribution. We argue that the proposed regression method is well suited to the task of causal inference in additive noise models. A practical disadvantage is that the resulting optimization problem is generally non-convex and can be difficult to solve. Nevertheless, we report good results on one of the tasks of the NIPS 2008 Causality Challenge, where the goal is to distinguish causes from effects in pairs of statistically dependent variables. In addition, we propose an algorithm for efficiently inferring causal models from observational data for more than two variables. The required number of regressions and independence tests is quadratic in the number of variables, which is a significant improvement over the simple method that tests all possible DAGs.

Session 2E: Learning Codebooks and Dictionaries Monday, 14:00 – 16:00, Leacock 132

Chair: Samy Bengio

Gradient Descent with Sparsification: An Iterative Algorithm for Sparse Recovery with Restricted Isometry Property

Rahul Garg and Rohit Khandekar

In this paper, we present an algorithm for finding an s-sparse vector x that minimizes the *square-error* $\|y-\Phi x\|^2$ where Φ satisfies the *restricted isometry property* (RIP). Our algorithm, called *GraDeS* (Gradient Descent with Sparsification) starts from an arbitrary s-sparse x and iteratively updates it as: $x \leftarrow H_s\left(x + \frac{1}{\gamma}\cdot\Phi^t(y-\Phi x)\right)$ where $\gamma > 1$ is a constant and H_s sets all but largest s coordinates in absolute value to zero.

We show that GraDeS, in constant number of iterations, computes the correct s-sparse solution to the system $y = \Phi x$ where Φ satisfies the condition that the *isometric constant* $\delta_{2s} < 1/3$. This is the most general condition for which, *near-linear time* algorithm is known. In comparison, the best condition under which any polynomial-time algorithm is known, is $\delta_{2s} < \sqrt{2}-1$. An important contribution of the paper is to analyze how the hard-thresholding function H_s acts w.r.t. the potential $\|y-\Phi x\|^2$. A special case of GraDeS, corresponding to $\gamma = 1$, called *Iterative Hard Thresholding (IHT)*, was previously shown to converge when $\delta_{3s} < 1/\sqrt{32}$.

Our Matlab implementation of GraDeS out-performs previously proposed algorithms like Subspace Pursuit, StOMP, OMP, and Lasso by an order of magnitude. Curiously, our experiments also uncovered several cases where L1-regularized regression (Lasso) fails but GraDeS finds the correct solution.

Learning Dictionaries of Stable Autoregressive Models for Audio Scene Analysis

Youngmin Cho and Lawrence Saul

In this paper, we explore an application of basis pursuit to audio scene analysis. The goal of our work is to detect when certain sounds are present in a mixed audio signal. We focus on the regime where out of a large number of possible sources, a small but unknown number combine and overlap to yield the observed signal. To infer which sounds are present, we decompose the observed signal as a linear combination of a small number of active sources. We cast the inference as a regularized form of linear regression whose sparse solutions yield decompositions with few active sources. We characterize the acoustic variability of individual sources by autoregressive models of their time domain waveforms. When we do not have prior knowledge of the individual sources, the coefficients of these autoregressive models must be learned from audio examples. We analyze the dynamical stability of these models and show how to estimate stable models by substituting a simple convex optimization for a difficult eigenvalue problem. We demonstrate our approach by learning dictionaries of musical notes and using these dictionaries to analyze polyphonic recordings of piano, cello, and violin.

Online Dictionary Learning for Sparse Coding

Julien Mairal, Francis Bach, Jean Ponce, and Guillermo Sapiro

Sparse coding—that is, modelling data vectors as sparse linear combinations of basis elements—is widely used in machine learning, neuroscience, signal processing, and statistics. This paper focuses on learning the basis set, also called dictionary, to adapt it to specific data, an approach that has recently proven to be very effective for signal reconstruction and classification in the audio and image processing domains. This paper proposes a new online optimization algorithm for dictionary learning, based on stochastic approximations, which scales up gracefully to large datasets with millions of training samples. A proof of convergence is presented, along with experiments with natural images demonstrating that it leads to faster performance and better dictionaries than classical batch algorithms for both small and large datasets.

Learning Non-Redundant Codebooks for Classifying Complex Objects

Wei Zhang, Akshat Surve, Xiaoli Fern, and Thomas Dietterich

Codebook-based representations are widely employed in the classification of complex objects such as images and documents. Most previous codebook-based methods construct a single codebook via clustering that maps a bag of low-level features into a fixed-length histogram that describes the distribution of these features. This paper describes a simple yet effective framework for learning multiple non-redundant codebooks that produces surprisingly good results. In this framework, each codebook is learned in sequence to extract discriminative information that was not captured by preceding codebooks and their corresponding classifiers. We apply this framework to two application domains: visual object categorization and document classification. Experiments on large classification tasks show substantial improvements in performance compared to a single codebook or codebooks learned in a bagging style.

Prototype Vector Machine for Large Scale Semi-supervised Learning

Kai Zhang, James T. Kwok, and Bahram Parvin

Practical data analysis and mining rarely falls exactly into the supervised learning scenario. Rather, the growing amount of unlabelled data from various scientific domains poses a big challenge to large-scale semi-supervised learning (SSL). We note that the computational intensiveness of graph-based SSL arises largely from the manifold or graph regularization, which may in turn lead to large models that are difficult to handle. To alleviate this, we proposed the prototype vector machine (PVM), a highly scalable, graph-based algorithm for large-scale SSL. Our key innovation is the use of "prototypes vectors" for efficient approximation on both the graph-based regularizer and the model representation. The choice of prototypes are grounded upon two important criterion: they not only perform effective low-rank approximation on the kernel matrix, but also span a model suffering the minimum information loss compared with the complete model. These criterion lead to consistent prototype selection scheme, allowing us to design a unified algorithm (PVM) that demonstrates encouraging performance while at the same time possessing appealing scaling properties (empirically linear with sample size).

Session 3A: Clustering Algorithms — Tuesday, 10:20 – 12:20, Leacock 15

Chair: Tony Jebara

Multi-Assignment Clustering for Boolean Data

Andreas Peter Streich, Mario Frank, David Basin, and Joachim M. Buhmann

Conventional clustering methods typically assume that each data item belongs to a single cluster. This assumption does not hold in general. In order to overcome this limitation, we propose a generative method for clustering vectorial data, where each object can be assigned to multiple clusters. Using a deterministic annealing scheme, our method decomposes the observed data into the contributions of individual clusters and infers their parameters.

Experiments on synthetic Boolean data show that our method achieves higher accuracy in the source parameter estimation and superior cluster stability compared to state-of-the-art approaches. We also apply our method to an important problem in computer security known as role mining. Experiments on real-world access control data show performance gains in generalization to new employees against other multi-assignment methods. In challenging situations with high noise levels, our approach maintains its good performance, while alternative state-of-the-art techniques lack robustness.

K-means in Space: A Radiation Sensitivity Evaluation

Kiri Wagstaff and Benjamin Bornstein

Spacecraft are increasingly making use of onboard data analysis to inform additional data collection and prioritization decisions. However, many spacecraft operate in high-radiation environments in which the reliability of data-intensive computation is not known. This paper presents the first study of radiation sensitivity for k-means clustering. Our key findings are that 1) k-means data structures differ in sensitivity, and sensitivity is not determined by the amount of memory exposed, 2) no special radiation protection is needed below a data-set-dependent radiation threshold, enabling the use of faster, smaller, and cheaper onboard memory in some cases, and 3) subsampling improves radiation tolerance slightly, but the use of kd-trees unfortunately reduces tolerance. Our conclusions can be used to tailor k-means for future use in high-radiation environments.

Information Theoretic Measures for Clusterings Comparison: Is a Correction for Chance Necessary?

Xuan Vinh Nguyen, Julien Epps, and James Bailey

Information theoretic based measures form a fundamental class of similarity measures for comparing clusterings, beside the class of pair-counting based and set-matching based measures. In this paper, we discuss the necessity of correction for chance for information theoretic based measures for clusterings comparison. We observe that the baseline for such measures, i.e. average value under random partitioning of a data set, does not take on a constant value, and tends to have larger variation when the ratio between the number of data points and the number of clusters is small. This effect is similar in some other non-information theoretic based measures such as the well-known Rand Index. Assuming a hypergeometric model of randomness, we derive the analytical formula for the expected mutual information value between a pair of clusterings, and then propose the adjusted version for several popular information theoretic based measures. Some examples are given to demonstrate the need and usefulness of the adjusted measures.

Fast Evolutionary Maximum Margin Clustering

Fabian Gieseke, Tapio Pahikkala, and Oliver Kramer

The maximum margin clustering approach is a recently proposed extension of the concept of support vector machines to the clustering problem. Briefly stated, it aims at finding an optimal partition of the data into two classes such that the margin induced by a subsequent application of a support vector machine is maximal. We propose a method based on stochastic search to address this hard optimization problem. While a direct implementation would be infeasible for large data sets, we present an efficient computational shortcut for assessing the "quality" of intermediate solutions. Experimental results show that our approach outperforms existing methods in terms of clustering accuracy.

Discriminative k Metrics

Arthur Szlam and Guillermo Sapiro

The k q-flats algorithm is a generalization of the popular k-means algorithm where q dimensional best fit affine sets replace centroids as the cluster prototypes. In this work, a modification of the k q-flats framework for pattern classification is introduced. The basic idea is to replace the original reconstruction only energy, which is optimized to obtain the k affine spaces, by a new energy that incorporates discriminative terms. This way, the actual classification task is introduced as part of the design and optimization. The presentation of the proposed framework is complemented with experimental results, showing that the method is computationally very efficient and gives excellent results on standard supervised learning benchmarks.

Session 3B: Inference in Probabilistic Models — Tuesday, 10:20 – 12:20, Leacock 210

Chair: Kevin Murphy

Orbit-Product Representation and Correction of Gaussian Belief Propagation

Jason Johnson, Vladimir Chernyak, and Michael Chertkov

We present a new view of Gaussian belief propagation (GaBP) based on a representation of the determinant as a product over orbits of a graph. We show that the GaBP determinant estimate captures totally backtracking orbits of the graph and consider how to correct this estimate. We show that the missing orbits may be grouped into equivalence classes corresponding to backtrackless orbits and the contribution of each equivalence class is easily determined from the GaBP solution. Furthermore, we demonstrate that this multiplicative correction factor can be interpreted as the determinant of a backtrackless adjacency matrix of the graph with edge weights based on GaBP. Finally, an efficient method is proposed to compute a truncated correction factor including all backtrackless orbits up to a specified length.

Convex Variational Bayesian Inference for Large Scale Generalized Linear Models

Hannes Nickisch and Matthias Seeger

We show how variational Bayesian inference can be implemented for very large generalized linear models. Our relaxation is proven to be a convex problem for any log-concave model. We provide a generic double loop algorithm for solving this relaxation on models with arbitrary super-Gaussian potentials. By iteratively decoupling the criterion, most of the work can be done by solving large linear systems, rendering our algorithm orders of magnitude faster than previously proposed solvers for the same problem. We evaluate our method on problems of Bayesian active learning for large binary classification models, and show how to address settings with many candidates and sequential inclusion steps.

Archipelago: Nonparametric Bayesian Semi-Supervised Learning

Ryan Adams and Zoubin Ghahramani

Semi-supervised learning (SSL), is classification where additional unlabeled data can be used to improve accuracy. Generative approaches are appealing in this situation, as good models of the data's probability density can assist in identifying clusters. Nonparametric Bayesian methods, while ideal in theory due to their principled motivations, have been difficult to apply to SSL in practice. In this work, we present a nonparametric Bayesian method that uses Gaussian processes for the generative model, avoiding many of the problems associated with Dirichlet process mixture models. Our model is fully generative and we take advantage of recent advances in Markov chain Monte Carlo algorithms to provide a practical inference method. Our method compares favorably to competing approaches on synthetic and real-world multi-class data.

