IERG 5130 Probabilistic Models and Inference Algorithms for Machine Larning

Term 1, 2018-19

Instructor: Prof. Dahua Lin

Scribe: Zihao FU

2018.9.7

Distribution

- Random Variable,

$x \in \{0, 1\}$

Variable
↑
Value

两个随机变量

X $\{0, 1\}$ Y $\{0, 1, 2\}$

X \ Y	0	1	2
0	0.2	0.2	0
1	0	0.1	0.5

Joint Distribution $P(X, Y)$

若已知 $X=0$

$$P(Y \mid X=0) = \frac{P(0, Y)}{\sum_{Y=0}^{2} (0, Y)}$$

↑

conditional distribution.

如果有n个变量，每个有2个值，
一共有2^n个状态，指数增长，

stats VS Machine Learning

complexity Direct calcute efficient

两个模: Bayes network & Markov

Independence between variables.

Model 和 Distribution 的区别

Normal Distribution

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

x: example
μ, σ: parameters.

改变 μ 和 σ 得到 parametric family

Model 指的是这整个 family.

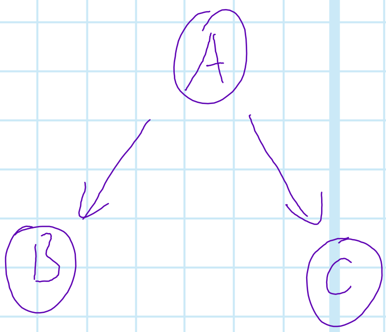
A Graphical model is use graph to represent the family

How can we use graph to represent the family ?

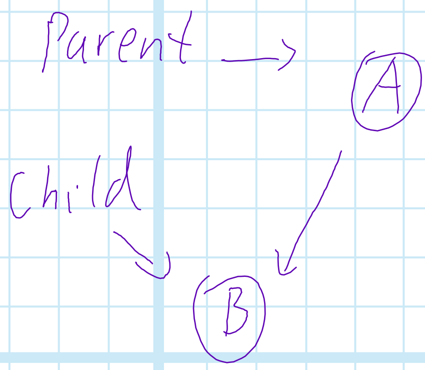
Graph

$$\underbrace{p(A, B, C) = p(A)\,p(B|A)\,p(C|A)}_{\text{factorization.}}$$

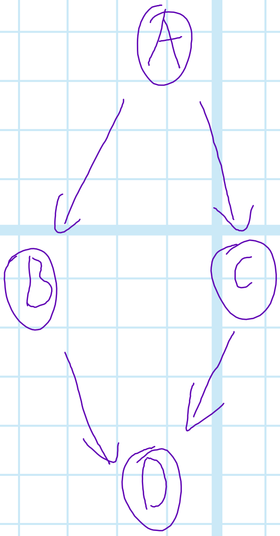
not always true here, $B \perp C \mid A$.



This is a Directed Acyclic Graph (DAG)



也叫做 Bayesian Network.



A C D B

A B C D

Topological ordering:

parents always before children

Inference alway follow Topolgical order

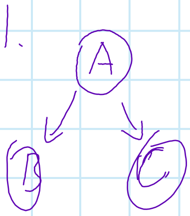
如何从 BN 来写 factorization

$$G = (V, E)$$

有几个变量就有几个乘项.

$$P(x_V) = \prod_{s \in V} p(x_s \mid x_{\pi(s)})$$

例1.

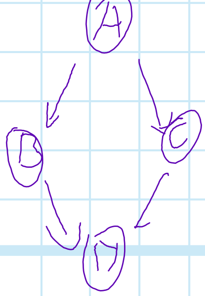


factorization formula

$$p(x_A \mid x_{\pi(A)})\, p(x_B \mid x_{\pi(B)})\, p(x_C \mid x_{\pi(C)})$$

$$= p(x_A)\, p(x_B \mid x_A)\, p(x_C \mid x_A)$$

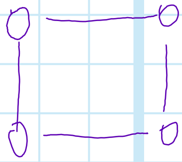
例2.



$$p(x_A \mid x_{\pi(A)})\, p(x_B \mid x_{\pi(B)})\, p(x_C \mid x_{\pi(C)})\, p(x_D \mid x_{\pi(D)})$$

$$= p(x_A)\, p(x_B \mid x_A)\, p(x_C \mid x_A)\, p(x_D \mid x_B, x_C)$$

有时候并不清楚谁是cause，这时就是无向图。

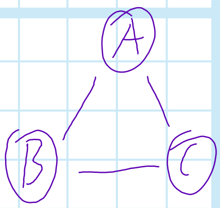


如何表示这种关系的 likelihood.

Markov Random Field (MRF)

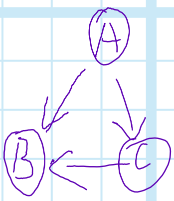
or Markov Network

用 undirected Graph 表示。



如何表示？

法1：加上依赖关系：



$p(A|B)\,p(A|C)\,p(B|C)$

但是强加了依赖关系.

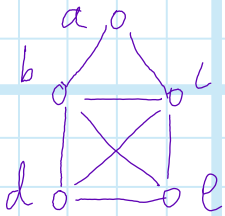
法2，

$$P(A,B,C)=\frac{1}{Z}\phi(A,B)\,\phi(B,C)\,\phi(A,C)$$

能否写成一个通式？

clique: A fully connected subset

$\{a,b\}$ $\{a,b,c\}$



不是团：$\{a,b,c,d\}$ ✗ ∵ a,d未连接.

Maximal Clique: add one more variable is not clique, it's a maximal clique.

$\{a,b\}$ 不是最大团，∵ 添加 c 了之后还是团。

$\{b,c,d,e\}$ 是最大团.

$\{b,c,d\}$ 是团，但不是最大团.

$$P(X_V) = \frac{1}{Z} \prod_{c \in C} \phi_c(X_c)$$

← compatibility function.

$$\frac{1}{Z} \phi(a,b,c)\, \phi(b,c,d,e)$$

$$\phi(a,b,c) = \phi(a,b)\, \phi(a,c)\, \phi(b,c)$$

如何定义 compatibility function,

can be arbitrary non-negative function.

例: Ⓐ — Ⓑ

A \ B	0	1
0	α	1
1	1	α

Independence: Most fundamental in probablistic theory.

can be used to simplify the computation.

$$P(A, B) = P(A)\, P(B) \qquad \text{Independence: } A \perp B$$

$$P(B|A) = P(B)$$

下面来看期望.

$$E[f(x)] = \sum_{x \in X} p(x) f(x)$$

$$A \perp B \Rightarrow E[f(A) \cdot f(B)] = E[f(A)]\, E[f(B)]$$

用处：减小计算复杂度

$$E[f(A) \cdot f(B)] : O(m^2)$$

$$E[f(A)] \cdot E[f(B)] : O(m).$$

Proof:

$$E(f(A) f(B)] = \sum_{A} \sum_{B} p(A, B)\, f(A)\, f(B)$$

$$= \sum_{A} \sum_{B} p(A)\, P(B)\, f(A)\, f(B)$$

$$= \sum_A P(A) f(A) \sum_B P(B) f(B)$$

$$= E[f(A)]\, E[f(B)]$$

推广: $E\left[\prod_{i=1}^{m} f_i(x_i)\right] = \prod_{i=1}^{m} E\left(f_i(x_i)\right)$

if x_i independent from each other.

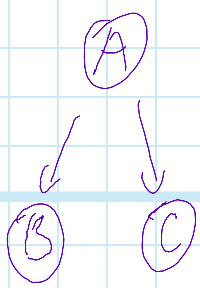
Conditional Independence

$$P(A, B \mid C) = P(A \mid C)\, P(B \mid C)$$

Conditional Independence.

It's very hard to verify.

But if we have graph, It's easy to see.



$$\Rightarrow P(B, C \mid A) = P(B \mid A)\, P(C \mid A)$$

$$\underbrace{B \perp C \mid A}_{\text{a statement of conditional independent}}$$

可见 Graphic Model 的作用:

1. give the factorization form.
2. encode conditional independent

I-map.

model $\longrightarrow$ a set of conditional independent.

$$P \longrightarrow I(P)$$

family distribution

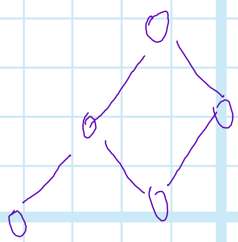
$$G \longrightarrow I(G)$$

I-map: $I(G) \subseteq I(P)$

We call G an I-map of P

G contains subset of the conditional independence of P.

G



$$P(x_v) = \frac{1}{Z} \prod_s \phi_s(x_s)$$

$$\Downarrow \qquad\qquad \Downarrow$$

$$I(G) \quad \overset{?}{\text{———}} \quad I(P)$$

what's the relation.

① Local Independence

$$X_s \perp X_{V \setminus (s \cup N_G(s))} \mid X_{N_G(s)}$$

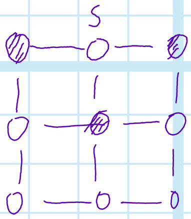
X_s is conditional independent of the rest of the world given its neighbour.

2018.9.10

$$\begin{array}{ccc} G & & P \\ \text{①}\Downarrow & & \\ I(G) & \overset{②}{\text{—}} & I(P) \\ I_g(G) & & \end{array}$$

1. Local Independence

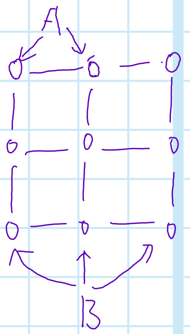
$$I_l(G) = \{ X_s \perp X_{V \setminus (\{s\} \cup N(s))} \mid X_{N(s)} \}$$

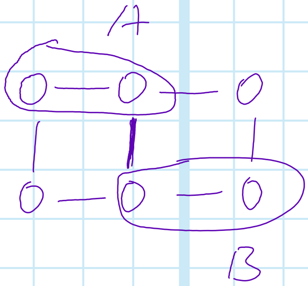


◍ : $N(s)$

2. Pairwise Independence. 是 Local Independe. 的推广

$$I_p(G) = \{ X_A \perp X_B \mid X_{V \setminus (A \cup B)} : A \text{ and } B \text{ disjoint } \& \text{ no direct edge between } A \text{ and } B \}$$





← 断不行，∵有直连的边

有 $I_l(G) \subseteq I_p(G)$

3. Global Independence.

$$I_g(G) = \{ X_A \perp X_B \mid X_C : C \text{ seperate } A \text{ and } B \}$$

Seperate: 从 A 到 B 无论怎么走都会经过 C.

$$I_l(G) \subseteq I_p(G) \subseteq I_g(G)$$

← 能从图中得到的所有 independence.

I-map:

G is an I-map of P if $I(G) \subseteq I(P)$

① Soundness

P factorize according to G

$G=(V,E)$ $\quad P(x)=\frac{1}{Z}\prod_{c\in C(G)}\phi_c(X_c)$

G is an I-map of P, i.e. $I(G)\subseteq I(P)$

证明:

把图分为三部分

Ⓐ—Ⓒ—Ⓑ $\qquad$ a,b,c: specific assigment of A,B,C

$$p(a,b,c)=\frac{1}{Z}\psi_A(a)\psi_B(b)\psi_C(c)\phi_{AC}(a,c)\phi_{BC}(b,c)$$

$$\phi'_{AC}(a,c)=\psi_A(a)\psi_C(c)\phi_{AC}(a,c)$$

$$\phi'_{BC}(b,c)=\psi_B(b)\phi_{BC}(b,c)$$

$$=\frac{1}{Z}\phi'_{AC}(a,c)\,\phi'_{BC}(b,c)$$

问题转化为证明: $A\perp B\mid C$ $\qquad p(a,b,c)=\frac{1}{Z}\phi_{AC}(a,c)\phi_{BC}(b,c)$

$$q_{A|C}(a,c)=\frac{\phi_{AC}(a,c)}{\sum_{a'}\phi_{AC}(a',c)} \qquad \psi_1(c)=\sum_{a'}\phi_{AC}(a',c)$$

$$\phi_{AC}(a,c)=q_{A|C}(a|c)\psi_1(c)$$

$$\phi_{BC}(b,c)=q_{B|C}(b|c)\psi_2(c)$$

$$\therefore p(a,b,c)=\frac{1}{Z}\phi_{AC}(a,c)\phi_{BC}(b,c)=\frac{1}{Z}\psi_1(c)\psi_2(c)q_{A|C}(a|c)q_{B|C}(b|c)$$

上面证明了如果一个概率分布能分解为一个图,即

么图中的 Independence 项都能在概率分布中满足.