The Bayesian Group-Lasso for Analyzing Contingency Tables

Sudhir Raman, Thomas Fuchs, Peter Wild, Edgar Dahl, and Volker Roth

Group-Lasso estimators, useful in many applications, suffer from lack of meaningful variance estimates for regression coefficients. To overcome such problems, we propose a full Bayesian treatment of the Group-Lasso, extending the standard Bayesian Lasso, using hierarchical expansion. The method is then applied to Poisson models for contingency tables using a highly efficient MCMC algorithm. The simulated experiments validate the performance of this method on artificial datasets with known ground-truth. When applied to a breast cancer dataset, the method demonstrates the capability of identifying the differences in interactions patterns of marker proteins between different patient groups.

Split Variational Inference

Guillaume Bouchard and Onno Zoeter

We propose a deterministic method to evaluate the integral of a positive function based on soft-binning functions that smoothly cut the integral into smaller integrals that are easier to approximate. The use of mean-field approximations for each individual sub-part leads to a tractable algorithm that alternates between the optimization of the bins and the approximation of the local integrals. We introduce suitable choices for the binning functions such that a standard mean field approximation can be extended to a split mean field approximation without the need for extra derivations. The method can be seen as a revival of the ideas underlying the mixture mean field approach. We discuss the relation between the two algorithms.

Session 3C: Reinforcement Learning with Temporal Differences

Tuesday, 10:20 – 12:20, Leacock 232

Chair: Shimon Whiteson

Proto-Predictive Representation of States with Simple Recurrent Temporal-Difference Networks

Takaki Makino

We propose a new neural network architecture, called Simple Recurrent Temporal-Difference Networks (SR-TDNs), that learns to predict future observations in partially observable environments. SR-TDNs incorporate the structure of simple recurrent neural networks (SRNs) into temporal-difference (TD) networks to use proto-predictive representation of states. Although they deviate from the principle of predictive representations to ground state representations on observations, they follow the same learning strategy as TD networks, i.e., applying TD-learning to general predictions. Simulation experiments revealed that SR-TDNs can correctly represent states with incomplete set of core tests (question networks), and consequently, SR-TDNs have better on-line learning capacity than TD networks in various environments.

Regularization and Feature Selection in Least Squares Temporal-Difference Learning

J. Zico Kolter and Andrew Ng

We consider the task of reinforcement learning with linear value function approximation. Temporal difference algorithms, and in particular the Least-Squares Temporal Difference (LSTD) algorithm, provide a method for learning the parameters of the value function, but when the number of features is large this algorithm can over-fit to the data and is computationally expensive. In this paper, we propose a regularization framework for the LSTD algorithm that overcomes these difficulties. In particular, we focus on the case of l1 regularization, which is robust to irrelevant features and also serves as a method for feature selection. Although the l1 regularized LSTD solution cannot be expressed as a convex optimization problem, we present an algorithm similar to the Least Angle Regression (LARS) algorithm that can efficiently compute the optimal solution. Finally, we demonstrate the performance of the algorithm experimentally.

Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation

Richard S. Sutton, Hamid Reza Maei, Doina Precup, Shalabh Bhatnagar, David Silver, Csaba Szepesvári, and Eric Wiewiora

Sutton, Szepesvari and Maei (2009) recently introduced the first temporal-difference learning algorithm compatible with both linear function approximation and off-policy training, and whose complexity scales only linearly in the size of the function approximator. Although their gradient temporal difference (GTD) algorithm converges reliably, it can be very slow compared to conventional linear TD (on on-policy problems where TD is convergent), calling into question its practical utility. In this paper we introduce two new related algorithms with better convergence rates. The first algorithm, GTD2, is derived and proved convergent just as GTD was, but uses a different objective function and converges significantly faster (but still not as fast as conventional TD). The second new algorithm, linear TD with gradient correction, or TDC, uses the same update rule as conventional TD except for an additional term which is initially zero. In our experiments on small test problems and in a Computer Go application with a million features, the learning rate of this algorithm was comparable to that of conventional TD. This algorithm appears to extend linear TD to off-policy learning with no penalty in performance while only doubling computational requirements.

Kernelized Value Function Approximation for Reinforcement Learning

Gavin Taylor and Ronald Parr

A recent surge in research in kernelized approaches to reinforcement learning has sought to bring the benefits of kernelized machine learning techniques to reinforcement learning. Kernelized reinforcement learning techniques are fairly new and different authors have approached the topic with different assumptions and goals. Neither a unifying view nor an understanding of the pros and cons of different approaches has yet emerged. In this paper, we offer a unifying view of the different approaches to kernelized value function approximation for reinforcement learning. We show that, except for different approaches to regularization, Kernelized LSTD (KLSTD) is equivalent to a model based approach that uses kernelized regression to find an approximate reward and transition model, and that Gaussian Process Temporal Difference learning (GPTD) returns a mean value function that is equivalent to these other approaches. We also demonstrate the relationship between our model based approach and the earlier Gaussian Processes in Reinforcement Learning (GPRL). Finally, we decompose the Bellman error into the sum of transition error and reward error terms, and demonstrate through experiments that this decomposition can be helpful in choosing regularization parameters.

Constraint Relaxation in Approximate Linear Programs

Marek Petrik and Shlomo Zilberstein

Approximate linear programming (ALP) is a reinforcement learning technique with nice theoretical properties, but it often performs poorly in practice. We identify some reasons for the poor quality of ALP solutions in problems where the approximation induces virtual loops. We then introduce two methods for improving solution quality. One method rolls out selected constraints of the ALP, guided by the dual information. The second method is a relaxation of the ALP, based on external penalty methods. The latter method is applicable in domains in which rolling out constraints is impractical. Both approaches show promising empirical results for simple benchmark problems as well as for a more realistic blood inventory management problem.

Session 3D: Structured Learning — Tuesday, 10:20 – 12:20, Leacock 26

Chair: Lawrence Carin

Large Margin Training for Hidden Markov Models with Partially Observed States

Trinh-Minh-Tri Do and Thierry Artiéres

Large margin learning of Continuous Density Hidden Markov Models with a partially labeled dataset has been extensively studied in the speech and handwriting recognition fields. Yet due to the non convexity of the optimization problem, previous works usually rely on severe approximations so that it is still an open problem. We propose a new learning algorithm that relies on non convex optimization and bundle methods and allows tackling the original optimization problem as is. It is proved to converge to a solution with accuracy ϵ with a rate $O(1/\epsilon)$. We provide experimental results gained on speech recognition and on handwriting recognition that demonstrate the potential of the method.

Matrix Updates for Perceptron Training of Continuous Density Hidden Markov Models

Chih-Chieh Cheng, Fei Sha, and Lawrence Saul

In this paper, we investigate a simple, mistake-driven learning algorithm for discriminative training of continuous density hidden Markov models (CD-HMMs). Most CD-HMMs for automatic speech recognition use multivariate Gaussian emission densities (or mixtures thereof) parameterized in terms of their means and covariance matrices. For discriminative training of CD-HMMs, we reparameterize these Gaussian distributions in terms of positive semidefinite matrices that jointly encode their mean and covariance statistics. We show how to explore the resulting parameter space in CD-HMMs with perceptron-style updates that minimize the distance between Viterbi decodings and target transcriptions. We experiment with several forms of updates, systematically comparing the effects of different matrix factorizations, initializations, and averaging schemes on phone accuracies and convergence rates. We present experimental results for context-independent CD-HMMs trained in this way on the TIMIT speech corpus. Our results show that certain types of perceptron training yield consistently significant and rapid reductions in phone error rates.

Unsupervised Search-Based Structured Prediction

Hal Daumé III

We describe an adaptation and application of a search-based structured prediction algorithm "Searn" to unsupervised learning problems. We show that it is possible to reduce unsupervised learning to supervised learning and demonstrate a high-quality unsupervised shift-reduce parsing model. We additionally show a close connection between unsupervised Searn and expectation maximization. Finally, we demonstrate the efficacy of a semi-supervised extension. The key idea that enables this development is an application of the *predict-self* idea for unsupervised learning.

Sparse Higher Order Conditional Random Fields for improved sequence labeling

Xian Qian, Xiaoqian Jiang, Qi Zhang, Xuanjing Huang, and Lide Wu

In real sequence labeling tasks, statistics of many higher order features are not sufficient due to the training data sparseness, very few of them are useful. We describe Sparse Higher Order Conditional Random Fields (SHO-CRFs), which are able to handle local features and sparse higher order features together using a novel tractable exact inference algorithm. Our main insight is that states and transitions with same potential functions can be grouped together, and inference is performed on the grouped states and transitions. Though the complexity is not polynomial, SHO-CRFs are still efficient in practice because of the feature sparseness. Experimental results on optical character recognition and Chinese organization name recognition show that with the same higher order feature set, SHO-CRFs significantly outperform previous approaches.

Detecting the Direction of Causal Time Series

Jonas Peters, Dominik Janzing, Arthur Gretton, and Bernhard Schölkopf

We propose a method that detects the true direction of time series, by fitting an autoregressive moving average model to the data. Whenever the noise is independent of the previous samples for one ordering of the observations, but dependent for the opposite ordering, we infer the former direction to be the true one. We prove that our method works in the population case as long as the noise of the process is not normally distributed (for the latter case, the direction is not identifiable). A new and important implication of our result is that it confirms a fundamental conjecture in causal reasoning — if after regression the noise is independent of signal for one direction and dependent for the other, then the former represents the true causal direction — in the case of time series. We test our approach on two types of data: simulated data sets conforming to our modeling assumptions, and real world EEG time series. Our method makes a decision for a significant fraction of both data sets, and these decisions are mostly correct. For real world data, our approach outperforms alternative solutions to the problem of time direction recovery.

Session 3E: Topic Models — Tuesday, 10:20 – 12:20, Leacock 219

Chair: Tom Diettrich

Evaluation Methods for Topic Models

Hanna Wallach, Iain Murray, Ruslan Salakhutdinov, and David Mimno

A natural evaluation metric for statistical topic models is the probability of held-out documents given a trained model. While exact computation of this probability is intractable, several estimators for this probability have been used in the topic modeling literature, including the harmonic mean method and empirical likelihood method. In this paper, we demonstrate experimentally that commonly-used methods are unlikely to accurately estimate the probability of held-out documents, and propose two alternative methods that are both accurate and efficient.

Accounting for Burstiness in Topic Models

Gabriel Doyle and Charles Elkan

Many different topic models have been used successfully for a variety of applications. However, even state-of-the-art topic models suffer from the important flaw that they do not capture the tendency of words to appear in bursts; it is a fundamental property of language that if a word is used once in a document, it is more likely to be used again. We introduce a topic model that uses Dirichlet compound multinomial (DCM) distributions to model this burstiness phenomenon. On both text and non-text datasets, the new model achieves better held-out likelihood than standard latent Dirichlet allocation (LDA). It is straightforward to incorporate the DCM extension into topic models that are more complex than LDA.

Topic-Link LDA: Joint Models of Topic and Author Community

Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc

Given a large-scale linked document collection, such as a collection of blog posts or a research literature archive, there are two fundamental problems that have generated a lot of interest in the research community. One is to identify a set of high-level topics covered by the documents in the collection; the other is to uncover and analyze the social network of the authors of the documents. So far these problems have been viewed as separate problems and considered independently from each other. In this paper we argue that these two problems are in fact inter-dependent and should be addressed together. We develop a Bayesian hierarchical approach that performs topic modeling and author community discovery in one unified framework. The effectiveness of our model is demonstrated on two blog data sets in different domains and one research paper citation data from CiteSeer.

MedLDA: Maximum Margin Supervised Topic Models for Regression and Classification

Jun Zhu, Amr Ahmed, and Eric Xing

Supervised topic models utilize document's side information for discovering predictive low dimensional representations of documents; and existing models apply likelihood-based estimation. In this paper, we present a max-margin supervised topic model for both continuous and categorical response variables. Our approach, the maximum entropy discrimination latent Dirichlet allocation (MedLDA), utilizes the max-margin principle to train supervised topic models and estimate predictive topic representations that are arguably more suitable for prediction. We develop efficient variational methods for posterior inference and demonstrate qualitatively and quantitatively the advantages of MedLDA over likelihood-based topic models on movie review and 20 Newsgroups data sets.

Independent Factor Topic Models

Duangmanee Putthividhya, Hagai Attias, and Srikantan Nagarajan

Topic models such as Latent Dirichlet Allocation (LDA) and Correlated Topic Model (CTM) have recently emerged as powerful statistical tools for text document modeling. In this paper, we improve upon CTM and propose Independent Factor Topic Models (IFTM) which use linear latent variable models to uncover the hidden sources of correlation between topics. There are 2 main contributions of this work. First, by using a sparse source prior model, we can directly visualize sparse patterns of topic correlations. Secondly, the conditional independence assumption implied in the use of latent source variables allows the objective function to factorize, leading to a fast Newton-Ralphson based variational inference algorithm. Experimental results on synthetic and real data show that IFTM runs on average 3-5 times faster than CTM, while giving competitive performance as measured by perplexity and log-likelihood of held-out data.

Session 3F: Transfer and MultiTask Learning — Tuesday, 10:20 – 12:20, Leacock 132

Chair: Kai Yu

Deep Transfer via Second-Order Markov Logic

Jesse Davis and Pedro Domingos

Standard inductive learning requires that training and test instances come from the same distribution. Transfer learning seeks to remove this restriction. In shallow transfer, test instances are from the same domain, but have a different distribution. In deep transfer, test instances are from a different domain entirely (i.e., described by different predicates). Humans routinely perform deep transfer, but few learning systems, if any, are capable of it. In this paper we propose an approach based on a form of second-order Markov logic. Our algorithm discovers structural regularities in the source domain in the form of Markov logic formulas with predicate variables, and instantiates these formulas with predicates from the target domain. Using this approach, we have successfully transferred learned knowledge between molecular biology, social network and Web domains. The discovered patterns include broadly useful properties of predicates, like symmetry and transitivity, and relations among predicates, like various forms of homophily.

Feature Hashing for Large Scale Multitask Learning

Kilian Weinberger, Anirban Dasgupta, Josh Attenberg, John Langford, and Alex Smola

Empirical evidence suggests that hashing is an effective strategy for dimensionality reduction and practical nonparametric estimation. In this paper we provide exponential tail bounds for feature hashing and show that the interaction between random subspaces is negligible with high probability. We demonstrate the feasibility of this approach with experimental results for a new use case — multitask learning with hundreds of thousands of tasks.

A Convex Formulation for Learning Shared Structures from Multiple Tasks

Jianhui Chen, Lei Tang, Jun Liu, and Jieping Ye

Multi-task learning (MTL) aims to improve generalization performance by learning multiple related tasks simultaneously. In this paper, we consider the problem of learning shared structures from multiple related tasks. We present an improved formulation (iASO) for multi-task learning based on the non-convex alternating structure optimization (ASO) algorithm, in which all tasks are related by a shared feature representation. We convert iASO, a non-convex formulation, into a relaxed convex one, which is, however, not scalable to large data sets due to its complex constraints. We propose an alternating optimization (cASO) algorithm which solves the convex relaxation efficiently, and further show that cASO converges to a global optimum. In addition, we present a theoretical condition, under which cASO can find a globally optimal solution to iASO. Experiments on several benchmark data sets confirm our theoretical analysis.

EigenTransfer: A Unified Framework for Transfer Learning

Wenyuan Dai, Ou Jin, Gui-Rong Xue, Qiang Yang, and Yong Yu

This paper proposes a general framework, called EigenTransfer, to tackle a variety of transfer learning problems, e.g. cross-domain learning, self-taught learning, etc. Our basic idea is to construct a graph to represent the target transfer learning task. By learning the spectra of a graph which represents a learning task, we obtain a set of eigenvectors that reflect the intrinsic structure of the task graph. These eigenvectors can be used as the new features which transfer the knowledge from auxiliary data to help classify target data. Given an arbitrary non-transfer learner (e.g. SVM) and a particular transfer learning task, EigenTransfer can produce a *transfer learner* accordingly for the target transfer learning task. We apply EigenTransfer on three different transfer learning tasks, cross-domain learning, cross-category learning and self-taught learning, to demonstrate its unifying ability, and show through experiments that EigenTransfer can greatly outperform several representative non-transfer learners.

Domain Adaptation from Multiple Sources via Auxiliary Classifiers

Lixin Duan, Ivor W. Tsang, Dong Xu, and Tat-Seng Chua

We propose a multiple source domain adaptation method, referred to as Domain Adaptation Machine (DAM), to learn a robust decision function (referred to as target classifier) for label prediction of patterns from the target domain by leveraging a set of pre-computed classifiers (referred to as auxiliary/source classifiers) independently learned with the labeled patterns from multiple source domains. We introduce a new datadependent regularizer based on smoothness assumption into Least-Squares SVM (LS-SVM), which enforces that the target classifier shares similar decision values with the auxiliary classifiers from relevant source domains on the unlabeled patterns of the target domain. In addition, we employ a sparsity regularizer to learn a sparse target classifier. Comprehensive experiments on the challenging TRECVID 2005 corpus demonstrate that DAM outperforms the existing multiple source domain adaptation methods for video concept detection in terms of effectiveness and efficiency.

Session 4A: Weak Supervision — Tuesday, 14:00 – 16:00, Leacock 219

Chair: Fei Sha

Semi-Supervised Learning Using Label Mean

Yu-Feng Li, James T. Kwok, and Zhi-Hua Zhou

Semi-Supervised Support Vector Machines (S3VMs) typically directly estimate the label assignments for the unlabeled instances. This is often inefficient even with recent advances in the efficient training of the (supervised) SVM. In this paper, we show that S3VMs, with knowledge of the means of the class labels of the unlabeled data, is closely related to the supervised SVM with known labels on all the unlabeled data. This motivates us to first estimate the label means of the unlabeled data. Two versions of the meanS3VM, which work by maximizing the margin between the label means, are proposed. The first one is based on multiple kernel learning, while the second one is based on alternating optimization. Experiments show that both of the proposed algorithms achieve highly competitive and sometimes even the best performance as compared to the state-of-the-art semi-supervised learners. Moreover, they are more efficient than existing S3VMs.

Partially Supervised Feature Selection with Regularized Linear Models

Thibault Helleputte and Pierre Dupont

This paper addresses feature selection techniques for classification of high dimensional data, such as those produced by microarray experiments. Some prior knowledge may be available in this context to bias the selection towards some dimensions (genes) a priori assumed to be more relevant. We propose a feature selection method making use of this partial supervision. It extends previous works on embedded feature selection with linear models including regularization to enforce sparsity. A practical approximation of this technique reduces to standard SVM learning with iterative rescaling of the inputs. The scaling factors depend here on the prior knowledge but the final selection may depart from it. Practical results on several microarray data sets show the benefits of the proposed approach in terms of the stability of the selected gene lists with improved classification performances.

Optimal Reverse Prediction: A Unified Perspective on Supervised, Unsupervised and Semi-supervised Learning

Linli Xu, Martha White, and Dale Schuurmans

Training principles for unsupervised learning are often derived from motivations that appear to be independent of supervised learning, causing a proliferation of semisupervised training methods. In this paper we present a simple unification of several supervised and unsupervised training principles through the concept of optimal reverse prediction: predict the inputs from the target labels, optimizing both over model parameters and any missing labels. In particular, we show how supervised least squares, principal components analysis, k-means clustering and normalized graph-cut clustering can all be expressed as instances of the same training principle, differing only in constraints made on the target labels. Natural forms of semi-supervised regression and classification are then automatically derived, yielding semi-supervised learning algorithms for regression and classification that, surprisingly, are novel and refine the state of the art. These algorithms can all be combined with standard regularizers and made non-linear via kernels.

Supervised Learning from Multiple Experts: Whom to Trust When Everyone Lies a Bit

Vikas Raykar, Shipeng Yu, Linda Zhao, Anna Jerebko, Charles Florin, Gerardo Valadez, Luca Bogoni, and Linda Moy

We describe a probabilistic approach for supervised learning when we have multiple experts/annotators providing (possibly noisy) labels but no absolute gold standard. The proposed algorithm evaluates the different experts and also gives an estimate of the actual hidden labels. Experimental results indicate that the proposed method clearly beats the commonly used majority voting baseline.

Good Learners for Evil Teachers

Ofer Dekel and Ohad Shamir

We consider a supervised machine learning scenario where labels are provided by a heterogeneous set of teachers, some of which are mediocre, incompetent, or perhaps even malicious. We present an algorithm, built on the SVM framework, that explicitly attempts to cope with low-quality and malicious teachers by decreasing their influence on the learning process. Our algorithm does not receive any prior information on the teachers, nor does it resort to repeated labeling (where each example is labeled by multiple teachers). We provide a theoretical analysis of our algorithm and demonstrate its merits empirically. Finally, we present a second algorithm with promising empirical results but without a formal analysis.

Session 4B: Learning Structures — Tuesday, 14:00 – 16:00, Leacock 132

Chair: Pedro Domingos

Structure Learning with Independent Non-Identically Distributed Data

Robert Tillman

There are well known algorithms for learning the structure of directed and undirected graphical models from data, but nearly all assume that the data consists of a single i.i.d. sample. In contexts such as fMRI analysis, data may consist of an ensemble of independent samples from a common data generating mechanism which may not have identical distributions. Pooling such data can result in a number of well known statistical problems so each sample must be analyzed individually, which offers no increase in power due to the presence of multiple samples. We show how existing constraint based methods can be modified to learn structure from the aggregate of such data in a statistically sound manner. The prescribed method is simple to implement and based on existing statistical methods employed in metaanalysis and other areas, but works surprisingly well in this context where there are increased concerns due to issues such as retesting. We report results for directed models, but the method given is just as applicable to undirected models.

Structure Learning of Bayesian Networks using Constraints

Cassio P. de Campos, Zhi Zeng, and Qiang Ji

This paper addresses exact learning of Bayesian network structure from data and expert's knowledge based on score functions that are decomposable. First, it describes useful properties that strongly reduce the time and memory costs of many known methods such as hill-climbing, dynamic programming and sampling variable orderings. Secondly, a branch and bound algorithm is presented that integrates parameter and structural constraints with data in a way to guarantee global optimality with respect to the score function. It is an any-time procedure because, if stopped, it provides the best current solution and an estimation about how far it is from the global solution. We show empirically the advantages of the properties and the constraints, and the applicability of the algorithm to large data sets (up to one hundred variables) that cannot be handled by other current methods (limited to around 30 variables).

Learning Structurally Consistent Undirected Probabilistic Graphical Models

Sushmita Roy, Terran Lane, and Margaret Werner-Washburne

In many real-world domains, undirected graphical models such as Markov random fields provide a more natural representation of the statistical dependency structure than directed graphical models. Unfortunately, structure learning of undirected graphs using likelihood-based scores remains difficult because of the intractability of computing the partition function. We describe a new Markov random field structure learning algorithm, motivated by canonical parameterization of Abbeel et al. We provide computational improvements on their parameterization by learning per-variable canonical factors, which makes our algorithm suitable for domains with hundreds of nodes. We compare our algorithm against several algorithms for learning undirected and directed models on simulated and real datasets from biology. Our algorithm frequently outperforms existing algorithms, producing higher-quality structures, suggesting that enforcing consistency during structure learning is beneficial for learning undirected graphs.

Sparse Gaussian Graphical Models with Unknown Block Structure

Benjamin M. Marlin and Kevin P. Murphy

Recent work has shown that one can learn the structure of Gaussian Graphical Models by imposing an L1 penalty on the precision matrix, and then using efficient convex optimization methods to find the penalized maximum likelihood estimate. This is similar to performing MAP estimation with a prior that prefers sparse graphs. In this paper, we use the stochastic block model as a prior. This prefer graphs that are blockwise sparse, but unlike previous work, it does not require that the blocks or groups be specified a priori. The resulting problem is no longer convex, but we devise an efficient variational Bayes algorithm to solve it. We show that our method has better test set likelihood on two different datasets (motion capture and gene expression) compared to independent L1, and can match the performance of group L1 using manually created groups.