下面来看逆命题.

Hammersley - Clifford

- P: positive distribution. ← density value is positive almost everywhere
- $G = (V, E)$ is an I-map of P.

  $I(G) \subseteq I(P)$

  $\Rightarrow$ P factorize according to G.

我们能否用相同的方法处理 Bayes Network? 可以.

无向: $A - B - C$

有向:

① $A \rightarrow B \rightarrow C$

$A \leftarrow B \leftarrow C$

$A \rightarrow B \leftarrow C$

$A \leftarrow B \rightarrow C$

$A \perp C \mid B$ 在哪些里面成立.

2018.9.14

下面研究 BNs 的 conditional independence

	Graph	Formulation	
①	$A \to C \to B$	$p(a)p(c\|a)p(b\|c)$	√
②	$A \leftarrow C \leftarrow B$	$p(b)p(c\|b)p(a\|c)$	√
③	$A \leftarrow C \to B$	$p(c)p(a\|c)p(b\|c)$	√
④	$A \to C \leftarrow B$	$p(a)p(b)p(c\|a,b)$	X

$A \perp B \mid C$ 是否成立?

即 $p(a,b|c)=p(a|c)p(b|c)$

$p(a)p(c|a)p(b|c) = p(a,c)\,p(b|c)$

$= p(c)\,p(a|c)p(b|c)$ ← when c is given a, and b are independent from each other

∴成立

同理成立

$p(c)p(a|c)p(b|c) = p(c)p(a,b|c)$

成立

不成立. 反例:

cause: A: work hard B: sleep well

effect: C: pass exam

A	B	P(C|A,B)
0	0	0.3
0	1	0.8
1	0	0.8
1	1	0.9

Suppose C=1 observed

$P(A,B|C=1)$

A \ B	0	1
0	$\frac{3}{20}$	$\frac{8}{20}$
1	$\frac{8}{20}$	$\frac{9}{20}$

把20 改成28

$P(A=1|C=1)=\frac{17}{20}$

$P(B=1|C=1)=\frac{17}{20}$

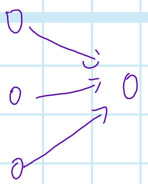
$P(A=1,B=1|C=1)=\frac{9}{20}$

$\therefore P(A=1|C=1)P(B=1|C=1) > P(A=1,B=1|C=1)$

Trick:

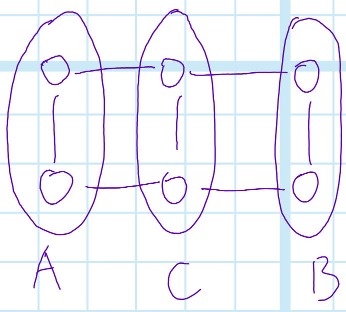
写 formulation 的时候, 一般按拓扑排序来写.

Explain Away:



一个结果有多个原因, 如果一个原因的概率高, 那么其它的原因概率则低. 则不是条件独立了.

Active Trial



$A \perp B \mid C$

A and B are blocked by C since all path from A to B is blocked/inactive by C

无向图中, observe 的变量 block 住该 path, block 住了就是 inactive, 如果所有 path inactive 则条件独立.

有向图中 block 和 active 的关系等. 四个特殊见课件 29 页.

下面看有向图如何描述 block

①②③ C block 而 ④ active when C is observed, inactive when C is not observed.

用 d-separation 来描述.

D-seperation

observed

$A \to B \to C \leftarrow D \to E$

$A \perp D \mid C$? 找到A到D的所有path. 如果path很长, 就 sub sesseion by sub session 地分析

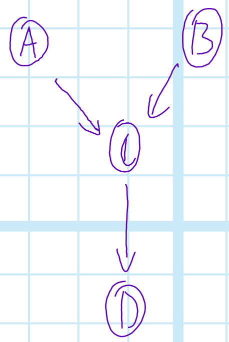
两个 subsession
① $A \to B \to C$ active
② $B \to C \leftarrow D$ active

∴ $B - B - C - D$ is active

∴ $A \perp D \mid C$ 不成立.

observe 的变量就是 blocked, 不能通过. 而 active 则根据 block 的情况由上面四种来看.

所有 BNs 都能变成 MRF

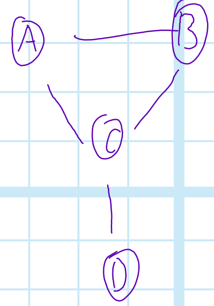


$p(a)p(b)p(c|a,b)p(d|c)$

⇓ 可以直接把概率定义为 compatibility function

$\frac{1}{Z}\phi_a(a)\phi_b(b)\phi_{c,a,b}(c,a,b)\phi_{d,c}(d,c)$

⇓ 画成 MRF



下面研究如何从有向图直接画无向图.

Moralization

给定 effect, 把所有的 parent 连起来的过程.

有向 G ⟹ 无向 M

问题: $I(G) = I(M)$?

答案：不相等.

实际上 $I(M) \subseteq I(G)$ 例如图中 $A \perp B \in I(G)$

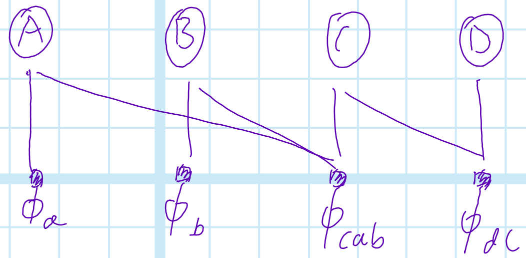
$A \perp B \notin I(M)$

转成无向图，conditional Independence 信息会丢失.

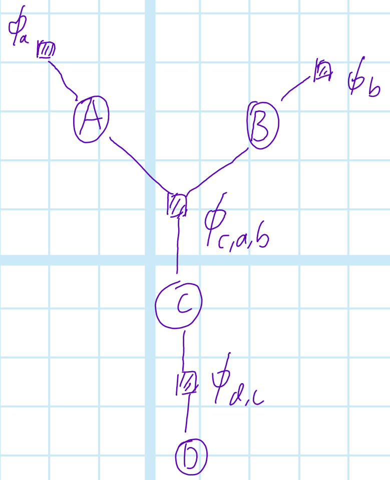
为了不丢失信息，需要研究 factor graph.

factor graph 是二部图.

Factor Graph:



也可以画成：



解决问题流程.

1. Formulate the model
2. estimate the model based on data
3. apply model

下面讲如何 formulate a graphical model.

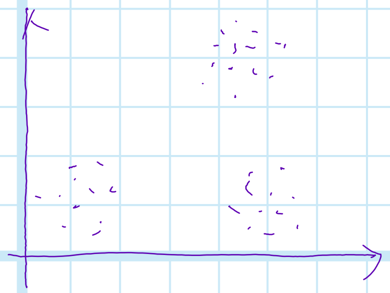
如何建模.

理解问题，用如下描述：

- Variables
- Relations
- Constraints & Assumptions.

Gaussian Mixture Model.

sample 都有 vector space representation



这种生成模型和GAN
那种生成模型是两回事

cluster
↓
components

prior
↓
components
↓
points

tradeoff: complexity
expressivity

GMM

Assumptions:

1. Each component is a Gaussian Distribution (μ, Σ)

$$p(x|\mu,\Sigma) = \frac{1}{\sqrt{(2\pi)^d|\Sigma|}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

2、k: index of the component $k=1:K$

$\{(\mu_k, \Sigma_k)\}$ ← parameters

Prior for component choice: π $\pi=(\pi_1,\cdots,\pi_K)$

Given the Model, the generate Process:

1. Choose a component:

$$z_i \in \{1,\cdots,K\} \sim \pi$$

2、$x_i \sim N(\mu_k, \Sigma_k)$ where $k=z_i$

或者写作 $x_i \sim N(\mu_{z_i}, \Sigma_{z_i})$

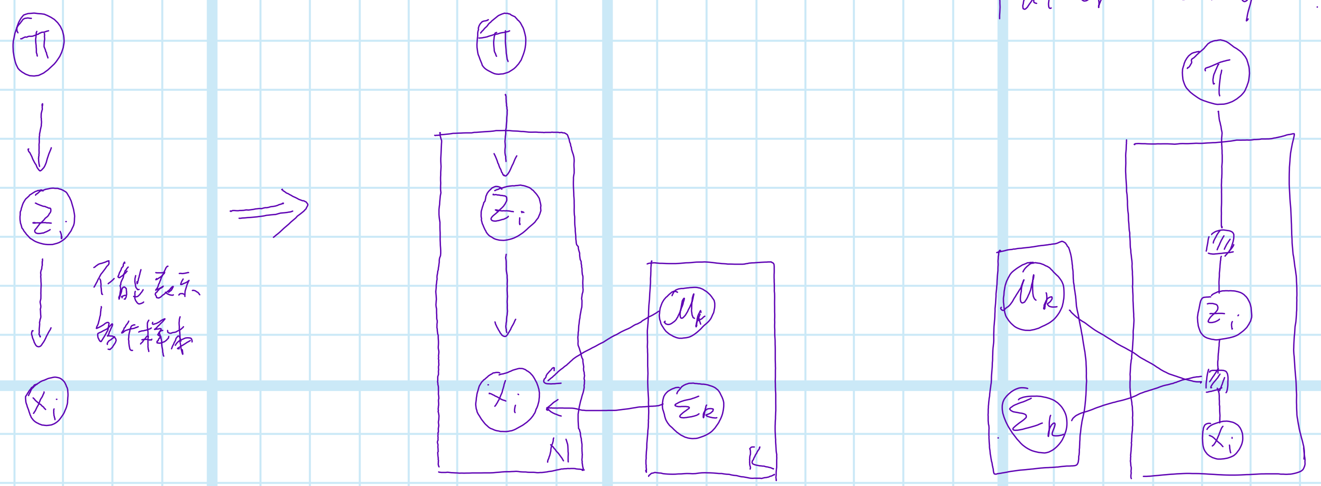
$X=(x_1,\cdots x_n)$ $Z=(z_1,\cdots,z_n)$

$$P(X,Z\mid\theta)=\prod_{i=1}^{n}p(x_i,z_i\mid\theta)=\prod_{i=1}^{n}P(z_i\mid\pi)\,p(x_i\mid\mu_{z_i},\Sigma_{z_i})$$

$(\theta, \mu_{z_i}, \Sigma_{z_i})$ 是 parameters. parameter 是从数据里学得, 学完了就固定的.

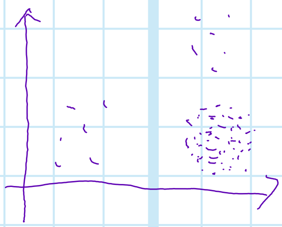
看如何画图

Factor Graph

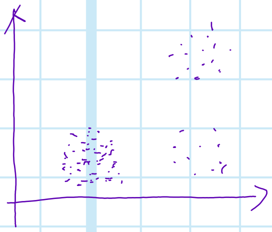


Question :

G_1

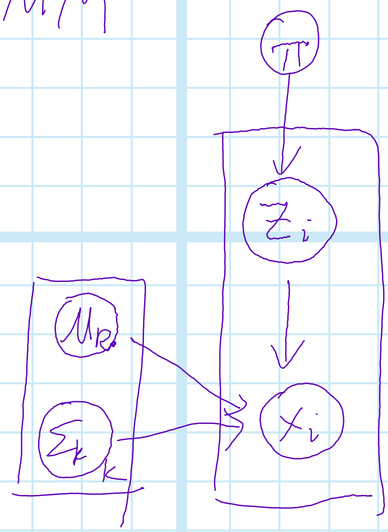


G_2



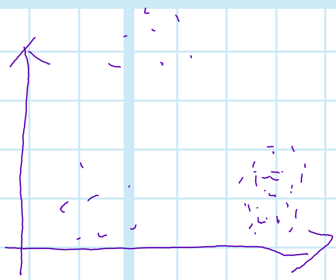
2018.9.21

GMM

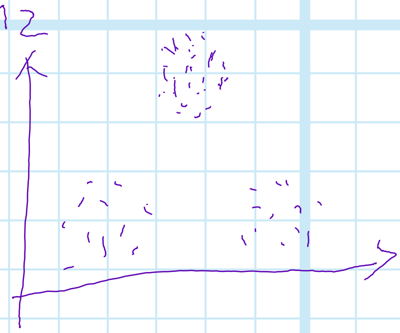


π, μ_k, Σ_k 是 parameters

G_1



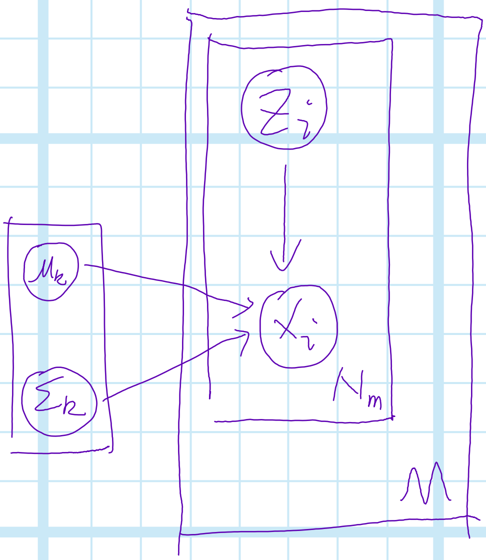
G_2



如何区别 G_1 G_2 这种不同类型.

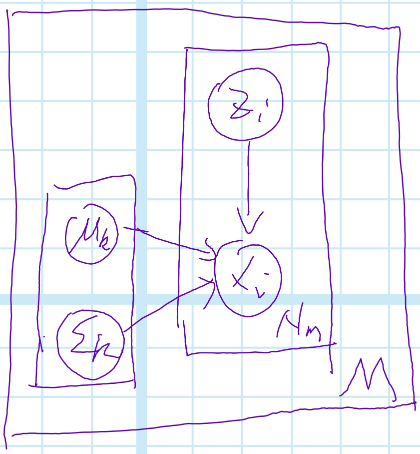
Group-wise GMM

①



m: index of group

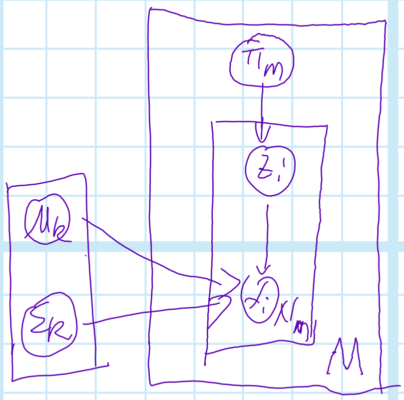
②



①: All component 共用一套参数

②: 每个 component 有自己的参数

选①还是②靠观察 dataset 来决定.



parameter:

$\{\pi_1, \pi_2, \cdots, \pi_M\}$

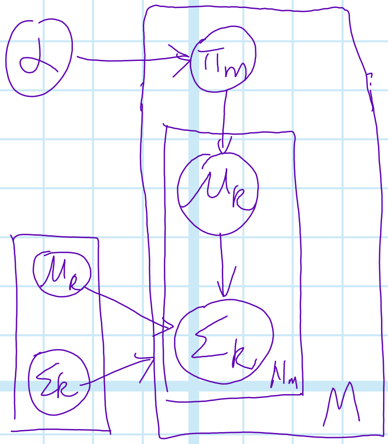
$\{(\mu_1, \Sigma_1), \cdots, (\mu_k, \Sigma_k)\}$

问题: train: $G_1, \cdots, G_M$

如果有一个新 group G' 我们没有参数 π'

解决方案: 把参数 $\{\pi_1, \pi_2, \cdots \pi_M\}$
变为 random variables.
那么问题变成了如何生成 r.v. π_m?
π_m is a probability vector
i.e. $\sum_k \pi_m(k) = 1$
$\pi_m(k) \geq 0 \quad k = 1, \cdots, k$.
Dirichlet Distribution.

Group-wise GMM v2

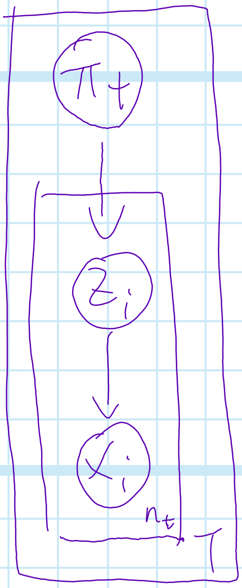


设计原则：

parameter 不要放在框里，因为这样新的 group 来了没法弄。解法是把 parameter 变成 variable，再引入先验参数

下面看 temprol 的建模方式：

Temporal Dynamic GMM



① π_t

② z_i

③ x_i

dynamics: how things change over time.

Dynamic 可以在 x_i，z_i，π_t 三个 level，下面分别讨论

① Dynamic on x_i

dynamic model：已知当前点，下一个点的出现方式。如何描述点的变化？

物理中 Motion:

location

$$\frac{dx_t}{dt} = \mu_t$$

velocity

$$dx_t = \mu_t \cdot dt$$

$$dx_t = \mu(x_t, t)\,dt$$ 速度是位置和时间的函数

→ DE: differential Equation (deterministic)

用微分方程给出部分性质，然后能确定计算。下面引入 uncertainty

把 DE 变成 stochastic Differential Equation

$$dX_t = \mu(X_t, t)\cdot dt + \sigma(X_t, t)\, \underline{dB_t}$$

Brownian Motion

$$dX_t = \sigma \cdot dB_t$$

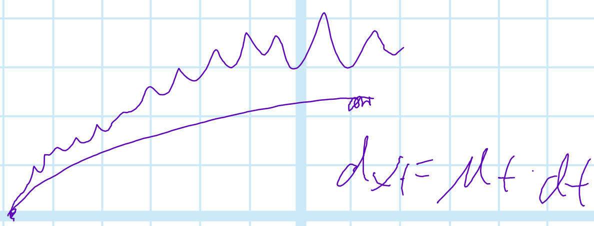
叫做 perfect Brownian Motion

$$dX_t = X_{t+dt} - X_t$$

虽然 X_t 无法明确地计算，但是有分布：

$$dX_t \sim \mathcal{N}(0, \sigma \cdot dt)$$

越往后方差越大。



那么怎么写成 conditional distribution 的形式呢

Discretize：只观察整点的 time step.

$X_t, \; X_{t+1} \; \ldots$

$P(X_{t+1} | X_t)$ 为了建模，做如下简化：

- Stationary：去掉和时间相关

$$dx_t = \mu(x_t)\,dt + \sigma \cdot dB_t$$

- Linear:

$$\mu(x_t) = Ax_t + b$$

deterministic part：决定性项　　　布朗运动 决定性项

$$\therefore dx_t = x_{t+1} - x_t \sim \mathcal{N}(\mu(x_t), \sigma^2 I)$$

$$\sim \mathcal{N}(Ax_t + b, \sigma^2 I)$$

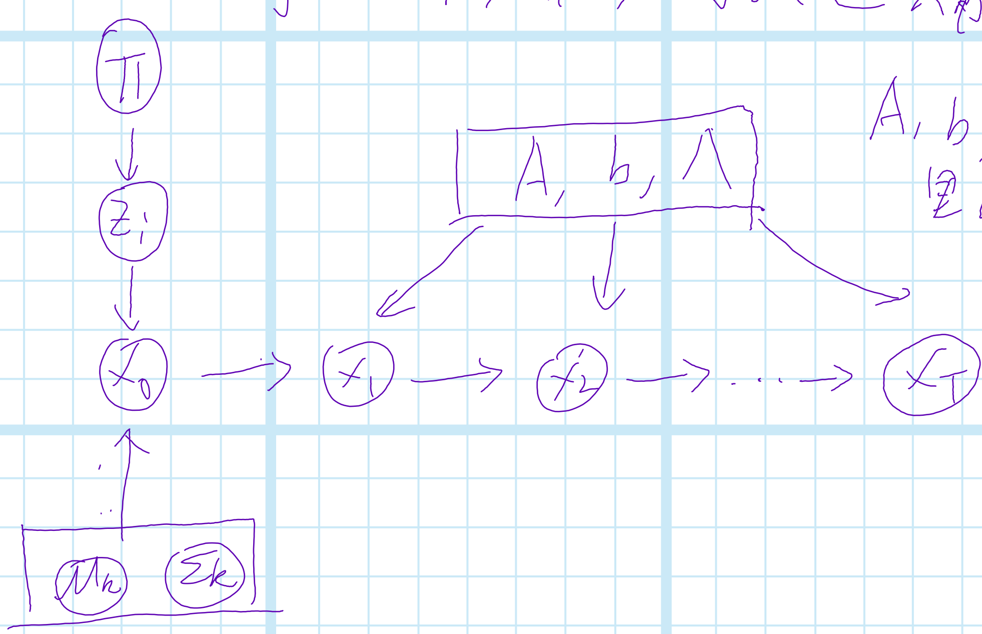
$$\therefore p(x_{t+1} \mid x_t) \sim \mathcal{N}(x_t + Ax_t + b, \sigma^2 I)$$

$$= \mathcal{N}(A' x_t + b, \sigma^2 I)$$

这样完成了布朗运动的建模(条件分布)

卡尔曼滤波也是用这个建模。

回到之前的GMM，可以这么做.



A, b, Λ 决定布朗运动里的那个参数.

详细说明其分布怎么来.

$$dx_t = \mu(x_t)\cdot dt + \sigma\cdot dB_t$$

$$x_{t+1} - x_t$$

$$= \underbrace{\int_t^{t+1}\mu(x_t)\,dt}_{\text{数学积分}} + \underbrace{\int_t^{t+1}\sigma\cdot dB_t}_{\text{伊藤积分}}$$

数学积分 $\downarrow$ 确定的积分

伊藤积分 $\downarrow$ 积出来是 r.v.

$$\int_t^{t+1} dB_t \sim \mathcal{N}(0, I)$$

$$\mu(x_{t+\Delta t}) = \mu(x_t) = \mu(x_t)\int_t^{t+1} dt$$

$$\Delta t \in (0,1) \qquad = \mu(x_t) + \sigma\cdot y$$

之前的 dynamic 是在 x_i 上，上面看 dynamic 在 z_i 上，时序的 HMM

② Dynamics on z_i

$$\begin{tikzcd}
z_0 \arrow[r] \arrow[d] & z_1 \arrow[r] \arrow[d] & z_2 \arrow[r] \arrow[d] & \cdots \arrow[r] & z_T \arrow[d] \\
x_0 & x_1 & x_2 & & x_4
\end{tikzcd}$$

这里 dynamic 不直接在 x_i 上，而是在 z_i 上。

$$P(z_{t+1}|z_t)$$

step:

$z_0 \sim \pi$

$x_0 \sim \mathcal{N}(\mu_{z_0}, \Sigma_{z_0})$

$z_{t+1}|z_t \sim$

$$\begin{array}{c|ccc} z_t \backslash z_{t+1} & 1 & \cdots & K \\ \hline 1 & & & \\ \vdots & & \text{conditional table} & \\ K & & & \end{array}$$

$$x_t \sim \mathcal{N}(\mu_{z_t}, \Sigma_{z_t})$$

∴ HMM 是 GMM dynamic 在 z_i 上的扩展

下面看能否把 dynamic 再上升一个 level

③ Dynamics on π_t.

$$\pi_0 \longrightarrow \pi_1 \longrightarrow \cdots \longrightarrow \pi_T$$

我们希望 π_{t+1} 与 π_t 接近.

可以用高斯模型来建模，但是无法保证新的 π_{t+1} 是概率分布

即 $P(\pi_{t+1}|\pi_t) \sim N(\pi_t, \sigma)$

不保证 $\sum_k \pi_{t+1}(k) = 1$

可以用 Tranformed (Warped) Model.

$$\begin{tikzcd} n_0 \arrow[r] \arrow[d] & n_1 \arrow[r] \arrow[d] & \cdots \arrow[r] & n_T \arrow[d] \\ \pi_0 & \pi_1 & & \pi_T \end{tikzcd}$$

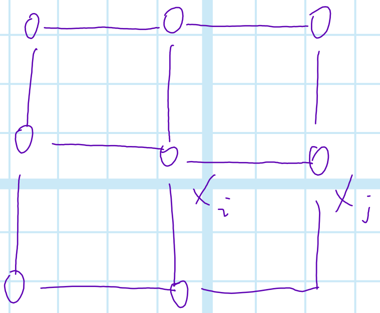
$$\pi(i) = \frac{e^{n(i)}}{\sum_{i'} e^{n(i')}}$$ softmax transform.

Softmax 可以把所有的向量变成 probability vector.