Learning Markov Logic Network Structure via Hypergraph Lifting

Stanley Kok and Pedro Domingos

Markov logic networks (MLNs) combine logic and probability by attaching weights to first-order clauses, and viewing these as templates for fea- tures of Markov networks. Learning MLN struc- ture from a relational database involves learn- ing the clauses and weights. The state-of-the-art MLN structure learners all involve some element of greedily generating candidate clauses, and are susceptible to local optima. To address this prob- lem, we present an approach that directly utilizes the data in constructing candidates. A relational database can be viewed as a hypergraph with con- stants as nodes and relations as hyperedges. We find paths of true ground atoms in the hypergraph that are connected via their arguments. To make this tractable (there are exponentially many paths in the hypergraph), we lift the hypergraph by jointly clustering the constants to form higher- level concepts, and find paths in it. We variabilize the ground atoms in each path, and use them to form clauses, which are evaluated using a pseudo- likelihood measure. In our experiments on three real-world datasets, we find that our algorithm outperforms the state-of-the-art approaches.

Session 4C: Active Learning — Tuesday, 14:00 – 16:00, Leacock 26

Chair: Yoav Freund

Learning to Segment from a Few Well-Selected Training Images

Alireza Farhangfar, Russell Greiner, and Csaba Szepesvári

We address the task of actively learning a segmentation system: given a large number of unsegmented images, and access to an oracle that can segment a given image, decide which images to provide, to quickly produce a segmenter (here, a discriminative random field) that is accurate over this distribution of images. We extend the standard models for active learner to define a system for this task that first selects the image whose expected label will most reduce the uncertainty of the other unlabeled images, and thenafter greedily selects, from the pool of unsegmented images, the most informative image. The results of our experiments, over two real-world datasets (segmenting brain tumors within magnetic resonance images; and segmenting the sky in real images) show that training on very few informative images (here, as few as 2) can produce a segmenter that is as good as training on the entire dataset.

Importance Weighted Active Learning

Alina Beygelzimer, Sanjoy Dasgupta, and John Langford

We present a practical and statistically consistent scheme for actively learning binary classifiers under general loss functions. Our algorithm uses importance weighting to correct sampling bias, and by controlling the variance, we are able to give rigorous label complexity bounds for the learning process.

Experiments on passively labeled data show that this approach reduces the label complexity required to achieve good predictive performance on many learning problems.

Learning from Measurements in Exponential Families

Percy Liang, Michael I. Jordan, and Dan Klein

Given a model family and a set of unlabeled examples, one could either label specific examples or state general constraints—both provide information about the desired model. In general, what is the most cost-effective way to learn? To address this question, we introduce measurements, a general class of mechanisms for providing information about a target model. We present a Bayesian decision-theoretic framework, which allows us to both integrate diverse measurements and choose new measurements to make. We use a variational inference algorithm, which exploits exponential family duality. The merits of our approach are demonstrated on two sequence labeling tasks.

Online Feature Elicitation in Interactive Optimization

Craig Boutilier, Kevin Regan, and Paolo Viappiani

Most models of utility elicitation in decision support and interactive optimization assume a predefined set of "catalog" features over which user preferences are expressed. However, users may differ in the features over which they are most comfortable expressing their preferences. In this work we consider the problem of feature elicitation: a user's utility function is expressed using features whose definition (in terms of "catalog" features) is unknown. We cast this as a problem of concept learning, but whose goal is to identify only enough about the concept to enable a good decision to be recommended. We describe computational procedures for identifying optimal alternatives w.r.t. minimax regret in the presence of concept uncertainty; and describe several heuristic query strategies that focus on reduction of relevant concept uncertainty.

Uncertainty Sampling and Transductive Experimental Design for Active Dual Supervision

Vikas Sindhwani, Prem Melville, and Richard Lawrence

Dual supervision refers to the general setting of learning from both labeled examples as well as labeled features. Labeled features are naturally available in tasks such as text classification where it is frequently possible to provide domain knowledge in the form of words that associate strongly with a class. In this paper, we consider the novel problem of active dual supervision, or, how to optimally query an example and feature labeling oracle to simultaneously collect two different forms of supervision, with the objective of building the best classifier in the most cost effective manner. We apply classical uncertainty and experimental design based active learning schemes to graph/kernel-based dual supervision models. Empirical studies confirm the potential of these schemes to significantly reduce the cost of acquiring labeled data for training high-quality models.

Session 4D: Lassos and Other L1s — Tuesday, 14:00 – 16:00, Leacock 232

Chair: Volker Roth

Stochastic Methods for ℓ_1 Regularized Loss Minimization

Shai Shalev-Shwartz and Ambuj Tewari

We describe and analyze two stochastic methods for ℓ_1 regularized loss minimization problems, such as the Lasso. The first method updates the weight of a single feature at each iteration while the second method updates the entire weight vector but only uses a single training example at each iteration. In both methods, the choice of feature/example is uniformly at random. Our theoretical runtime analysis suggests that the stochastic methods should outperform state-of-the-art deterministic approaches, including their deterministic counterparts, when the size of the problem is large. We demonstrate the advantage of stochastic methods by experimenting with synthetic and natural data sets.

Blockwise Coordinate Descent Procedures for the Multi-Task Lasso, with Applications to Neural Semantic Basis Discovery

Han Liu, Mark Palatucci, and Jian Zhang

We develop a cyclical blockwise coordinate descent algorithm for the multi-task Lasso that efficiently solves problems with thousands of features and tasks. The main result shows that a closed-form Winsorization operator can be obtained for the sup-norm penalized least squares regression. This allows the algorithm to find solutions to very large-scale problems far more efficiently than existing methods. This result complements the pioneering work of Friedman, et al. (2007) for the single-task Lasso. As a case study, we use the multi-task Lasso as a variable selector to discover a semantic basis for predicting human neural activation. The learned solution outperforms the standard basis for this task on the majority of test participants, while requiring far fewer assumptions about cognitive neuroscience. We demonstrate how this learned basis can yield insights into how the brain represents the meanings of words.

An Efficient Projection for $l_{1,\infty}$ Regularization

Ariadna Quattoni, Xavier Carreras, Michael Collins, and Trevor Darrell

In recent years the L1,Infinity norm has been proposed for joint regularization. In essence, this type of regularization aims at extending the L1 framework for learning sparse models to a setting where the goal is to learn a set of jointly sparse models. In this paper we derive a simple and effective projected gradient method for optimization of L1,Infinity regularized problems. The main challenge in developing such a method resides on being able to compute efficient projections to the L1,Infinity ball. We present an algorithm that works in O(n log n) time and O(n) memory where n is the number of parameters. We test our algorithm in a multi-task image annotation problem. Our results show that L1,Infinity leads to better performance than both L2 and L1 regularization and that it is is effective in discovering jointly sparse solutions.

An Accelerated Gradient Method for Trace Norm Minimization

Shuiwang Ji and Jieping Ye

We consider the minimization of a smooth loss function regularized by the trace norm of the matrix variable. Such formulation finds applications in many machine learning tasks including multi-task learning, matrix classification, and matrix completion. The standard semidefinite programming formulation for this problem is computationally expensive. In addition, due to the non-smoothness nature of the trace norm, the optimal first-order black-box method for solving such class of problems converges as O(1/sqrt(k)), where k is the iteration counter. In this paper, we exploit the special structure of the trace norm, based on which we propose an extended gradient algorithm that converges as O(1/k). We further propose an accelerated gradient algorithm, which achieves the optimal convergence rate of O($1/k^2$) for smooth problems. Experiments on multi-task learning problems demonstrate the efficiency of the proposed algorithms.

Group Lasso with Overlaps and Graph Lasso

Laurent Jacob, Guillaume Obozinski, and Jean-Philippe Vert

We propose a new penalty function which, when used as regularization for empirical risk minimization procedures, leads to sparse estimators. The support of the sparse vector is typically a union of potentially overlapping groups of covariates defined a priori, or a set of covariates which tend to be connected to each other when a graph of covariates is given. We study theoretical properties of the estimator, and illustrate its behavior on simulated and breast cancer gene expression data.

Session 4E: Document Collections — Tuesday, 14:00 – 16:00, Leacock 210

Chair: Johannes Fuernkranz

Bayesian Clustering for Email Campaign Detection

Peter Haider and Tobias Scheffer

We discuss the problem of clustering elements according to the sources that have generated them. For elements that are characterized by independent binary attributes, a closed-form Bayesian solution exists. We derive a solution for the case of dependent attributes that is based on a transformation of the instances into a space of independent feature functions. We derive an optimization problem that produces a mapping into a space of independent binary feature vectors; the features can reflect arbitrary dependencies in the input space. This problem setting is motivated by the application of spam filtering for email service providers. Spam traps deliver a real-time stream of messages known to be spam. If elements of the same campaign can be recognized reliably, entire spam and phishing campaigns can be contained. We present a case study that evaluates Bayesian clustering for this application.

A Novel Lexicalized HMM-based Learning Framework for Web Opinion Mining

Wei Jin and Hung Hay Ho

Merchants selling products on the Web often ask their customers to share their opinions and hands-on experiences on products they have purchased. As e-commerce is becoming more and more popular, the number of customer reviews a product receives grows rapidly. This makes it difficult for a potential customer to read them to make an informed decision on whether to purchase the product. In this research, we aim to mine customer reviews of a product and extract highly specific product related entities on which reviewers express their opinions. Opinion expressions and sentences are also identified and opinion orientations for each recognized product entity are classified as positive or negative. Different from previous approaches that have mostly relied on natural language processing techniques or statistic information, we propose a novel machine learning framework using lexicalized HMMs. The approach naturally integrates linguistic features, such as part-of-speech and surrounding contextual clues of words into automatic learning. The experimental results demonstrate the effectiveness of the proposed approach in web opinion mining and extraction from product reviews.

Learning Spectral Graph Transformations for Link Prediction

Jérôme Kunegis and Andreas Lommatzsch

We present a unified framework for learning link prediction and edge weight prediction functions in large networks, based on the transformation of a graph's algebraic spectrum. Our approach generalizes several graph kernels and dimensionality reduction methods and provides a method to estimate their parameters efficiently. We show how the parameters of these prediction functions can be learned by reducing the problem to a one-dimensional regression problem whose runtime only depends on the method's reduced rank and that can be inspected visually. We derive variants that apply to undirected, weighted, unweighted, unipartite and bipartite graphs. We evaluate our method experimentally using examples from social networks, collaborative filtering, trust networks, citation networks, authorship graphs and hyperlink networks.

Interactively Optimizing Information Retrieval Systems as a Dueling Bandits Problem

Yisong Yue and Thorsten Joachims

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

Transfer Learning for Collaborative Filtering via a Rating-Matrix Generative Model

Bin Li, Qiang Yang, and Xiangyang Xue

Cross-domain collaborative filtering solves the sparsity problem by transferring rating knowledge across multiple domains. In this paper, we propose a rating-matrix generative model (RMGM) for effective cross-domain collaborative filtering. We first show that the relatedness across multiple rating matrices can be established by finding a shared implicit cluster-level rating matrix, which is next extended to a cluster-level rating model. Consequently, a rating matrix of any related task can be viewed as drawing a set of users and items from a user-item joint mixture model as well as drawing the corresponding ratings from the cluster-level rating model. The combination of these two models gives the RMGM, which can be used to fill the missing ratings for both existing and new users. A major advantage of RMGM is that it can share the knowledge by pooling the rating data from multiple tasks even when the users and items of these tasks do not overlap. We evaluate the RMGM empirically on three real-world collaborative filtering data sets to show that RMGM can outperform the individual models trained separately.

Chair: Léon Bottou and Michael Littman

Curriculum Learning

Yoshua Bengio, Jérôme Louradour, Ronan Collobert and Jason Weston

Humans and animals learn much better when the examples are not randomly presented but organized in a meaningful order which illustrates gradually more concepts, and more complex ones. Here, we formalize such training strategies in the context of machine learning, and call them "curriculum learning". In the context of recent research studying the difficulty of training in the presence of non-convex training criteria (for deep deterministic and stochastic neural networks), we explore curriculum learning in various set-ups. The experiments show that significant improvements in generalization can be achieved by using a particular curriculum, i.e., the selection and order of training examples. We hypothesize that curriculum learning has both an effect on the speed of convergence of the training process to a minimum and, in the case of non-convex criteria, on the quality of the local minima obtained: curriculum learning can be seen as a particular form of continuation method (a general strategy for global optimization of non-convex functions).