这个模型叫 Warped Gaussian Copula Process

上面讲了 Bayes Network，下面讲如何对 Markov Random Field 建模

Image as grid



$j \in N_i$ neighbour.

Smoothness Assumption.

如何用 compability function 来写 smoothness Assumption?

$$\frac{1}{Z} \prod_{(i,j) \in P_N} \phi_{ij}(x_i, x_j)$$

$$\downarrow$$

$$\frac{1}{Z} \exp\left(\sum_{(i,j) \in P_N} e_{ij}(x_i, x_j) \right)$$

Gibbs Distribution

← Neighbood 集合.

basic principle of energy function:

e_{ij} ↗ large value (undesirable)

e_{ij} ↘ small value (desirable)

$$e_{ij}(x_i, x_j) = \frac{1}{2}(x_i - x_j)^2$$

为什么要选这个? 为了 mathematical convinient.

Limitation: Smoothness Assumption

在图像中的边界处并不成立.

如何去解：采用 Gated Markov Random Field.

Gated MRF:

z_{ij} : i j 间是否有boundary.

z_{ij} 是一个Gate.

$$\frac{\alpha(z_{ij})}{2}(x_i - x_j)^2$$

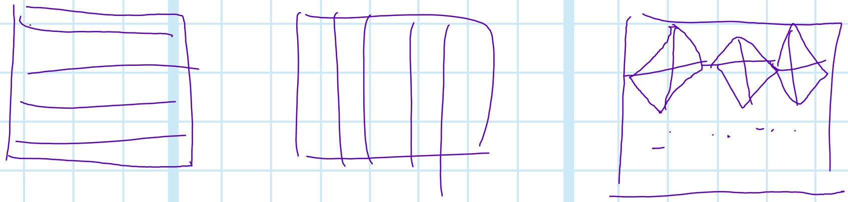
如何 develop model:

先有个 simple model, 然后观察 Assumption 是否不完全符合, 如果不是所有的都符合, 就引入 latent variable 来 indicate 每个变量的类型.

Field of Experts.

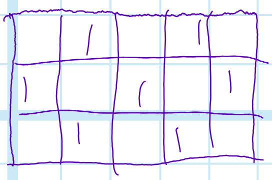
2018.9.24

Textures:



这种用 MRF smoothness 不好建模.

Field of Experts



Kernals:

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix} \quad \Bigg| \quad \begin{bmatrix} & 1 & \\ 1 & & 1 \\ & 1 & \end{bmatrix} \quad \begin{bmatrix} & & \\ 1 & 1 & 1 \\ & & \end{bmatrix} \quad \begin{bmatrix} 1 & & 1 \\ & 1 & \\ 1 & & 1 \end{bmatrix}$$

pattern	k_1	k_2	k_3
	4	2	0

Single Expert:

$$\exp(k_i^T I)$$

Product of Experts

$k_1, \dots, k_m$

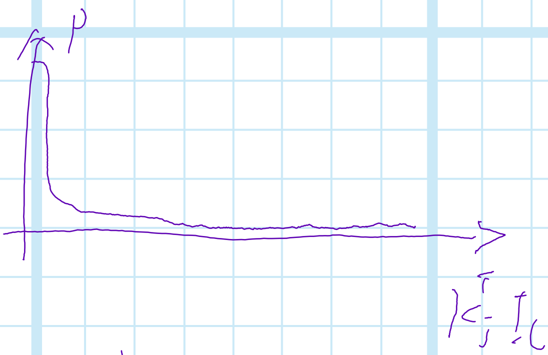
$$\exp\left(\sum_{j=1}^{n} \alpha_j k_j^T I\right)$$

但是这种只能用于 Kernel 和原图一样大的情况，如何扩展？

Field of Experts

$$\frac{1}{Z}\exp\left(\sum_c \sum_{j=1}^{m} \alpha_j \cdot K_j^T \cdot I_c\right)$$

Long Tail Distribution 问题



解决方法：把

$$f(u)=\exp(u) \Rightarrow f(u)=\frac{1}{1+\frac{1}{2}u^2}$$

$$\exp\left(\log\frac{1}{1+\frac{u^2}{2}}\right)$$

Exponential Family

$$P_\theta(x)=\frac{h(x)}{Z(\theta)}\exp\left(\eta^T(\theta)\,\phi(x)\right)$$

normalizing constant (partition function)

function of parameter

function of data.

$$= h(x)\cdot \exp\left(\eta(\theta)^T\phi(x) - \log Z(\theta)\right)$$

$$= h(x)\cdot \exp\left(\eta(\theta)^T\phi(x) - \underbrace{A(\theta)}_{\text{log partition function}}\right)$$

$$Z(\theta) = \int \exp\left(\eta(\theta)^T\phi(x)\right) h(x)\,\nu(dx)$$

$$A(\theta) = \int \exp\left(\eta(\theta)^T\phi(x)\right) h(x)\,\underbrace{\nu(dx)}_{\text{base measure}}$$

* $\eta^T(\theta)\phi(x)$ 可以的例子:

$f(\theta,x) = \theta\cdot x^2$ ✓

$f(\theta,x) = \dfrac{1}{1+\theta\cdot x}$ ✗ 不能分解成连乘的形式.

* base measure: 统一离散和连续的表达形式.

在概率中有两种分布 离散和连续

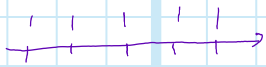
Discrete Distribution

$p_1 \dots p_n$

$f: X \to \mathbb{R}$

$E_p[f] = \sum_{i=1}^{n} p_i f_i$ 如何写成统一的积分形式.

$= \int f(x)p(x)\underbrace{\mu(dx)}_{\text{counting measure}}$ →



连续情况下:

$E_p[f] = \int f(x)p(x)\,dx$

$= \int f(x)p(x)\,\underbrace{\nu(dx)}$ 用勒贝格测度 转化为标准数学积分.

*

$$\underbrace{h(x)}_{\text{base density}}\exp\left(\underbrace{\eta(\theta)}_{\text{canonical parameter}}^T \underbrace{\phi(x)}_{\text{sufficient statistics}} - \underbrace{A(\theta)}_{\text{log partition function}}\right)$$

sufficient statistics: 只要知道期望 $E_p[\phi(x)]$ 整个分布就确定了

canonical parameter: 就是说所有的 parameter 化成这种标准形式

为什么 exp-family 重要?

∵ 很多常见分布都是 exp family

下面看如何写成 exp family

Bernoulli Distribution: (最简单的分布)

$$x \in \{0, 1\}$$

$$p(x) = \begin{cases} p_0 & (x=0) \\ p_1 & (x=1) \end{cases}$$

$$p_0 + p_1 = 1$$

$$p(x) = \begin{cases} \exp(\log(p_0)) & (x=0) \\ \exp(\log(p_1)) & (x=1) \end{cases}$$

$$= \exp\left(\mathbb{1}(x=0)\log(p_0) + \mathbb{1}(x=1)\log(p_1)\right)$$

$$= \exp\left((1-x)\log(p_0) + x \cdot \log(p_1)\right)$$

$$= \exp\left(\begin{bmatrix} 1-x \\ x \end{bmatrix} \cdot \begin{bmatrix} \log p_0 \\ \log p_1 \end{bmatrix}\right)$$

2018.9.28

Exponential Family

$$P_\theta(x) = \underbrace{h(x)}_{\text{base measure}} \exp\big(\underbrace{\eta(\theta)^T}_{\text{canonical parameter}} \underbrace{\phi(x)}_{\text{sufficient statistics}} - \underbrace{A(\theta)}_{\text{log partition function}}\big)$$

下面来看 Poisson Distribution

自然数的分布

$$P_\lambda(x) = \frac{\lambda^x}{x!} e^{-\lambda} \qquad x \in \{0, 1, \cdots\}$$

$\lambda^x = \exp(x \cdot \log \lambda)$ 代入原式：

$$P_\lambda(x) = \underbrace{\frac{1}{x!}}_{h(x)} \exp\big(\underbrace{x}_{\phi(x)} \cdot \underbrace{\log \lambda}_{\eta(\theta)} - \underbrace{\lambda}_{A(\theta)}\big)$$

Exponential Distribution

$$P_\lambda(x) = \lambda e^{-\lambda x}$$

used to capture the time is high related to Poisson Distr.

rate: in unit time, how many happens.

the rate is the λ in Poisson,

$$P_\lambda(x) = \lambda e^{-\lambda x}$$

$$= \exp(-\lambda x + \log \lambda)$$

$-\lambda : \eta(\theta)$ 式: $\lambda : \eta(\theta)$

$x : \phi(x)$ $-x : \phi(x)$

Normal Distribution

$$P_{(\mu,\sigma^2)}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\frac{(x-\mu)^2}{2\sigma^2} = \frac{x^2-2\mu x+\mu^2}{2\sigma^2}$$

$$= \frac{1}{2\sigma^2}x^2 - \frac{\mu}{\sigma^2}x + \frac{\mu^2}{2\sigma^2}$$

$$= \begin{bmatrix} \frac{1}{\sigma^2} \\ \frac{\mu}{\sigma^2} \end{bmatrix} \cdot \begin{bmatrix} \frac{x^2}{2} \\ -x \end{bmatrix} + \frac{\mu^2}{2\sigma^2}$$

$\frac{1}{\sigma^2}$: precision 记作 a

$\frac{\mu}{\sigma^2}$: potential coefficient 记作 b

$\therefore P(x) = \exp\left(-\frac{a}{2}x^2 + bx + A(a,b)\right)$

$\therefore$ 只要 exp 里面是个二次函数

那么就是正态分布!

Uniform Distribution 不属于 Exp Family 的特例.

canonical parameter: 普通参数经过 canonical transform 变为标准参数
（变换后更简洁的形式为:

$$p_\theta(x) = h(x)\exp(\theta^T x - A(\theta))$$

Ω: Domain of parameters.

$$Z(\theta) = \int_X \exp(\theta^T \phi(x))\, h(dx)$$

我们希望 $Z(\theta)$ 是 finite value.

The set of valid parameters:

$$\Omega = \left\{\theta : \int_{\underset{\text{sample space}}{X}} \exp(\theta^T \underbrace{\phi(x)}_{\text{sufficient statistics}})\, h(dx) < +\infty\right\}$$

Ω 由 sample space 和 sufficient statistics 共同决定.

Regular Family

Ω is an open subset of $\mathbb{R}^d$

open subset 是个拓扑概念

开集的好处：邻域都在 Ω，所以在任一地方都可以求导.

幸运的是，几乎所有常见分布都是 Regular Family

事实上，很多 paper 引入各种条件都是为了引入某种性质，这里是可导.

Idenitiable.

$$p(x) = \begin{cases} p & (x=1) \\ 1-p & (x=0) \end{cases} \qquad \text{(Bernulli Distr.)}$$

$$p(x) = \exp\Big((1-x)\underbrace{\log(1-p)}_{a_0} + x\underbrace{\log p}_{a_1} \Big)$$

$$= \exp\big(a_0(1-x) + a_1 x - A(a_0, a_1)\big)$$

当 $a_0=a_1=1$ 或 $a_0=a_1=2$ 时

都有 $p=0.5$

Identifiable problem: is a fundamentable problem. Two sets of parameters are undisdinguishable.

形式上:

$\theta_1 \neq \theta_2 \Rightarrow p_{\theta_1} \neq p_{\theta_2}$: Identifiable.

$\exists\, \theta_1 \neq \theta_2$ 时, $p_{\theta_1} = p_{\theta_2}$: Unidentifiable.

∴ 刚才那个 Bernulli Distr. is unidentifiable.

下面研究如何把 unidentifiable 转变为 identifiable.

$$p(x) = \exp(a_0(1-x) + a_1 x - A(a_0, a_1)) \quad \text{B1}$$

$$\Downarrow$$

$$p(x) = \exp(\theta x - A(\theta)) \quad \text{B2}$$

下面的问题是为一个 Family 是否 identifiable 取决于如何设充分统计量.

Over complete representation

$\exists\, a \neq 0 \quad a^T \phi(x) = b$

Minimal representation

otherwise

对于 B1 :

$$\phi(x) = \begin{pmatrix} 1-x \\ x \end{pmatrix} \quad a = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$$

$$a^T \phi(x) = (1-x) + x = 1$$ over complete

$(b=1)$

对于 B2

$\phi(x) = x$ $\therefore$ Minimal.

那么 over complete 和 Identifiable 的关系是?

Regular Family

$\exists\ a \neq 0$: $a^T \phi(x) = b$ a.e.

$\rightarrow$ unidentifiable.

注意 有两个条件

① Regular Family

② $a^T \phi(x) = b$.

证明!

由上式决定

$$\theta : \quad p_\theta(x) = h(x) \exp\big(\underbrace{\theta^T \phi(x)}_{\text{This component determin the distribution}} - A(\theta)\big)$$

$\theta' = \theta + \lambda a$ where $\lambda \neq 0$ $\theta' \neq \theta$.

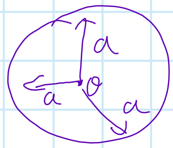
$$P_{\theta'}(x) = h(x) \cdot \exp(\theta'^T \phi(x) - A(\theta'))$$

$$= h(x) \cdot \exp(\theta^T \phi(x) + \lambda a^T \phi(x) - A(\theta'))$$

$$= h(x) \exp(\theta^T \phi(x) + \lambda b - A(\theta'))$$

$$\Rightarrow P_\theta(x) = P_{\theta'}(x)$$

Q.E.D



这里用 Regular 性质，θ 往任一方向都没出界.

$\{\theta + \lambda a, \forall \lambda \in R\}$ 都能得到相同分布.

那么 unidentifiable 能否推出 Overcomplete?

证 unidentifiable 比证 identifiable 容易，

∵ 只要举反例即可. 这个证明以后再证.

Categorical Distribution

$$X = \{1, \dots, k\}$$

$$P(x) = \begin{cases} p_1 & (x=1) \\ \vdots \\ p_k & (x=k) \end{cases} \qquad \sum_{k=1}^{K} p_k = 1$$

和 Bernulli 一样:

$$P(x) = \exp\left(\sum_{k=1}^{K} \theta_k \mathbb{I}(x=k) - A(\theta)\right)$$

这个不是 Minimal Distr.

例如

$$\theta_1=\theta_2=\cdots\theta_K=1$$

$$\theta_1=\theta_2=\cdots\theta_K=2$$

是相同分布.

那么如何写成 Minimal Representation 作业.

下面看充分统计量为什么是充分, 以及如何决定整个分布.

P_θ $\phi(x)$ 充分统计量的均值:

$$\mu=E_p[\phi(x)]=\int P_\theta(x)\phi(x)\nu(dx)$$

如果知道了 μ, 那么整个分布即确定.

We call it mean parameter.

$\mu\in\mathbb{R}^d$

$P_\theta \xrightarrow{\text{realize}} \mu$:

μ 可以由一套参数来 realize.

M_ϕ realizable mean

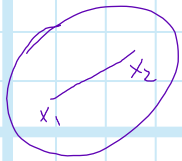
$$=\{\mu:\exists\theta\in\Omega,\ s.t.\ E_{P_\theta}[\phi(x)]=\mu\}$$

也写作: $\{\mu : \exists p \in \mathcal{P}(x) \; E_p[\phi(x)] = \mu\}$

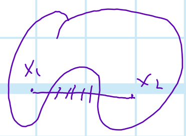
M_ϕ is convex set

Convex:

is a subset of real vector space



Convex



non-convex

S is a convex set when

$\forall x_1, x_2 \in S$

$(1-\lambda)x_1 + \lambda x_2 \in S \quad \forall \lambda \in [0,1]$

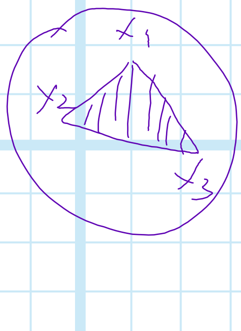
任取两点，则整条线段都在S中.

可以当作两个向量的加权平均.

$\forall x_1, \dots, x_n \in S$

$p_1, \dots, p_n$ s.t. $\sum_i p_i = 1 \quad p_i \geq 0$

$\sum_{i=1}^{n} p_i x_i \in S$



下面来证 $\mathcal{M}_\phi$ is a convex set

$\mu_1 \in \mathcal{M}_\phi$. $\mu_2 \in \mathcal{M}_\phi$

证：$(1-\lambda)\mu_1 + \lambda\mu_2 \in \mathcal{M}_\phi$.

证明：

$P_1 \longrightarrow \mu_1$

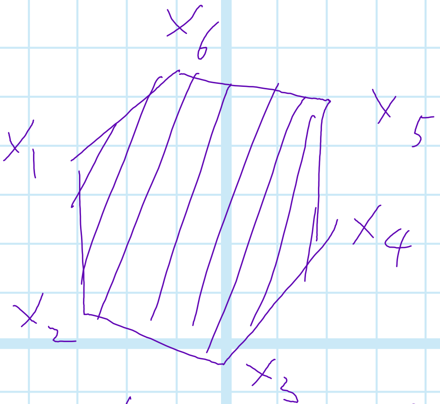
$P_2 \longrightarrow \mu_2$.

$p' = (1-\lambda)p_1 + \lambda p_2$ 仍是分布.

$p' \longrightarrow (1-\lambda)\mu_1 + \lambda\mu_2$

可以用 μ 的定义证？

$C = \{x_1, \cdots, x_6\}$



$\mathrm{conv}(C)$: convex hull

包含了所有的 convex combination of C

$$\mathrm{conv}(C) = \left\{ \sum_{i=1}^{n} p_i x_i \,\middle|\, \sum p_i = 1,\ p_i \geq 0 \right\}$$

if $\mathrm{conv}(C)$ is finite it's called a convex polytope.

$\mathcal{X}$: finite space $\{x_1, \cdots, x_k\}$

$$P_\theta(x) = \exp(\theta^T \phi(x) - A(\theta))$$

如何 charestic realise mean?

$$M_\phi = conv(\phi(x_1), \cdots, \phi(x_n))$$

可由有限个充分统计量的凸组合表示.

$$\sum p_i \phi(x_i)$$

Log Partition Funtion.

很多论文都关注如何计算 log Partition function

$$P_\theta(x) = h(x)\exp(\theta^T \phi(x) - A(\theta))$$

$$A(\theta) = \log \int_X \exp(\theta^T \phi(x)) h(dx)$$

$$\boxed{\underbrace{\nabla_\theta A(\theta)}_{\text{canonical parameter}} = \underbrace{E_{p_\theta}[\phi(x)]}_{\text{mean parameter}}}$$

下面证明这个重要的等式:

$$\nabla_\theta A(\theta) = \nabla_\theta \log \int_{\mathcal{X}} \exp(\theta^T \phi(x))\, h(dx)$$

$$= \frac{1}{\int_{\mathcal{X}} \exp(\theta^T \phi(x))\, h(dx)} \nabla_\theta \int_{\mathcal{X}} \exp(\theta^T \phi(x))\, h(dx)$$

$$h(dx) = h(x)\,dx.$$

$$= \frac{1}{Z_\theta} \int_{\mathcal{X}} \nabla_\theta \exp(\theta^T \phi(x))\, h(dx)$$

$$= \frac{1}{Z_\theta} \int_{\mathcal{X}} \exp(\theta^T \phi(x))\, \nabla_\theta [\theta^T \phi(x)]\, h(dx)$$

$$= \frac{1}{Z_\theta} \int_{\mathcal{X}} \exp(\theta^T \phi(x))\, \phi(x)\, h(dx)$$

$$= \int_{\mathcal{X}} \frac{\exp(\theta^T \phi(x))}{Z_\theta}\, \phi(x)\, h(dx)$$

$$\underbrace{\frac{\exp(\theta^T \phi(x))}{Z(\theta)}\, h(x)}_{p(x)}.$$

$$= \int_{\mathcal{X}} \phi(x)\, p(x)\, \nu(dx) = E_p[\phi(x)]$$

$\nabla_\theta A$: Gradient Map

Map the canonical parameter

to mean parameter.

下面关联 $\nabla_\theta A$, Identifable, 和 Minimal Rep.

Oct 19. In-class discuss & Presentation.

Exponential Family

$p(x) = h(x)\exp(\theta^T\phi(x) - A(\theta))$

$A(\theta) = \int_X \exp(\theta^T\phi(x))\, h(dx)$. $\nabla_\theta A$

Canonical parameter θ. Mean parameter $\mu = E_\theta[\phi(x)]$

M_ϕ . realizable mean (convex) Ω dataset

Gradient Map.

$$\boxed{\nabla_\theta A(\theta) = E_\theta[\phi(x)]}$$

injective? one-to-one.

$x_1 \neq x_2 \Rightarrow f(x_1) \neq f(x_2)$.

surjective?

$\nabla_\theta A(\Omega) \overset{?}{=} M_\phi$?

① $\nabla_\theta A$ is injective iff minimal representation

Proof:

— minimal $\Rightarrow \nabla_\theta A$ injective $\Leftarrow$ Q

② $$\boxed{H = \nabla^2 f \geq 0 \quad \text{semi positive definite}}$$

symmetric matrix A. eigenvalue & eigenvector
λ e

$Ae = \lambda e$

$\lambda_1 \geq \cdots \geq \lambda_n \geq 0 \Rightarrow A$ is ~~semi~~ positive semidefinite

$\downarrow$

$> 0 \Rightarrow A$ is positive definite.

Hessian Matrix

$f(x_1, \ldots, x_n)$

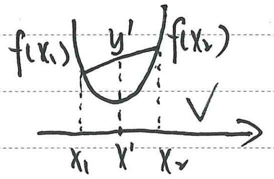
$\nabla f = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)$ Gradient.

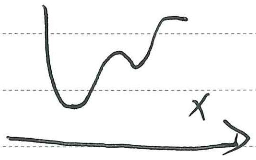
$$\nabla^2 f \begin{pmatrix} \frac{\partial^2 f}{\partial x_1 \partial x_1} & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots \\ & \ddots & \\ & & \frac{\partial^2 f}{\partial x_i \partial x_j} \end{pmatrix} \Leftarrow \text{Hessian}$$

$$\boxed{\nabla_\theta^2 A(\theta) = \underbrace{Cov_\theta}_{\text{covariance}}(\phi(x))}$$

$$Cov(x) = \begin{bmatrix} C_{ij} \\ = E[(x_i - Ex_i)(x_j - Ex_j)] \end{bmatrix}$$

$f: \Omega \to \mathbb{R}$. convex function.





$x' = \alpha x_1 + (1-\alpha)x_2$

$y' = \alpha f(x_1) + (1-\alpha) f(x_2)$

$y' \geq f(x')$.

$$p(x) = \begin{cases} \alpha & (x = x_1) \\ 1-\alpha & (x = x_2) \end{cases}$$

$E_p(x) = \alpha x_1 + (1-\alpha) x_2$

$f(E_p(x)) \leq E_p[f(x)] \Rightarrow$ ① $\boxed{f(E(x)) \leq E[f(x)] \text{ for all convex function}}$

熵

A is semi-definite

$x^TAx \geq 0 \quad \forall x$

$x = \sum_{i=1}^{n} c_i e_i$

$x^TAx = x^T\left(\sum_{i=1}^{n} c_i A e_i\right)$

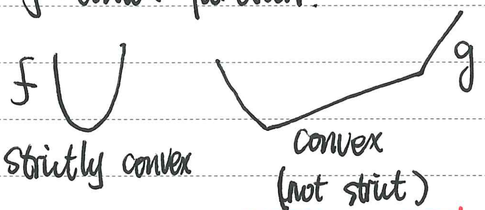
$= \left(\sum_{i=1}^{n} c_i e_i\right) \cdot \left(\sum_{i=1}^{n} c_i \lambda_i e_i\right)$

$= \underbrace{\sum_{i=1}^{n} \lambda_i c_i^2}_{\geq 0} + \underbrace{\sum_{i \neq j} c_i c_j \lambda_j e_i^T e_j}_{=0}$

$e_i^T e_j = 0$

A is definite, $x^TAx > 0 \quad \forall x \neq 0$.

f: convex function



$H = \nabla^2 f \to$ $\boxed{\text{positive definite}}$ $\Leftrightarrow f$ is strictly convex $\Leftrightarrow \nabla f$ injective.

↳ semi-definite $\Leftrightarrow f$ is convex

Answer.

∇f is injective.

∇g is not injective.

$\nabla^2 A(\theta) = Cov_\theta(\phi(x))$.

if overcomplete $\to$ A is $\boxed{\text{not}}$ strictly convex

$\to \nabla A$ is $\boxed{\text{not}}$ injective

X ~~Cov~~ $Cov(X) = C$.