Herding Dynamical Weights to Learn

Max Welling

A new "herding" algorithm is proposed which directly converts observed moments into a sequence of pseudo-samples. The pseudo-samples respect the moment constraints and may be used to estimate (unobserved) quantities of interest. The procedure allows us to sidestep the usual approach of first learning a joint model (which is intractable) and then sampling from that model (which can easily get stuck in a local mode). Moreover, the algorithm is fully deterministic, avoiding random number generation) and does not need expensive operations such as exponentiation.

Sequential Bayesian Prediction in the Presence of Changepoints

Roman Garnett, Michael Osborne, and Stephen Roberts

We introduce a new sequential algorithm for making robust predictions in the presence of changepoints. Unlike previous approaches, which focus on the problem of detecting and locating changepoints, our algorithm focuses on the problem of making predictions even when such changes might be present. We introduce nonstationary covariance functions to be used in Gaussian process prediction that model such changes, then proceed to demonstrate how to effectively manage the hyperparameters associated with those covariance functions. By using Bayesian quadrature, we can integrate out the hyperparameters, allowing us to calculate the marginal predictive distribution. Furthermore, if desired, the posterior distribution over putative changepoint locations can be calculated as a natural byproduct of our prediction algorithm.

Model-Free Reinforcement Learning as Mixture Learning

Nikos Vlassis and Marc Toussaint

We cast model-free reinforcement learning as the problem of maximizing the likelihood of a probabilistic mixture model via sampling, addressing both the infinite and finite horizon cases. We describe a Stochastic Approximation EM algorithm for likelihood maximization that, in the tabular case, is equivalent to a non-bootstrapping optimistic policy iteration algorithm like Sarsa(1) that can be applied both in MDPs and POMDPs. On the theoretical side, by relating the proposed stochastic EM algorithm to the family of optimistic policy iteration algorithms, we provide new tools that permit the design and analysis of algorithms in that family. On the practical side, preliminary experiments on a POMDP problem demonstrated encouraging results.

Active Learning for Directed Exploration of Complex Systems

Michael Burl and Esther Wang

Physics-based simulation codes are widely used in science and engineering to model complex systems that would be infeasible or impossible to study otherwise. While such codes generally provide the highest-fidelity representation of system behavior, they are often so slow to run that it is difficult to gain significant insight into the system. For example, conducting an exhaustive sweep over a d-dimensional input parameter space with k-steps along each dimension requires k**d simulation trials (translating into k**d CPU-days for one of our current simulations). An alternative is directed exploration in which the next simulation trials are cleverly chosen at each step. Given the results of previous trials, standard supervised learning techniques (SVM, KDE, GP) are applied to build up simplified predictive models of system behavior. These models are then used within an active learning framework to identify the most valuable trials to run next. Several active learning strategies are examined including a recently-proposed information-theoretic approach. Performance is evaluated on a set of thirteen challenging oracles, which serve as surrogates for the more expensive simulations and enable easy replication of the experiments by other researchers.

Session 5A: Algorithms — Wednesday, 10:20 – 12:20, Leacock 26

Chair: Alex Smola

Proximal Regularization for Online and Batch Learning

Chuong Do, Quoc Le, and Chuan-Sheng Foo

Many learning algorithms rely on the curvature (in particular, strong convexity) of regularized objective functions to provide good theoretical performance guarantees. In practice, the choice of regularization penalty that gives the best testing set performance may result in objective functions with little or even no curvature. In these cases, algorithms designed specifically for regularized objectives often either fail completely or require some modification that involves a substantial compromise in performance.

We present new online and batch algorithms for training a variety of supervised learning models (such as SVMs, logistic regression, structured prediction models, and CRFs) under conditions where the optimal choice of regularization parameter results in functions with low curvature. We employ a technique called proximal regularization, in which we solve the original learning problem via a sequence of modified optimization tasks whose objectives are chosen to have greater curvature than the original problem. Theoretically, our algorithms achieve low regret bounds in the online setting and fast convergence in the batch setting. Experimentally, our algorithms improve upon state-of-the-art techniques, including Pegasos and bundle methods, on medium and large-scale SVM and structured learning tasks.

A Majorization-Minimization Algorithm for (Multiple) Hyperparameter Learning

Chuan-Sheng Foo, Chuong Do, and Andrew Ng

We present a general Bayesian framework for hyperparameter tuning in L_2-regularized supervised learning models. Paradoxically, our algorithm works by first analytically integrating out the hyperparameters from the model. We find a local optimum of the resulting nonconvex optimization problem efficiently using a majorization-minimization (MM) algorithm, in which the non-convex problem is reduced to a series of convex L_2-regularized parameter estimation tasks. The principal appeal of our method is its simplicity: the updates for choosing the L_2-regularized subproblems in each step are trivial to implement (or even perform by hand), and each subproblem can be efficiently solved by adapting existing solvers. Empirical results on a variety of supervised learning models show that our algorithm is competitive with both grid-search and gradient-based algorithms, but is more efficient and far easier to implement.

A Least Squares Formulation for a Class of Generalized Eigenvalue Problems in Machine Learning

Liang Sun, Shuiwang Ji, and Jieping Ye

Many machine learning algorithms can be formulated as a generalized eigenvalue problem. One major limitation of such formulation is that the generalized eigenvalue problem is computationally expensive to solve especially for large-scale problems. In this paper, we show that under a mild condition, a class of generalized eigenvalue problems in machine learning can be formulated as a least squares problem. This class of problems include classical techniques such as Canonical Correlation Analysis (CCA), Partial Least Squares (PLS), Linear Discriminant Analysis (LDA), as well as Hypergraph Spectral Learning (HSL). As a result, various regularization techniques such as the 1-norm and 2-norm regularization can be readily incorporated into the formulation to improve model sparsity and generalization ability. In addition, the least squares formulation leads to efficient and scalable implementations based on the iterative conjugate gradient type algorithms. We report experimental results that confirm the established equivalence relationship. We also demonstrate the efficiency and effectiveness of the equivalent least squares formulations on large-scale problems. The presented analysis provides significant new insights into the relationship between the generalized eigenvalue and least squares problems in machine learning.

On Sampling-based Approximate Spectral Decomposition

Sanjiv Kumar, Mehryar Mohri, and Ameet Talwalkar

This paper addresses the problem of approximate singular value decomposition of large dense matrices that arises naturally in many machine learning applications. We discuss two recently introduced sampling-based spectral decomposition techniques: the Nystrom and the Column-sampling methods. We present a theoretical comparison between the two methods and provide novel insights regarding their suitability for various applications. We then provide experimental results motivated by this theory. Finally, we propose an efficient adaptive sampling technique to select informative columns from the original matrix. This novel technique outperforms standard sampling methods on a variety of datasets.

Efficient Euclidean Projections in Linear Time

Jun Liu and Jieping Ye

We consider the problem of computing the Euclidean projection of a vector of length n onto a closed convex set including the ℓ_1 ball and the specialized polyhedra employed in (Shalev-Shwartz & Singer, 2006). These problems have played building block roles in solving several ℓ_1-norm based sparse learning problems. Existing methods have a worst-case time complexity of $O(n \log n)$. In this paper, we propose to cast both Euclidean projections as root finding problems associated with specific auxiliary functions, which can be solved in linear time via bisection. We further make use of the special structure of the auxiliary functions, and propose an improved bisection algorithm. Empirical studies demonstrate that the proposed algorithms are much more efficient than the competing ones for computing the projections.

Session 5B: Priors in Probabilistic Models — Wednesday, 10:20 – 12:20, Leacock 132

Chair: Max Welling

Bayesian Inference for Plackett-Luce Ranking Models

John Guiver and Edward Snelson

This paper gives an efficient Bayesian method for inferring the parameters of a Plackett-Luce ranking model. Such models are parameterised distributions over rankings of a finite set of objects, and have typically been studied and applied within the psychometric, sociometric and econometric literature. The inference scheme is an application of Power EP (expectation propagation). The scheme is robust and can be readily applied to large scale data sets. The inference algorithm extends to variations of the basic Plackett-Luce model, including partial rankings. We show a number of advantages of the EP approach over the traditional maximum likelihood method. We apply the method to aggregate rankings of NASCAR racing drivers over the 2002 season, and also to model rankings of movie genres.

Incorporating Domain Knowledge into Topic Modeling via Dirichlet Forest Priors

David Andrzejewski, Xiaojin Zhu, and Mark Craven

Users of topic modeling methods often have knowledge about the composition of words that should have high or low probability in various topics. We incorporate such domain knowledge using a novel Dirichlet forest prior in a Latent Dirichlet Allocation framework. The prior is a mixture of Dirichlet tree distributions with special structures. We present its construction, and inference via collapsed Gibbs sampling. Experiments on synthetic and real datasets demonstrate our model's ability to follow and generalize beyond user-specified domain knowledge.

Nonparametric Factor Analysis with Beta Process Priors

John Paisley and Lawrence Carin

We propose a nonparametric extension to the factor analysis problem using a beta process prior. This beta process factor analysis (BP-FA) model allows for a dataset to be decomposed into a linear combination of a sparse set of factors, providing information on the underlying structure of the observations. As with the Dirichlet process, the beta process is a fully Bayesian conjugate prior, which allows for analytical posterior calculation and straightforward inference. We derive a variational Bayes inference algorithm and demonstrate the model on the MNIST handwritten digits and HGDP-CEPH cell line panel datasets.

Accelerated Gibbs Sampling for the Indian Buffet Process

Finale Doshi-Velez and Zoubin Ghahramani

We often seek to identify co-occurring hidden features in a set of observations. The Indian Buffet Process (IBP) provides a non-parametric prior on the features present in each observation, but current inference techniques for the IBP often scale poorly. The collapsed Gibbs sampler for the IBP has a running time cubic in the number of observations, and the uncollapsed Gibbs sampler, while linear, is often slow to mix. We present a new linear-time collapsed Gibbs sampler for conjugate models and demonstrate its efficacy on several large real-world data-sets.

A Stochastic Memoizer for Sequence Data

Frank Wood, Cédric Archambeau, Jan Gasthaus, Lancelot James, and Yee Whye Teh

We propose an unbounded-depth, hierarchical, Bayesian nonparametric model for discrete sequence data. This model can be estimated from a single training sequence, yet shares statistical strength between subsequent symbol predictive distributions in such a way that predictive performance generalizes well. The model builds on a specific parameterization of an unbounded-depth hierarchical Pitman-Yor process. We introduce analytic marginalization steps (using coagulation operators) to reduce this model to one that can be represented in time and space linear in the length of the training sequence. We show how to perform inference in such a model without truncation approximation and introduce fragmentation operators necessary to do predictive inference. We demonstrate the sequence memoizer by using it as a language model, achieving state-of-the-art results.

Session 5C: RL in High Order Environments Wednesday, 10:20 – 12:20, Leacock 219

Chair: Kurt Driessens

Binary Action Search for Learning Continuous-Action Control Policies

Jason Pazis and Michail Lagoudakis

Reinforcement Learning methods for controlling stochastic processes typically assume a small and discrete action space. While continuous action spaces are quite common in real-world problems, the most common approach still employed in practice is coarse discretization of the action space. This paper presents a novel method, called Binary Action Search, for realizing continuous-action policies by searching efficiently the entire action range through increment and decrement modifications to the values of the action variables according to an internal binary policy defined over an augmented state space. The proposed approach essentially approximates any continuous action space to arbitrary resolution and can be combined with any discrete-action reinforcement learning algorithm for learning continuous-action policies. Binary Action Search eliminates the restrictive modification steps of Adaptive Action Modification and requires no temporal action locality in the domain. Our approach is coupled with two well-known reinforcement learning algorithms (Least-Squares Policy Iteration and Fitted Q-Iteration) and its use and properties are thoroughly investigated and demonstrated on the continuous state-action Inverted Pendulum, Double Integrator, and Car on the Hill domains.

Predictive Representations for Policy Gradient in POMDPs

Abdeslam Boularias and Brahim Chaib-draa

We consider the problem of estimating the policy gradient in Partially Observable Markov Decision Processes (POMDPs) with a special class of policies that are based on Predictive State Representations (PSRs). We compare PSR policies to Finite-State Controllers (FSCs), which are considered as a standard model for policy gradient methods in POMDPs. We present a general actor-critic algorithm for learning both FSCs and PSR policies. The critic part computes a value function that has as variables the parameters of the policy. These latter parameters are gradually updated to maximize the value function. We show that the value function is polynomial for both FSCs and PSR policies, with a potentially smaller degree in the case of PSR policies. Therefore, the value function of a PSR policy can have less local optima than the equivalent FSC, and consequently, the gradient algorithm is more likely to converge to a global optimal solution.