$Y = a^TX$

↳ variance $Var(a^TX) = a^TCa = 0$

overcomplete: $a^TX = b$

$\downarrow$

$Var(a^TX) = 0$.

if minimal $\to a^TCa > 0 \to A$ is strictly convex

$\to \nabla A$ is injective.

Another view

∇A: Canonical parameter $\to$ mean parameter

overcomplete $\to \exists\, \theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} = P_{\theta_2} \Rightarrow \mu_1 = \mu_2 \to$ not injective

minimal $\to \forall \theta_1 \neq \theta_2 \Rightarrow P_{\theta_1} \neq P_{\theta_2} \Rightarrow \mu_1 \neq \mu_2 \Rightarrow$ injective.

Surjective. $f: \Omega \to \mathcal{M}$ $\quad \mathcal{M}_\phi(P_\theta) = \mathcal{M}_\phi$?

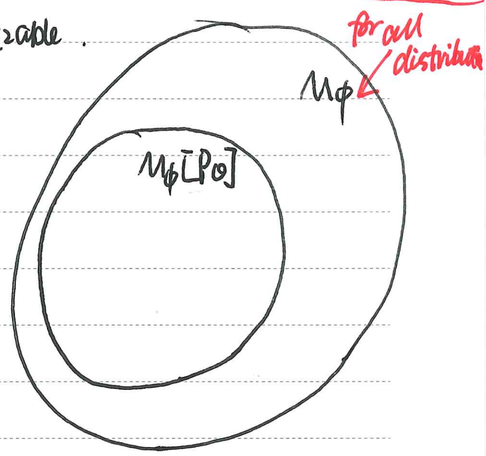
True for all regular exponential family

X. $\phi(x)$ arbitrary distribution P

$\mu = E_p(x)$

$\uparrow$

realizable.



$\mathcal{X}$: P

entropy $\to H(p) = -\int_{\mathcal{X}} p(x) \log p(x) \mu(dx)$.

$P_1, P_2, \cdots, P_n \quad \mu$

$E_{P_1}[\phi(x)] = \cdots E_{P_n}[\phi(x)] = \mu$.

Maximum Entropy $\leftarrow$ is the best choice.

Usually, more entropy means less information we have

Maximize $H(p)$ s.t. $E_p[\phi(x)] = \mu$.

$\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$ finite

$p = (p_1, p_2, \cdots p_k)$.

~~maximize~~ minimize $\sum_{i=1}^{k} p_i \log p_i$ s.t. $\sum_{i=1}^{k} p_i \phi(x_i) = \mu$ $\quad \lambda_i$

$\sum_{i=1}^{k} p_i = 1$, $p_i \geq 0$. $\quad \nu$

$L(p, \lambda, \nu) = \sum_{i=1}^{k} p_i \log p_i - \lambda_i \sum_{i=1}^{k} p_i \phi_i - \nu \sum_{i=1}^{k} p_i$

EASY MATE

$$\frac{\partial L}{\partial p_i} = (\log p_i + 1) - \lambda_i \phi_i - v = 0$$

$$\log p_i = \lambda_i \phi_i + v - 1$$

$$p_i = \exp(\lambda_i \phi_i + v - 1)$$

$$p = \frac{\exp(\lambda^T \phi)}{\exp(1-v)} = \frac{1}{Z}\exp(\lambda^T \phi).$$

$$p(x) = \frac{1}{Z}\exp(\lambda^T \phi(x)).$$

$$\begin{tikzcd} \underset{\in \Omega}{\theta} \arrow[r, bend left, "\nabla A \text{ inference}"] & \underset{\in M_\phi}{\mu} \arrow[l, bend left, "\nabla_\mu A^* \text{ learning}"] \end{tikzcd}$$

Convex Conjugate

$f: \Omega \to \mathbb{R}$. $\Omega \in \mathbb{R}^d$

$$\boxed{f^*(y) = \sup_{x\in\Omega}(y^T x - f(x)) \Rightarrow f^*(y) \geq y^T x - f(x)}$$

$$y \to \underset{x\in\Omega}{\text{maximize}}\ y^T x - f(x) \to \hat{x}$$

$$y^T \hat{x} - f(\hat{x}) = \sup_{x\in\Omega}(y^T x - f(x)).$$

$\boxed{f^* \text{ is alway convex}}$

Fenchel's inequality

$$f(x) + f^*(y) \geq y^T x.$$

Fenchel – Moreau Therom

$f^{**} = f$ iff f is convex and continuous

$$\begin{cases} f^*(y) = \sup_x (y^T x - f(x)) \\ f(x) = \sup_y (y^T x - f^*(y)). \end{cases}$$

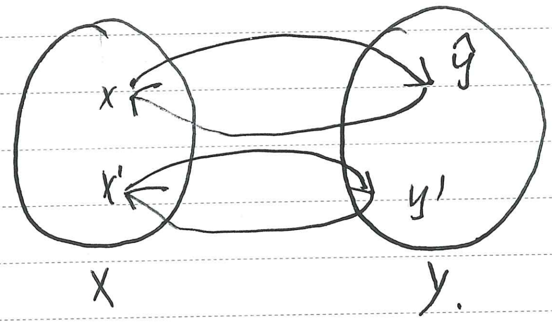
$$f^*(y) = y^T \hat{x} - f(\hat{x}).$$

$$f(x) = \hat{y}^T x - f^*(\hat{y}).$$

$$y \to \hat{x} \to \hat{y}$$

$\hat{x}, \hat{y}$ dually coupled

$$f(\hat{x}) + f^*(\hat{y}) = \hat{x}^T \hat{y}$$



$$A^*(\mu) = \sup_\theta \{\theta^T \mu - A(\theta)\}.$$

maximize $L(\theta) = \theta^T \mu - A(\theta)$

$$\nabla_\theta L(\theta) = \mu - \nabla A(\theta) = 0.$$

$$\mu = \nabla_\theta A(\theta) = E_\theta[\phi(x)].$$

$$A(\theta) = \sup_\mu \{\theta^T \mu - A^*(\mu)\}.$$

$$\hat{\theta} = \nabla_\mu A^*$$

negative entropy

$$A^*(\mu) = \begin{cases} -H(p_\theta) & \mu \in M_\phi \\ +\infty & \text{otherwise}. \end{cases}$$

$$\begin{tikzcd} \boxed{\theta} \arrow[r] \arrow[d, shift left, "\text{dually coupled}"'] & \boxed{A(\theta)} \arrow[d, "\text{convex conjugate}"] & \text{log-partition} \\ \boxed{\mu} \arrow[u, shift left] \arrow[r, "\{\theta^T\mu - A(\theta)\}"] & \boxed{A^*(\mu)} \arrow[u] & \text{negative entropy} \end{tikzcd}$$

$\hat{\mu} = \arg\max_\mu \{\theta^T\mu - A^*(\mu)\}$

$\hat{\theta} = \arg\max_\theta$

[illegible]

2018.10.8

Conjugate Priors.

$p(x|\theta)$ observe, parameter　　在Bayesian中 估计 $p(\theta)$ prior

$p(\theta|x_1,\cdots x_m)=\frac{1}{Z}p(\theta)\prod_{i=1}^{n}p(x_i|\theta)$ 计算Z很复杂，

选择合适的 $p(\theta)$，简化Z的计算。

参数的贝叶斯估计，需要选择合适的prior.

Bernolli distribution.

$$p(x|\theta)=\begin{cases}\theta & (x=1)\\ 1-\theta & (x=0)\end{cases}=\theta^x(1-\theta)^{1-x}$$

来找一个先验 $p(\theta)$

如果已经有 样本 $X=(x_1,\dots x_n)$

$$p(\theta|X)=\frac{1}{Z}\prod_{i=1}^{n}p(x_i|\theta)p(\theta)$$

$$Z=\int_{\theta\in\Omega}\prod_{i=1}^{n}p(x_i|\theta)p(\theta)d\theta$$

如果先验 $p(\theta|\alpha,\beta)=\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}\theta^{\beta-1}$ Beta分布.

$$p(\theta|X)=\frac{1}{Z}p(\theta|\alpha,\beta)\prod_{i=1}^{n}p(x_i|\theta)$$

$$=\frac{1}{Z}\cdot\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1}\prod_{i=1}^{n}\theta^{x_i}(1-\theta)^{1-x_i}$$

$$=\frac{1}{Z}\frac{1}{B(\alpha,\beta)}\theta^{\alpha-1+\sum_{i=1}^{n}x_i}(1-\theta)^{\beta-1+\sum_{i=1}^{n}(1-x_i)}$$

$$=\underbrace{\frac{1}{Z}\frac{1}{B(\alpha,\beta)}\theta^{\alpha'-1}(1-\theta)^{\beta'-1}}_{\text{posterior distribution}}$$

$$\underbrace{\alpha'=\alpha+\sum_{i=1}^{n}x_i\quad\beta'=\beta+\sum_{i=1}^{n}(1-x_i)}_{\text{updating formula.}}$$

$$B(\alpha,\beta)=\int_\Omega\theta^{\alpha-1}(1-\theta)^{\beta-1}d\theta.$$

把Z合并 $B(\alpha',\beta')$

$$Z\cdot B(\alpha,\beta)=\int_\Omega\theta^{\alpha'-1}(1-\theta)^{\beta'-1}d\theta$$

$$=B(\alpha',\beta')$$

用了conjugate后，posterior 和likelihood 形式相同，只是改变了参数.

下面 总结 conjugate prior.

和 exp family 紧密相关，如果 likehood 不是 exp family，则 没法使用conjugate.

Conjugate prior.

$p(\theta|\alpha)$

$p(x|\theta)$ likehood model.

$D=\{x_1,\cdots,x_n\}$

$$p(D|\theta)=\prod_{i=1}^{n}p(x_i|\theta)$$

$p(\theta|D)=p(\theta|\alpha')$ posteri.

$\alpha'=\alpha\oplus D$ 用given samples D更新 α.

$=\alpha\oplus\{x_1,\cdots,x_n\}$.

如何选择 conjugate prior.

$$f(x|\theta)=h(x)\exp(\eta(\theta)^T\phi(x)-r\cdot a(\theta))$$

$$p(\theta|\alpha,\beta)=\exp(\alpha^T\eta(\theta)-\beta\cdot a(\theta)-A(\alpha,\beta))$$

$p(\theta|D)\propto$ ~~$\exp(\eta(\theta)^T\phi(x)-r\cdot a(\theta))$~~

prior $\rightarrow\exp(\alpha^T\eta(\theta)-\beta\cdot a(\theta)-A(\alpha,\beta))$

likelihood $\cdot\prod_{i=1}^{n}\exp(\eta(\theta)^T\phi(x_i)-r\cdot a(\theta))$

$$\propto\exp\Big(\big(\underbrace{\alpha+\sum_{i=1}^{n}\phi(x_i)}_{\alpha'}\big)^T\eta(\theta)-(\underbrace{\beta+n r}_{\beta'})a(\theta)-A(\alpha',\beta')\Big)$$

构造方式：直接猜 $\phi(x)$ 的充分统计量.