Stochastic Search using the Natural Gradient

Sun Yi, Daan Wierstra, Tom Schaul, and Jürgen Schmidhuber

To optimize unknown 'fitness' functions, we introduce Natural Search, a novel stochastic search method that constitutes a principled alternative to standard evolutionary methods. It maintains a multinormal distribution on the set of solution candidates. The Natural Gradient is used to update the distribution's parameters in the direction of higher expected fitness, by efficiently calculating the inverse of the exact Fisher information matrix whereas previous methods had to use approximations. Other novel aspects of our method include optimal fitness baselines and importance mixing, a procedure adjusting batches with minimal numbers of fitness evaluations. The algorithm yields competitive results on a number of benchmarks.

Approximate Inference for Planning in Stochastic Relational Worlds

Tobias Lang and Marc Toussaint

Relational world models that can be learned from experience in stochastic domains have received significant attention recently. However, efficient planning using these models remains a major issue. We propose to convert learned noisy probabilistic relational rules into a structured dynamic Bayesian network representation. Predicting the effects of action sequences using approximate inference allows for planning in complex worlds. We evaluate the effectiveness of our approach for online planning in a 3D simulated blocksworld with an articulated manipulator and realistic physics. Empirical results show that our method can solve problems where existing methods fail.

Discovering Options from Example Trajectories

Peng Zang, Peng Zhou, David Minnen, and Charles Isbell

We present a novel technique for automated problem decomposition to address the problem of scalability in Reinforcement Learning. Our technique makes use of a set of near-optimal trajectories to discover *options* and incorporates them into the learning process, dramatically reducing the time it takes to solve the underlying problem. We run a series of experiments in two different domains and show that our method offers up to 30 fold speedup over the baseline.

Session 5D: Learning Theory — Wednesday, 10:20 – 12:20, Leacock 15

Chair: Sham Kakade

Nonparametric Estimation of the Precision-Recall Curve

Stéphan Clémençon and Nicolas Vayatis

The Precision-Recall (PR) curve is a widely used visual tool to evaluate the performance of scoring functions in regards to their capacities to discriminate between two populations. The purpose of this paper is to examine both theoretical and practical issues related to the statistical estimation of PR curves based on classification data. Consistency and asymptotic normality of the empirical counterpart of the PR curve in sup norm are rigorously established. Eventually, the issue of building confidence bands in the PR space is considered and a specific resampling procedure based on a smoothed and truncated version of the empirical distribution of the data is promoted. Arguments of theoretical and computational nature are presented to explain why such a bootstrap is preferable to a "naive" bootstrap in this setup.

Surrogate Regret Bounds for Proper Losses

Mark Reid and Robert Williamson

We present tight surrogate regret bounds for the class of proper (i.e., Fisher consistent) losses. The bounds generalise the margin-based bounds due to Bartlett et al. (2006). The proof uses Taylor's theorem and leads to new representations for loss and regret and a simple proof of the integral representation of proper losses. We also present a different formulation of a duality result of Bregman divergences which leads to a demonstration of the convexity of composite losses using canonical link functions.

Robust Bounds for Classification via Selective Sampling

Nicolò Cesa-Bianchi, Claudio Gentile, and Francesco Orabona

We introduce a new algorithm for binary classification in the selective sampling protocol. Our algorithm uses Regularized Least Squares (RLS) as base classifier, and for this reason it can be efficiently run in any RKHS. Unlike previous margin-based semi-supervised algorithms, our sampling condition hinges on a simultaneous upper bound on bias and variance of the RLS estimate under a simple linear label noise model. This fact allows us to prove performance bounds that hold for an arbitrary sequence of instances. In particular, we show that our sampling strategy approximates the margin of the Bayes optimal classifier to any desired accuracy ε by asking $\widetilde{O}\left(d/\varepsilon^2\right)$ queries (in the RKHS case d is replaced by a suitable spectral quantity). While these are the standard rates in the fully supervised i.i.d. case, the best previously known result in our harder setting was $\widetilde{O}\left(d^3/\varepsilon^4\right)$. Preliminary experiments show that some of our algorithms also exhibit a good practical performance.

PAC-Bayesian Learning of Linear Classifiers

Pascal Germain, Alexandre Lacasse, François Laviolette, and Mario Marchand

We present a general PAC-Bayes theorem from which all known PAC-Bayes bounds are simply obtained as particular cases. We also propose different learning algorithms for finding linear classifiers that minimize these PAC-Bayes risk bounds. These learning algorithms are generally competitive with both AdaBoost and the SVM.

Piecewise-Stationary Bandit Problems with Side Observations

Jia Yuan Yu and Shie Mannor

We consider a sequential decision problem where the rewards are generated by a piecewise-stationary distribution. However, the different reward distributions are unknown and may change at unknown instants. Our approach uses a limited number of side observations on past rewards, but does not require prior knowledge of the frequency of changes. In spite of the adversarial nature of the reward process, we provide an algorithm whose regret, with respect to the baseline with perfect knowledge of the distributions and the changes, is $O(k \log(T))$, where k is the number of changes up to time T. This is in contrast to the case where side observations are not available, and where the regret is at least $\Omega(\sqrt{T})$.

Session 5E: Paul Utgoff Memorial Session: Learning for Discrete Problems

Wednesday, 10:20 – 12:20, Leacock 232

Chair: Carla Brodley

Bandit-Based Optimization on Graphs with Application to Library Performance Tuning

Frédéric de Mesmay, Arpad Rimmel, Yevgen Voronenko, and Markus Püschel

The problem of choosing fast implementations for a class of recursive algorithms such as the fast Fourier transforms can be formulated as an optimization problem over the language generated by a suitably defined grammar. We propose a novel algorithm that solves this problem by reducing it to maximizing an objective function over the sinks of a directed acyclic graph. This algorithm valuates nodes using Monte-Carlo and grows a subgraph in the most promising directions by considering local maximum k-armed bandits. When used inside an adaptive linear transform library, it cuts down the search time by an order of magnitude compared to the existing algorithm. In some cases, the performance of the implementations found is also increased by up to 10% which is of considerable practical importance since it consequently improves the performance of all applications using the library.

Robust Feature Extraction via Information Theoretic Learning

Xiao-Tong Yuan and Bao-Gang Hu

In this paper, we present a robust feature extraction framework based on information-theoretic learning. Its formulated objective aims at dual targets, motivated by the Renyi's quadratic entropy of the features and the Renyi's cross entropy between features and class labels, respectively. This objective function reaps the advantages in robustness from both redescending M-estimator and manifold regularization, and can be efficiently optimized via half-quadratic optimization in an iterative manner. In addition, the popular algorithms LPP, SRDA and LapRLS for feature extraction are all justified to be the special cases within this framework. Extensive comparison experiments on several real-world data sets, with contaminated features or labels, well validate the encouraging gain in algorithmic robustness from this proposed framework.

Block-Wise Construction of Acyclic Relational Features with Monotone Irreducibility and Relevancy Properties

Ondřej Kuželka and Filip Železný

We describe an algorithm for constructing a set of acyclic conjunctive relational features by combining smaller conjunctive blocks. Unlike traditional level-wise approaches which preserve the monotonicity of frequency, our block-wise approach preserves a form of monotonicity of the irreducibility and relevancy feature properties, which are important in propositionalization employed in the context of classification learning. With pruning based on these properties, our block-wise approach efficiently scales to features including tens of first-order literals, far beyond the reach of state-of-the art propositionalization or inductive logic programming systems.

Rule Learning with Monotonicity Constraints

Wojciech Kotlowski and Roman Slowiński

In the ordinal classification with monotonicity constraints, it is assumed that the class label should increase with increasing values on the attributes. In this paper we aim at formalizing the approach to learning with monotonicity constraints from statistical point of view, which results in the algorithm for learning rule ensembles. The algorithm first "monotonizes" the data using a nonparametric classification procedure and then generates rule ensemble consistent with the training set. The procedure is justified by a theoretical analysis and verified in a computational experiment.

Grammatical Inference as a Principal Component Analysis Problem

Raphaël Bailly, François Denis, and Liva Ralaivola

One of the main problems in probabilistic grammatical inference consists in inferring a stochastic language, i.e. a probability distribution, in some class of probabilistic models, from a sample of words independently drawn according to a fixed unknown target distribution p. Here we consider the class of rational stochastic languages composed of stochastic languages that can be computed by muliplicity automata, which can be viewed as a generalization of probabilistic automata. Rational stochastic languages p have a useful algebraic characterization: all the mappings up:v-¿p(uv) lie in a finite dimensional vector subspace Vp of the vector space R(E) composed of all real-valued functions defined over E. Hence, a first step in the grammatial inference process can consist in identifying the subspace Vp. In this paper, we study the possibility of using principal component analysis to achieve this task. We provide an inference algorithm which computes an estimate of the target distribution. We prove some theoretical properties of this algorithm and we provide results from numerical simulations that confirm the relevance of our approach.

Session 6A: Structured Learning: Metric LearningWednesday, 14:00 – 16:00, Leacock 219

Chair: Sofus Attila Macskassy

Polyhedral Outer Approximations with Application to Natural Language Parsing

André Martins, Noah Smith, and Eric Xing

Recent approaches to learning structured predictors often require approximate inference for tractability; yet its effects on the learned model are unclear. Meanwhile, most learning algorithms act as if computational cost was constant within the model class. This paper sheds some light on the first issue by establishing risk bounds for max-margin learning with LP relaxed inference, and addresses the second issue by proposing a new paradigm that attempts to penalize "time-consuming" hypotheses. Our analysis relies on a geometric characterization of the outer polyhedra associated with the LP relaxation. We then apply these techniques to the problem of dependency parsing, for which a concise LP formulation is provided that handles non-local output features. A significant improvement is shown over arc-factored models.

On Primal and Dual Sparsity of Markov Networks

Jun Zhu and Eric Xing

Sparsity is a desirable property in high dimensional learning. The ℓ_1-norm regularization can lead to primal sparsity, while max-margin methods achieve dual sparsity; but achieving both in a single structured prediction model remains difficult. This paper presents an ℓ_1-norm max-margin Markov network (ℓ_1-M^{3}N), which enjoys both primal and dual sparsity, and analyzes its connections to the Laplace max-margin Markov network (LapM^{3}N), which inherits the dual sparsity of max-margin models but is pseudo-primal sparse. We show that ℓ_1-M^{3}N is an extreme case of LapM^{3}N when the regularization constant is infinity. We also show an equivalence between ℓ_1-M^{3}N and an adaptive M^{3}N, from which we develop a robust EM-style algorithm for ℓ_1-M^{3}N. We demonstrate the advantages of the simultaneously (pseudo-) primal and dual sparse models over the ones which enjoy either primal or dual sparsity on both synthetic and real data sets.

Learning Structural SVMs with Latent Variables

Chun-Nam Yu and Thorsten Joachims

We present a large-margin formulation and algorithm for structured output prediction that allows the use of latent variables. The paper identifies a particular formulation that covers a large range of application problems, while showing that the resulting optimization problem can generally be addressed using Concave-Convex Programming. The generality and performance of the approach is demonstrated on a motif-finding application, noun-phrase coreference resolution, and optimizing precision at k in information retrieval.

An Efficient Sparse Metric Learning in High-Dimensional Space via ℓ_1-Penalized Log-Determinant Regularization

Guo-Jun Qi, Jinhui Tang, Tat-Seng Chua, and Hong-Jiang Zhang

This paper proposes an efficient sparse metric learning algorithm in high dimensional space via an ℓ_1-penalized log-determinant regularization. Compare to the most existing distance metric learning algorithms, the proposed algorithm exploits the sparsity nature underlying the intrinsic high dimensional feature space. This sparsity prior of learning distance metric serves to regularize the complexity of the distance model especially in the "less example number p and high dimension d" setting. Theoretically, by analogy to the covariance estimation problem, we find the proposed distance learning algorithm has a consistent result at rate $\mathcal{O}\left(\sqrt{(m^2 \log d)/n}\right)$ to the target distance matrix with at most m nonzeros per row. Moreover, from the implementation perspective, this ℓ_1-penalized log-determinant formulation can be efficiently optimized in a block coordinate descent fashion which is much faster than the standard semi-definite programming which has been widely adopted in many other advanced distance learning algorithms. We compare this algorithm with other state-of-the-art ones on various datasets and competitive results are obtained.

Learning Instance Specific Distances Using Metric Propagation

De-Chuan Zhan, Ming Li, Yu-Feng Li, and Zhi-Hua Zhou

In many real-world applications, such as image retrieval, it would be natural to measure the distances from one instance to others using *instance specific distance* which captures the distinctions from the perspective of the concerned instance. However, there is no complete framework for learning instance specific distances since existing methods are incapable of learning such distances for test instance and unlabeled data. In this paper, we propose the ISD method to address this issue. The key of ISD is *metric propagation*, that is, propagating and adapting metrics of individual labeled examples to individual unlabeled instances. We formulate the problem into a convex optimization framework and derive efficient solutions. Experiments show that ISD can effectively learn instance specific distances for labeled as well as unlabeled instances. The metric propagation scheme can also be used in other scenarios.

Session 6B: Deep Architectures — Wednesday, 14:00 – 16:00, Leacock 132

Chair: Yann LeCun

Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representations

Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Ng

There has been much interest in unsupervised learning of hierarchical generative models such as deep belief networks. Scaling such models to full-sized, high-dimensional images remains a difficult problem. To address this problem, we present the convolutional deep belief network, a hierarchical generative model which scales to realistic image sizes. This model is translation-invariant and supports efficient bottom-up and top-down probabilistic inference. Key to our approach is probabilistic max-pooling, a novel technique which shrinks the representations of higher layers in a probabilistically sound way. Our experiments show that the algorithm learns useful high-level visual features, such as object parts, from unlabeled images of objects and natural scenes. We demonstrate excellent performance on several visual recognition tasks and show that our model can perform hierarchical (bottom-up and top-down) inference over full-sized images.

Using Fast Weights to Improve Persistent Contrastive Divergence

Tijmen Tieleman and Geoffrey Hinton

The most commonly used learning algorithm for restricted Boltzmann machines is contrastive divergence which starts a Markov chain at a data point and runs the chain for only a few iterations to get a cheap, low variance estimate of the sufficient statistics under the model. Tieleman (2008) showed that better learning can be achieved by estimating the model's statistics using a small set of persistent "fantasy particles" that are not reinitialized to data points after each weight update. With sufficiently small weight updates, the fantasy particles represent the equilibrium distribution accurately but to explain why the method works with much larger weight updates it is necessary to consider the interaction between the weight updates and the Markov chain. We show that the weight updates force the Markov chain to mix fast, and using this insight we develop an even faster mixing chain that uses an auxiliary set of "fast weights" to implement a temporary overlay on the energy landscape. The fast weights learn rapidly but also decay rapidly and do not contribute to the normal energy landscape that defines the model.

Large-Scale Deep Unsupervised Learning using Graphics Processors

Rajat Raina, Anand Madhavan, and Andrew Ng

The promise of unsupervised learning methods lies in their potential to use vast amounts of unlabeled data to learn complex, highly nonlinear models with millions of free parameters. We consider two well-known unsupervised learning methods, deep belief networks (DBNs) and sparse coding, that have recently been applied to a flurry of machine learning applications (Hinton & Salakhutdinov, 2006; Raina et al., 2007). Unfortunately, learning algorithms for both DBNs and sparse coding are too slow for large-scale applications, forcing researchers to focus on smaller-scale models, or smaller number of training examples.

In this paper, we suggest massively parallel methods to help resolve these problems. We argue that modern graphics processors far surpass the computational capabilities of multicore CPUs, and have the potential to revolutionize the applicability of deep unsupervised learning methods. We develop general principles for massively parallelizing unsupervised learning tasks using graphics processors. We show that these principles can be applied to successfully scaling up learning algorithms for both DBNs and sparse coding. Our implementation of DBN learning is upto 55 times faster than a dual-core CPU implementation. For example, we are able to reduce the time required to learn a four-layer DBN with 100 million free parameters from several weeks to around a single day. For sparse coding, we develop a simple, inherently parallel algorithm, that leads to a 5 to 16-fold speedup over previous methods.

Factored Conditional Restricted Boltzmann Machines for Modeling Motion Style

Graham Taylor and Geoffrey Hinton

The Conditional Restricted Boltzmann Machine (CRBM) is a recently proposed model for time series that has a rich, distributed hidden state and permits simple, exact inference. We present a new model, based on the CRBM that preserves its most important computational properties and includes multiplicative three-way interactions that allow the effective interaction weight between two units to be modulated by the dynamic state of a third unit. We factorize the three-way weight tensor implied by the multiplicative model, reducing the number of parameters from $O(N^3)$ to $O(N^2)$. The result is an efficient, compact model whose effectiveness we demonstrate by modeling human motion. Like the CRBM, our model can capture diverse styles of motion with a single set of parameters, and the three-way interactions greatly improve the model's ability to blend motion styles or to transition smoothly between them.

Deep Learning from Temporal Coherence in Video

Hossein Mobahi, Ronan Collobert, and Jason Weston

This work proposes a learning method for deep architectures that takes advantage of sequential data, in particular from the temporal coherence that naturally exists in unlabeled video recordings. That is, two successive frames are likely to contain the same object or objects. This coherence is used as a supervisory signal over the unlabeled data, and is used to improve the performance on a supervised task of interest. We demonstrate the effectiveness of this method on some pose invariant object and face recognition tasks.

Session 6C: Learning Actions and Sequences Wednesday, 14:00 – 16:00, Leacock 232

Chair: Nikos Vlassis

Robot Trajectory Optimization using Approximate Inference

Marc Toussaint

The general stochastic optimal control (SOC) problem in robotics scenarios is often too complex to be solved exactly and in near real time. A classical approximate solution is to first compute an optimal (deterministic) trajectory and then solve a local linear-quadratic-gaussian (LQG) perturbation model to handle the system stochasticity. We present a new algorithm for this approach which improves upon previous algorithms like iLQG. We consider a probabilistic model for which the maximum likelihood (ML) trajectory coincides with the optimal trajectory and which, in the LQG case, reproduces the classical SOC solution. The algorithm then utilizes approximate inference methods (similar to expectation propagation) that efficiently generalize to non-LQG systems. We demonstrate the algorithm on a simulated 39-DoF humanoid robot.

Trajectory Prediction: Learning to Map Situations to Robot Trajectories

Nikolay Jetchev and Marc Toussaint

Trajectory planning and optimization is a fundamental problem in articulated robotics. Algorithms used typically for this problem compute optimal trajectories from scratch in a new situation. In effect, extensive data is accumulated containing situations together with the respective optimized trajectories - but this data is in practice hardly exploited. The aim of this paper is to learn from this data. Given a new situation we want to predict a suitable trajectory which only needs minor refinement by a conventional optimizer. Our approach has two essential ingredients. First, to generalize from previous situations to new ones we need an appropriate situation descriptor - we propose a sparse feature selection approach to find such well-generalizing features of situations. Second, the transfer of previously optimized trajectories to a new situation should not be made in joint angle space - we propose a more efficient task space transfer of old trajectories to new situations. Experiments on trajectory optimization for a simulated humanoid reaching problem show that we can predict reasonable motion prototypes in new situations for which the refinement is much faster than an optimization from scratch.

Learning Complex Motions by Sequencing Simpler Motion Templates

Gerhard Neumann, Wolfgang Maass, and Jan Peters

Abstraction of complex, longer motor tasks into simpler elemental movements enables humans and animals to exhibit motor skills which have not yet been matched by robots. Humans intuitively decompose complex motions into smaller, simpler segments. For example when describing simple movements like drawing a triangle with a pen, we can easily name the basic steps of this movement.

Surprisingly, such abstractions have rarely been used in artificial motor skill learning algorithms. These algorithms typically choose a new action (such as a torque or a force) at a very fast time-scale. As a result, both policy and temporal credit assignment problem become unnecessarily complex - often beyond the reach of current machine learning methods.

We introduce a new framework for temporal abstractions in reinforcement learning (RL), i.e. RL with motion templates. We present a new algorithm for this framework which can learn high-quality policies by making only few abstract decisions.

Learning When to Stop Thinking and Do Something!

Barnabas Poczos, Yasin Abbasi-Yadkori, Csaba Szepesvri, Russell Greiner, and Nathan Sturtevant

An anytime algorithm is capable of returning a response to the given task at essentially any time; typically the quality of the response improves as the time increases. Here, we consider the challenge of learning when we should terminate such algorithms on each of a sequence of iid tasks, to optimize the expected average reward per unit time. We provide an algorithm for answering this question. We combine the global optimizer Cross Entropy method and the local gradient ascent, and theoretically investigate how far the estimated gradient is from the true gradient. We empirically demonstrate the applicability of the proposed algorithm on a toy problem, as well as on a real-world face detection task.

Monte-Carlo Simulation Balancing

David Silver and Gerald Tesauro

In this paper we introduce the first algorithms for efficiently learning a simulation policy for Monte-Carlo search. Our main idea is to optimise the balance of a simulation policy, so that an accurate spread of simulation outcomes is maintained, rather than optimising the direct strength of the simulation policy. We develop two algorithms for balancing a simulation policy by gradient descent. The first algorithm optimises the balance of complete simulations, using a policy gradient algorithm; whereas the second algorithm optimises the balance over every two steps of simulation. We compare our algorithms to reinforcement learning and supervised learning algorithms for maximising the strength of the simulation policy. We test each algorithm in the domain of 5x5 Computer Go, using a softmax policy that is parameterised by weights for a hundred simple patterns. When used in a simple Monte-Carlo search, the policies learnt by simulation balancing achieved significantly better performance, with half the mean squared error of a uniform random policy, and equal overall performance to a sophisticated Go engine.

Session 6D: Learning Kernels — Wednesday, 14:00 – 16:00, Leacock 15

Chair: Francis Bach

More Generality in Efficient Multiple Kernel Learning

Manik Varma and Bodla Rakesh Babu

Recent advances in Multiple Kernel Learning (MKL) have positioned it as an attractive tool for tackling many supervised learning tasks. The development of efficient gradient descent based optimization schemes has made it possible to tackle large scale problems. Simultaneously, MKL based algorithms have achieved very good results on challenging real world applications. Yet, despite their successes, MKL approaches are limited in that they focus on learning a linear combination of given base kernels.

In this paper, we observe that existing MKL formulations can be extended to learn general kernel combinations subject to general regularization. This can be achieved while retaining all the efficiency of existing large scale optimization algorithms. To highlight the advantages of generalized kernel learning, we tackle feature selection problems on benchmark vision and UCI databases. It is demonstrated that the proposed formulation can lead to better results not only as compared to traditional MKL but also as compared to state-of-the-art wrapper and filter methods for feature selection.

Multiple Indefinite Kernel Learning with Mixed Norm Regularization

Matthieu Kowalski, Marie Szafranski, and Liva Ralaivola

We address the problem of learning classifiers using several kernel functions. On the contrary to many contributions in the field of learning from different sources of information using kernels, we here do not assume that the kernels used are positive definite. The learning problem that we are interested in involves a misclassification loss term and a regularization term that is expressed by means of a mixed norm. The use of a mixed norm allows us to enforce some sparsity structure, a particular case of which is, for instance, the Group Lasso. We solve the convex problem by employing proximal minimization algorithms, which can be viewed as refined versions of gradient descent procedures capable of naturally dealing with nondifferentiability. A numerical simulation on a UCI dataset shows the soundness of our approach.

Learning Kernels from Indefinite Similarities

Yihua Chen, Maya Gupta, and Benjamin Recht

Similarity measures in many real applications generate indefinite similarity matrices. In this paper, we consider the problem of classification based on such indefinite similarities. These indefinite kernels cannot be used in standard kernel-based algorithms as the optimization problems become non-convex. In order to adapt kernel methods for similarity-based learning, we introduce a method that aims to simultaneously find a reproducing kernel Hilbert space based on the given similarities and train a classifier with good generalization in that space. The method is formulated as a convex optimization problem. We propose a simplified version, that can reduce overfitting and whose associated convex conic program can be solved in a very efficient way due to its special structure. We compare the proposed methods with five other methods on a collection of real data sets.