下面看如何从 likelihood 直接来构造

$$f(x|\theta)=\theta^{x}(1-\theta)^{1-x}$$

$$=\exp(\underset{\downarrow}{x}\cdot\log\theta+\underset{\downarrow}{(1-x)}\log(1-\theta))$$

$$p(\theta|\alpha,\beta)\propto\exp(\alpha\cdot\log\theta+\beta\log(1-\theta))$$

$$=\frac{1}{Z(\alpha,\beta)}\cdot\exp(\alpha\cdot\log(\theta)+\beta\cdot\log(1-\theta))$$

$$=\frac{1}{Z(\alpha,\beta)}\cdot\theta^{\alpha}\cdot(1-\theta)^{\beta}$$

$$\therefore P_{beta}(\theta|\alpha,\beta)\propto\theta^{\alpha-1}(1-\theta)^{\beta-1}$$

Dirichlet Distribution

Beta Distribution 的扩展.

$$\begin{array}{ccc} \text{Bernoulli} & \xrightarrow{\text{两类变多类}} & \text{Categorical} \\ \uparrow \text{prior} & & \uparrow \text{prior} \\ \text{Beta} & \longrightarrow & \text{Dirichlete} \end{array}$$

$$p(x|\pi)=\pi_k.$$

$$=\exp\Big(\sum_{k=1}^{K}\underbrace{\mathbb{1}(x=k)}\log\pi_k\Big)$$

就是充分统计量.

$$\text{prior: } p(\pi|\alpha)=\frac{1}{Z(\alpha)}\exp\Big(\sum_{k=1}^{K}\alpha_k\log\pi_k\Big)$$

$$=\frac{1}{Z(\alpha)}\cdot\prod_{k=1}^{K}\pi_k^{\alpha_k}.$$

规范化：

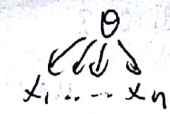
$$P_{prior}(\pi|\alpha)=\frac{1}{B(\alpha)}\prod_{k=1}^{K}\pi_k^{\alpha_k-1}$$

下面看 Dir 的性质.

$$P_{prior}(x|\alpha)=\frac{1}{B(\alpha)}=\prod_{k=1}^{K}x_k^{\alpha_k-1}$$

所有的 x 都是一个概率分布 ∴ Dir 是分布的分布.

$$E[x_k]=\frac{\alpha_k}{\sum_i\alpha_i}$$



$$p(x|D)=\int_{\Omega}p(x|\theta)\underbrace{p(\theta|D)}d\theta \longrightarrow p(\theta|\alpha',\beta')$$

$$\doteq\int_{\Omega}\exp(\eta(\theta)^{T}\phi(x)-a\cdot r(\theta))$$

$$\exp(\alpha'^{T}\eta(\theta)-\beta' r(\theta)-A(\alpha',\beta'))d\theta.$$

$$=\int_{\Omega}\exp([\alpha'-\phi(x)]^{T}\eta(\theta)-(\beta'+a)r(\theta)-A(\alpha',\beta'))d\theta$$

$$=\frac{1}{\exp(A(\alpha',\beta'))}\underbrace{\int_{\Omega}\exp[(\alpha'-\phi(x))^{T}\eta(\theta)-(\beta'+a)r(\theta)]d\theta}_{=\exp(A(\alpha'-\phi(x),\beta'+a))}$$

下一节讲 inference, given model, 做推断.

2018.10.12

Topic 2 Inference

$$P(x;\theta) \Leftarrow \text{Model}$$

$$\downarrow$$

$$(x_1, \cdots, x_n)$$

$$P(\underbrace{X_u}_{\text{query}} \mid \underbrace{X_B}_{\text{observation evidences}} ; \theta)$$

θ ← model, learnt in the past

已知模型参数 θ 和部分变量 x_B，求另一部分的条件概率

假设变量分为三部分: X Y Z 已知 $p(x,y,z|\theta)$

$$P(Y|X;\theta) = \frac{P(Y,X|\theta)}{P(X|\theta)}$$

本来是 conditional distr.
转化成了 Marginal distr.

要求两部分:

$$P(Y,X|\theta) = \sum_{z\in Z} p(y,x,z|\theta)$$

$$P(X|\theta) = \sum_{y\in Y} p(x,y|\theta)$$

计算会指数增长.

Evidence Absorption (一个 trick)

$$p(x,y,z) \propto \psi_x(x)\,\psi_y(y)\,\psi_z(z)\,\phi_{xy}(x,y)\,\phi_{yz}(y,z)\,\phi_{xz}(x,z)$$

$$P(X,Y|\underbrace{z}_{\text{observation}}) = \frac{p(x,y,z)}{\sum_{x,y} p(x,y,z)} = \frac{\psi_x(x)\psi_y(y)\psi_z(z)\phi_{xy}(x,y)\phi_{xz}(x,z)\phi_{yz}(y,z)}{\sum_{x'}\sum_{y'}\psi_x(x')\psi_y(y')\psi_z(z)\phi_{xy}(x',y')\phi_{xz}(x',z)\phi_{yz}(y',z)}$$

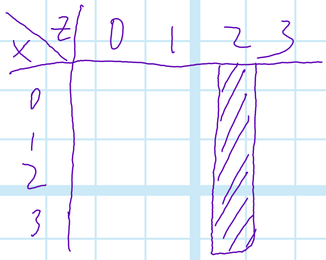
$$= \frac{\psi_x(x)\psi_y(y)\phi_{xy}(x,y)\phi_{x|z}(x)\phi_{y|z}(y)}{\sum_{x'}\sum_{y'}\cdots\cdots}$$

$$\propto \psi_x(x)\psi_y(y)\phi_{xy}(x,y)\phi_{x|z}(x)\phi_{y|z}(y)$$

∵ z 是给定 constant，分子分母的 $\psi_z(z)$ 可以约出来消掉。

$\phi_{xz}(x,z)$ 实际是个二维表：

∵ x 固定，∴ 只看其中一行



∴ $\phi_{xz}(x,z) \longrightarrow \phi_{x|z}(x)$

∴ 开始有三个变量 X, Y, Z，经过 Evidence Absorption 之后，变成了只有 X, Y 的 MRF，形式更简单。

下面看计算 Marginal probability

$P(X,Y)$ 求 $P(X)$

$$p(x) = \sum_{y} p(x,y)$$

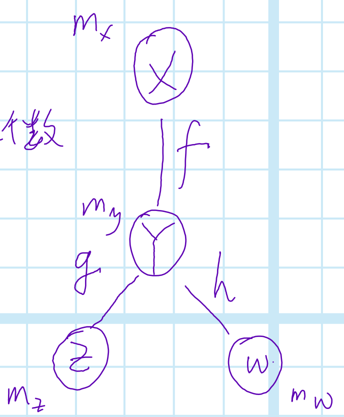
Complexity: key difference between statistics and machine learning

核心问题是如何计算 Marginal distr. efficiently

Y可能有很多个变量 $O(|Y|) \to K^n$

conditional independent is the key to reduce the complexity of the summation

m_i: 变量个数



$$P(x,y,z,w) = \frac{1}{Z} f(x,y)\, g(y,z) \cdot h(y,w)$$

$P(x)$?

marginalization: remove other varaibles only keep marginal distr.

$$P(x) = \sum_y \sum_z \sum_w \frac{1}{Z} f(x,y) g(y,z) h(y,w)$$

$$= \frac{1}{Z} \underbrace{\sum_y \sum_z \sum_w f(x,y)\, g(y,z)\, h(y,w)}_{\tilde{p}(x)}$$

一般只关注 $\tilde{p}(x)$ 然后 $p(x) = \frac{1}{Z}\tilde{p}(x)$; $Z = \sum_x \tilde{p}(x)$

$\tilde{p}(x)$ 的复杂度: $O(m_y \cdot m_z \cdot m_w)$ for a single x

∴ Overall complexity: $O(m_x, m_y \cdot m_z, m_w)$ $O(m^4)$

利用 structure 简化计算.

$\sum_x c \cdot f(x) = c \sum_x f(x)$ 利用这种性质来简化.

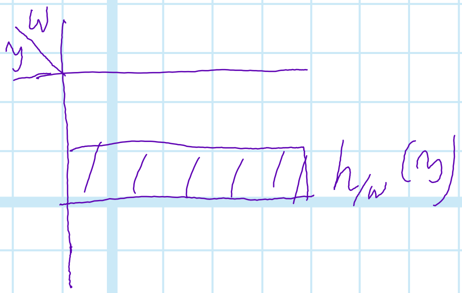
$$\tilde{p}(x) = \sum_y \sum_z \sum_w f(x,y) g(y,z) h(y,w)$$

$$= \sum_y \sum_z f(x,y) g(y,z) \sum_w h(y,w)$$

$$= \sum_y f(x,y) \sum_z g(y,z) \sum_w h(y,w)$$

下面看如何简化计算的.

令 $h_{/w}(y) = \sum_w h(y,w)$



$$= \underbrace{\sum_y f(x,y) \overset{①}{g_{/z}(y)} \overset{②}{h_{/w}(y)}}_{③}$$

复杂度: ①: $O(m_y \cdot m_z)$

+ ②: $O(m_y, m_z)$ ⇒ $O(m^2)$

+ ③: $O(m_x m_y)$

对于不同的 x, $h_{/w}(y)$ 都是相同的, ∴ 可以一次计算

以上的方法叫 Variable elimination

$g_{/z}(y)$ $\quad$ $h_{/w}(y)$

$\uparrow$ $\quad$ $\uparrow$

eliminate z $\quad$ eliminate w

条件独立的作用在于 例如 X 和 (z, w) 独立，对于任意 x，z, w 有影响其概率，可以一次计算。

Variable Eliminate

算法框架：$Y_0, Y_1, \cdots, Y_n$

$$f = \{\phi_1, \cdots, \phi_m\}$$

$$\mathcal{V} = \{Y_0, Y_1, \cdots, Y_n\}$$ 未消去的变量集合。

$j = 1, \cdots, n$

$i = \pi(j)$ 决定 Elimination 的顺序。

顺序对 complexity 影响很大。

$$f, \mathcal{V} = \text{VarEliminate}(f, \mathcal{V}, Y_i)$$

$f(Y_i)$: the set of factors involving Y_i

上例中 $f(w)=\{h\}$ $f(z)=\{g\}$

$\nu(\phi)$: active variables involved in ϕ

上例中 $\nu(h)=\{y,w\}$ $\nu(g)=\{y,z\}$

Neighbour of Y_i:

$$N_i=\{Y\neq Y_i : \exists\,\phi\in f(Y_i),\ Y\in\nu(\phi)\}$$

上例中: $f(y)=\{f,g,h\}$

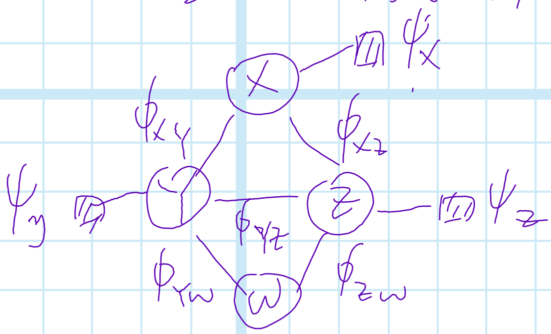
$N(y)=\{x,z,w\}$

$f(w)=\{h\}$ $N(w)=\{y\}$

Construct ψ_i on N_i

$$\psi_i(z)=\sum_{y}\prod_{\phi\in f(Y_i)}\phi(y,z\,|\,\nu(\phi))$$

下面看一个例子, 应用该算法.



$F = \{\psi_x, \psi_y, \psi_z, \psi_w, \phi_{xy}, \phi_{xz}, \phi_{yz}, \phi_{yw}, \phi_{zw}\}$

$V = \{x, y, z, w\}$

$\pi = \{w, z, y\}$

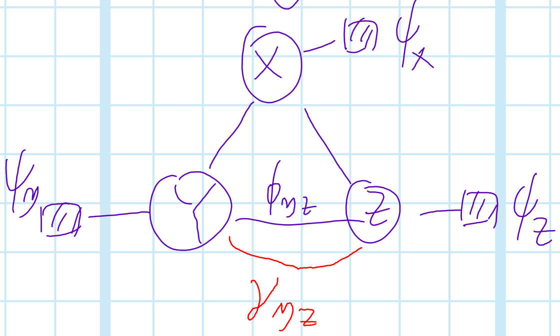
① Eliminate w

$F(w) = \{\psi_w, \phi_{yw}, \phi_{zw}\}$

$N_w = \{y, z\}$

$$\gamma(y,z) = \sum_w \psi(w)\,\phi_{yw}(y,w)\,\phi_{zw}(z,w)$$

$\Downarrow$



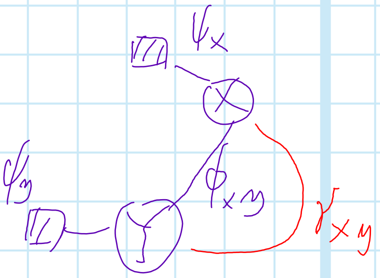
② Eliminate z

$F(z) = \{\phi_{xz}, \phi_{yz}, \gamma_{yz}, \psi_z\}$

$N(z) = \{x, y\}$

$$\gamma(x,y) = \sum_z \phi_{xz}(x,z)\,\phi_{yz}(y,z)\,\gamma_{yz}(y,z)\,\psi_z(z)$$

$\Downarrow$



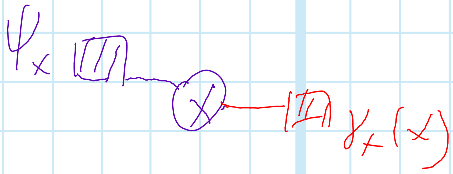
③ Eliminate y

$$f(y)=\{\phi_{xy},\ \psi_y,\ \gamma_{xy}\}$$

$$N(y)=\{x\}$$

$$\gamma_x(x)=\sum_y \phi_{xy}(x,y)\,\psi_y(y)\,\gamma_{xy}(x,y)$$

$$\Downarrow$$



$$\therefore P(x)\propto \psi_x(x)\,\gamma_x(x)$$

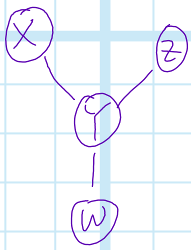
下面分析复杂度:

①: $O(m_y\cdot m_z\cdot m_w)$

②: $O(m_x\cdot m_y\cdot m_z)$

③ $O(m_x\cdot m_y)$

下面来看 order $\pi(j)$ 对 Complexity 的影响.