SimpleNPKL: Simple NonParametric Kernel Learning

Jinfeng Zhuang, Ivor Tsang, and Steven Hoi

Previous studies of Non-Parametric Kernel (NPK) learning usually reduce to solving some Semi-Definite Programming (SDP) problem by a standard SDP solver. However, time complexity of standard interior-point SDP solvers could be as high as $O(n^{6.5})$. Such intensive computation cost prohibits NPK learning applicable to real applications, even for data sets of moderate size. In this paper, we propose an efficient approach to NPK learning from side information, referred to as SimpleNPKL, which can efficiently learn non-parametric kernels from large sets of pairwise constraints. In particular, we show that the proposed SimpleNPKL with linear loss has a closed-form solution that can be simply computed by the Lanczos algorithm. Moreover, we show that the SimpleNPKL with square hinge loss can be re-formulated as a saddle-point optimization task, which can be further solved by a fast iterative algorithm. In contrast to the previous approaches, our empirical results show that our new technique achieves the same accuracy, but is significantly more efficient and scalable.

Geometry-Aware Metric Learning

Zhengdong Lu, Prateek Jain, and Inderjit Dhillon

In this paper, we introduce a generic framework for semi-supervised kernel learning. Given pairwise (dis-)similarity constraints, we learn a kernel matrix over the data that respects the provided side-information as well as the local geometry of the data. Our framework is based on metric learning methods, where we jointly model the metric/kernel over the data along with the underlying manifold. Furthermore, we show that for some important parameterized forms of the underlying manifold model, we can estimate the model parameters and the kernel matrix efficiently. Our resulting algorithm is able to incorporate local geometry into metric learning task, at the same time it can handle a wide class of constraints and can be applied to various applications. Finally, our algorithm is fast and scalable – unlike most of the existing methods, it is able to exploit the low dimensional manifold structure and does not require semi-definite programming. We demonstrate wide applicability and effectiveness of our framework by applying to various machine learning tasks such as semi-supervised classification, colored dimensionality reduction, manifold alignment etc. On each of the tasks our method performs competitively or better than the respective state-of-the-art method.

Chair: Shai Shalev-Schwartz

Boosting Products of Base Classifiers

Balázs Kégl and Róbert Busa-Fekete

In this paper we show how to boost products of simple base learners. Similarly to trees, we call the base learner as a subroutine but in an iterative rather than recursive fashion. The main advantage of the proposed method is its simplicity and computational efficiency. On benchmark datasets, our boosted products of decision stumps clearly outperform boosted trees, and on the MNIST dataset the algorithm achieves the second best result among no-domain-knowledge algorithms after deep belief nets. As a second contribution, we present an improved base learner for nominal features and show that boosting the product of two of these new subset indicator base learners solves the maximum margin matrix factorization problem used to formalize the collaborative filtering task. On a small benchmark dataset, we get experimental results comparable to the semi-definite-programming-based solution but at a much lower computational cost.

ABC-Boost: Adaptive Base Class Boost for Multi-class Classification

Ping Li

We propose ABC-Boost (Adaptive Base Class Boost) for multi-class classification and present ABC-MART, an implementation of ABC-Boost. The original MART (Multiple Additive Regression Trees) algorithm has been popular in certain industry applications (e.g., Web search). For binary classification, ABC-MART recovers MART. For multi-class classification, ABC-MART improves MART, as evaluated on several public data sets.

Compositional Noisy-Logical Learning

Alan Yuille and Songfeng Zheng

We describe a new method for learning the conditional probability distribution of a binary-valued variable from labelled training examples. Our proposed Compositional Noisy-Logical Learning (CNLL) approach learns a noisy-logical distribution in a compositional manner. CNLL is an alternative to the well-known AdaBoost algorithm which performs coordinate descent on an alternative error measure. We describe two CNLL algorithms and test their performance compared to AdaBoost on two types of problem: (i) noisy-logical data (such as noisy exclusive-or), and (ii) four standard datasets from the UCI repository. Our results show that we outperform AdaBoost while using significantly fewer weak classifiers, thereby giving a more transparent classifier suitable for knowledge extraction.

Boosting with Structural Sparsity

John Duchi and Yoram Singer

We derive generalizations of AdaBoost and related gradient-based coordinate descent methods that incorporate sparsity-promoting penalties for the norm of the predictor that is being learned. The end result is a family of coordinate descent algorithms that integrate forward feature induction and back-pruning through regularization and give an automatic stopping criterion for feature induction. We study penalties based on the ℓ_1, ℓ_2, and ℓ_∞ norms of the predictor and introduce mixed-norm penalties that build upon the initial penalties. The mixed-norm regularizers facilitate structural sparsity in parameter space, which is a useful property in multiclass prediction and other related tasks. We report empirical results that demonstrate the power of our approach in building accurate and structurally sparse models.

Learning with Structured Sparsity

Junzhou Huang, Tong Zhang, and Dimitris Metaxas

This paper investigates a new learning formulation called structured sparsity, which is a natural extension of the standard sparsity concept in statistical learning and compressive sensing. By allowing arbitrary structures on the feature set,this concept generalizes the group sparsity idea. A general theory is developed for learning with structured sparsity, based on the notion of coding complexity associated with the structure. Moreover, a structured greedy algorithm is proposed to efficiently solve the structured sparsity problem. Experiments demonstrate the advantage of structured sparsity over standard sparsity.

Student Scholarship Recipients

We are delighted this year to be able to support an extensive student scholarship program. Each student received Registration and Travel support. This program has been made possible by the very generous support of the National Science Foundation, with additional support from Microsoft Research, Google, Yahoo! Labs, Siemens, IBM Research, and AICML.

The following student scholarship recipients will be presenting posters based on conference papers (see program for titles and assignment to Tuesday or Wednesday poster sessions):

Amr Ahmed	David Andrzejewski	Bodla Rakesh Babu
Abdeslam Boularias	Jianhui Chen	Chih-Chieh Cheng
Weiwei Cheng	Youngmin Cho	Meghana Deodhar
Marc Deisenroth	Lixin Duan	Eduardo Gomes
Laurent Jacob	Stefanie Jegelka	Jonathan Huang
Tzu-Kuo Huang	Jens Hühn	Prateek Jain
Shuiwang Ji	Hetunandan Kamisetty	Nikos Karampatziakis
Yanyan Lan	Yu-Feng Li	André Martins
Brian McFee	Frédéric de Mesmay	Gerhard Neumann
Vincent Nguyen	Wei Pan	Jason Pazis
Jonas Peters	Arpad Rimmel	Sushmita Roy
Nino Shervashidze	Liang Sun	Ilya Sutskever
Gavin Taylor	Maksims Volkovs	Liu Yang
Xiao-Tong Yuan	Yisong Yue	De-Chuan Zhan
Jinfeng Zhuang		

The following student scholarship recipients will present posters describing their research:

Tuesday Poster Session, 6:45 to 11:00 p.m.

Todd Hester
An Empirical Comparison of Abstraction in Models of Markov Decision Processes

Shivaram Kalyanakrishnan
Integrating Value Function-Based and Policy Search Methods for Sequential Decision Making

George Kondaris
Value Function Approximation using the Fourier Basis

James MacGlashan
Hierarchical Skill Learning for High-Level Planning

Don Miner:
Learning Non-Explicit Control Parameters of Self-Organizing Systems

Wednesday Poster Session, 6:45 to 11:00 p.m.

Christopher Painter-Wakefield
Linear Value Function Approximation and Linear Models

Marc Pickett
Unsupervised Formation of Invariant Concepts from Unstructured Data

Patricia Ordonez Rozo
Multivariate Time Series Analysis of Physiological and Clinical Data

Alicia Wolfe
Finding Equivalences Among Abstract Actions

Tutorials (Sunday, June 14)

9:00 – 11:30	**T1: Reductions in Machine Learning** Alina Beygelzimer, John Langford, and Bianca Zadrozny	Leacock 26
9:00 – 11:30	**T2: Convergence of Natural Dynamics to Equilibria** Eyal Even-Dar and Vahab Mirrokni	Leacock 219
9:00 – 11:30	**T3: Learning with Dependencies between Several Response Variables** Volker Tresp and Kai Yu	Leacock 232
13:00 – 15:30	**T4: Survey of Boosting from an Optimization Perspective** Manfred K. Warmuth and S.V.N. Vishwanathan	Leacock 219
13:00 – 15:30	**T5: The Neuroscience of Reinforcement Learning** Yael Niv	Leacock 232
13:00 – 15:30	**T6: Machine Learning in IR: Recent Successes and New Opportunities** Paul Bennett, Misha Bilenko, and Kevyn Collins-Thompson	Leacock 26
16:00 – 18:30	**T7: Active Learning** Sanjoy Dasgupta and John Langford	Leacock 219
16:00 – 18:30	**T8: Large Social and Information Networks: Opportunities for ML** Jure Leskovec	Leacock 232
16:00 – 18:30	**T9: Structured Prediction for Natural Language Processing** Noah Smith	Leacock 26

Workshops (Thursday, June 18)

W1: Seventh Annual Workshop on Bayes Applications John Mark Agosta, Russell Almond, Dennis Buede, Marek J. Druzdzel, Judy Goldsmith, and Silja Renooij	Trottier 0070
W2: Automated Interpretation and Modelling of Cell Images Robert F. Murphy, Chun-Nan Hsu, and Loris Nanni	Trottier 2120
W3: Workshop on Learning Feature Hierarchies Kay Yu, Ruslan Salakhutdinov, Yann LeCun, Geoff Hinton, and Yoshua Bengio	Trottier 2110
W4: Results of the 2009 Reinforcement Learning Competition David Wingate, Carlos Diuk, Lihong Li, Matthew Taylor, and Jordan Frank	Rutherford 118
W5: The Fourth Workshop on Evaluation Methods for Machine Learning Chris Drummond, Nathalie Japkowicz, William Klement, and Sofus Macskassy	Trottier 2100
W6: On-line Learning with Limited Feedback Jean-Yves Audibert, Peter Auer, Alessandro Lazaric, Remi Munos, Daniil Ryabko, and Csaba Szepesvári	Trottier 0060
W7: Numerical Mathematics in Machine Learning Matthias Seeger, Suvrit Sra, and John P. Cunningham	Trottier 1100
W8: Abstraction in Reinforcement Learning Özgür Şimşek and George Konidaris	Rutherford 112
W9: Sparse Methods for Music Audio Douglas Eck, Dan Ellis, and Philippe Hamel	Trottier 1090

Organizing Committee

General Chair	Andrea Danyluk	Williams College
Programme Co-Chairs	Léon Bottou	NEC Laboratories America
	Michael Littman	Rutgers University
Workshop Chair	Chris Williams	University of Edinburgh
Tutorials Chair	Jennifer Neville	Purdue University
Publications Chair	Kiri Wagstaff	Jet Propulsion Laboratory, Calif. Inst. of Tech.
Volunteers Chair	Joelle Pineau	McGill University
Student Funding Co-Chairs	Drew Bagnell	Carnegie Mellon University
	Nicholas Roy	Massachusetts Institute of Technology
Fundraising Chair	Lise Getoor	University of Maryland, College Park
Local Arrangements Chair	Doina Precup	McGill University
Video Chair	Dunja Mladenic	Jozef Stefan Institute

Exhibitors

The following exhibitors will have booths in Leacock during the conference, from 8 a.m. to 6 p.m. each day.

- CRC/Taylor & Francis
- Cambridge University Press
- Now Publishers
- Springer

Sponsors

The organizers of ICML 2009 would like to thank the following sponsors:

Platinum

Microsoft Research

MITACS

NSF

Silver

GERAD Groupe d'études et de recherche en analyse des décisions

Google

YAHOO! LABS

Supporting

ALBERTA INGENUITY CENTRE FOR MACHINE LEARNING

CENTRE DE RECHERCHES MATHÉMATIQUES

PASCAL^2 Pattern Analysis, Statistical Modelling and Computational Learning

IBM

SIEMENS