求 $p(x)$.

$\pi_1 = (y, w, z) \longrightarrow O(m^4)$ ∵ y 有很多 neighbour.

$\pi_2 = (w, z, y) \longrightarrow O(m^2)$

那么如何选择 optimal order

是一个 NP 问题，一般靠经验.

一般优先选 neighbour 少的变量来消除.

例题：计算复杂度.

A chain of discreate varibe

$x_1 - x_2 - \cdots - x_n$ $\quad n > 3$

求 $p(x_1)$

space conditionality: m 每个 x_i 有 m 个取值.

— direct formulation $O(?)$ $\quad O(m^n)$

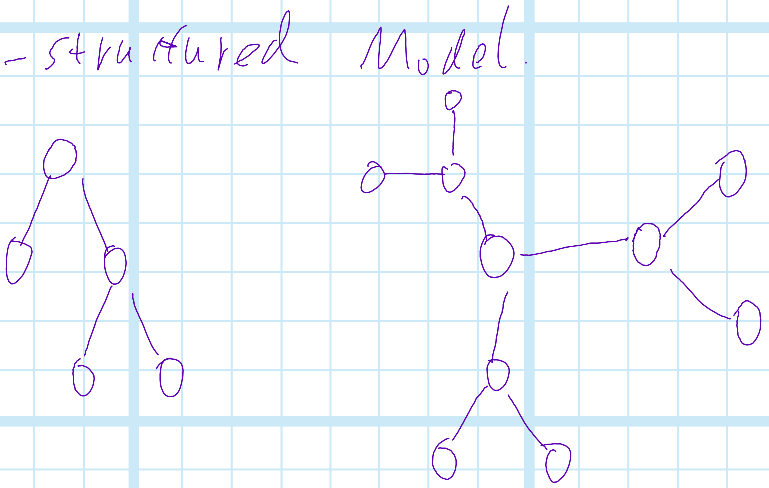
— variable elimination $O(?)$ $\quad O(n \cdot m^2)$

问题，如果要对很多个 x 计算 $p(x)$，里面也有很多相同的计算，能不能合并？下面讲 Belief Propagation

Belief Propagation

如何对所有的x来计算p(x)

Tree-structured Model

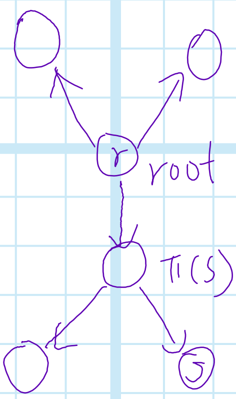


Tree: contains no cycle.

Pairwise MRF

$$p(x)=\frac{1}{Z}\prod_{s\in V}\underbrace{\psi_s(x_s)}_{\text{uniary term}}\prod_{(s,t)\in E}\underbrace{\psi_{st}(x_s,x_t)}_{\text{binary term}}$$

Tree-structured 的优点：存在很快的inference算法。



root 任选一个root.

每个edge连接了一个parent和children

$$p(x)=\frac{1}{Z}\overbrace{\prod_{s\in V}\psi_s(x_s)}^{n}\prod_{s\in V/\{root\}}\overbrace{\phi_s(x_{\pi(s)},x_s)}^{n-1}$$

$$=\frac{1}{Z}\psi_r(x_r)\prod_{s\in V/\{r\}}\psi_s(x_s)\,\phi_s(x_{\pi(s)},x_s)$$

下面看 root 的 marigional distr, 然后扩展
到所有的结点. 最后扩展到任意的 graph

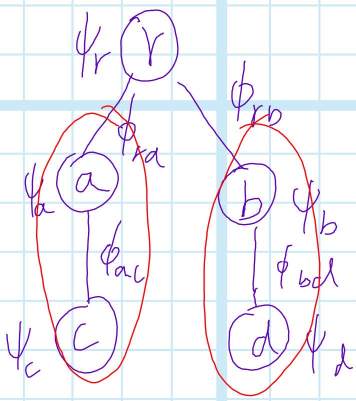
2018.10.15

Tree-Structured Model

$$p(x) = \frac{1}{Z}\psi_r(x_r)\prod_{s\in V(T)\backslash r}\psi_s(x_s)\,\phi_s\left(x_{\pi(s)}, x_s\right)$$

每个非根结点和父结点相连

先计算如下的 $p(x_r)$，然后 generalize 到一般模型



$$p(x_r) = \frac{1}{Z}\sum_{x_a}\sum_{x_b}\sum_{x_c}\sum_{x_d}\psi_r(x_r)\psi_a(x_a)\psi_b(x_b)\psi_c(x_c)\psi_d(x_d)$$
$$\phi_{ra}(x_r,x_a)\phi_{rb}(x_r,x_b)\phi_{ac}(x_a,x_c)\phi_{bd}(x_b,x_d)$$

$$= \frac{1}{Z}\psi_r(x_r)\sum_{x_a}\psi_a(x_a)\phi_{ra}(x_r,x_a)\sum_{x_c}(x_c)\phi_{ac}(x_a,x_c)$$ 第一行只和 a,c 有关

$$\sum_{x_b}\psi_b(x_b)\phi_{rb}(x_r,x_b)\sum_{x_d}\psi_d(x_d)\phi_{bd}(x_b,x_d)$$ 第二行只和 b,d 有关

可以发现是按 sub tree 分解的

decomposition along sub trees

先介绍几个术语:

T_s : sub-Tree rooted at s

ch(s): children of s $\quad ch(r)=\{a,b\},\ ch(a)=\{c\}$

$V(s)$: $V(T(s))$ 例: $V(a) = \{a, c\}$ all the vertices contained in $T(s)$

$V(r) = V$

$D(s) = V(s) / \{s\}$ decedents

定义 $w_s(x_{V(s)}) = \psi_s(x_s) \prod_{t \in D(s)} \psi_t(x_t) \phi_t(x_{\pi(t)}, x_t)$

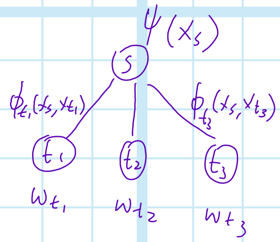
$w_r(x_r) = \tilde{p}(x_r)$

$p(x) \propto w_r(x_V)$ 只差了个Z

Leaf t: $w_t(x_t) = \psi_t(x_t)$

Non leaf s:

$$w_s(x_{V(s)}) = \psi(x_s) \prod_{t \in Ch(s)} \phi_t(x_s, x_t)\, w_t(x_{V(t)})$$



$w_r(x) = \psi_r \psi_a \psi_b \psi_c \psi_d$ 递归求解

$\phi_{ra} \phi_{rb} \phi_{ac} \phi_{bd}$

$= \psi_r (\phi_{ra} \underbrace{\psi_a \psi_c \phi_{ac}}_{w_a})$

$(\phi_{rb} \underbrace{\psi_b \psi_d \phi_{bd}}_{w_b})$

$= \psi_r (\phi_{ra} w_a)(\phi_{rb} w_b)$

有了 w_s, 就可以计算:

$$f_s(x_s) = \sum_{x_{D(s)}} w_s(x_s, x_{D(s)})$$

$D(s) = V(s) \setminus s$, $f_s(x_s)$ 就是把除 x_s 外的所有变量 Marginalize 掉

∴计算整个网络用 w_s, 而计算某个变量的marginal 用 f_s

然后 $p(x_r) \propto f_r(x_r)$

下面看 $f_s(x_s)$ 的计算

Leaf t:

$f_t(x_t) = \psi_t(x_t)$ ∵没有children

non-leaf s:

$$f_s(x_s) = \sum_{\substack{x_{D(s)} \\ 空-X}} w_s(x_s; x_{D(s)})$$

$D(r)$ 可以分解为不相交的 (a,c) 和 (b,d)

$ch(s) = \{t_1, \cdots, t_k\}$

$D(s) = V(t_1) \cup \cdots \cup V(t_k)$

$w_s(x_{V(s)}) = \psi_s(x_s) \prod_{t \in D(s)} \psi_t(x_t) \phi_t(x_{\pi(t)}, x_t)$

$$= \psi_s(x_s) \cdot \sum_{x_{D(s)}} \begin{matrix} \phi_{t_1}(s,t_1) w_{t_1}(x_{V(t_1)}) \\ \vdots \\ \phi_{t_k}(s,t_k) \cdot w_{t_k}(x_{V(t_k)}) \end{matrix}$$

$$= \psi_s(x_s) \cdot \begin{matrix} \sum_{x_{V(t_1)}} \phi_{t_1}(s,t_1) w_{t_1}(x_{V(t_1)}) \\ \vdots \\ \sum_{x_{V(t_k)}} \phi_{t_k}(s,t_k) \cdot w_{t_k}(x_{V(t_k)}) \end{matrix}$$

$$= \psi_s(x_s) \prod_{t \in ch(s)} \cdot \sum_{x_{V(t)}} \phi_t(x_s, x_t)\, w_t(x_{V(t)})$$

$$= \psi_s \prod_{t \in ch(s)} \sum_{x_t, x_{D(t)}} \phi(x_s, x_t)\, w_t(x_{V(t)})$$

$$= \psi_s \prod_{t \in ch(s)} \sum_{x_t} \phi(x_s, x_t) \sum_{x_{D(t)}} w_t(x_t, x_{D(t)})$$

$$= \psi_s \prod_{t \in ch(s)} \cdot \sum_{x_t} \phi_t(x_s, x_t) f_t(x_t)$$

Goal:

$$p(x_r) \propto f_r(x_r)$$

leaf t: $f_t(x_t) = \psi_t(x_t)$

Non-leaf:

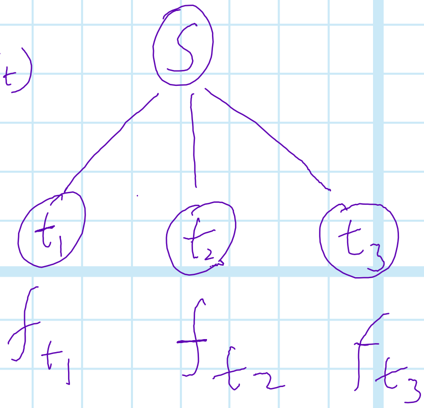
$$f_s(x_s) = \psi_s(x_s) \prod_{t \in ch(s)} \cdot \underbrace{\sum_{x_t} \phi_t(x_s, x_t) f_t(x_t)}$$

复杂度: $O(m_s) \cdot \left(O(|ch(s)|) \sum_{t \in ch(s)} O(m_s \cdot m_t) \right)$

下面从Message的角度来理解这个式子

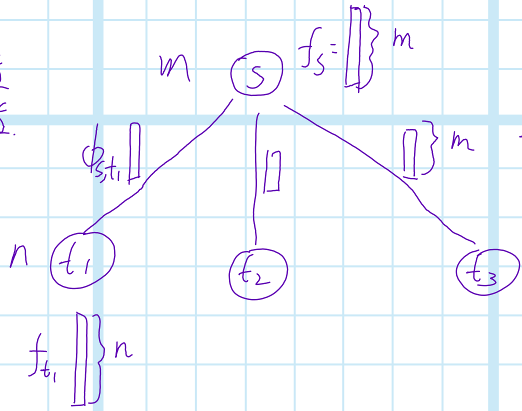
定义 $M_{t\to s}(x_s) \triangleq \sum_{x_t} \phi_t(x_s, x_t) f_t(x_t)$

那么 $f_s(x_s) = \psi_s(x_s) \prod_{t\in Ch(s)} M_{t\to s}(x_s)$



下面看如何实现.

s有m个取值
t有n个取值.



t固定, s有m种取法.

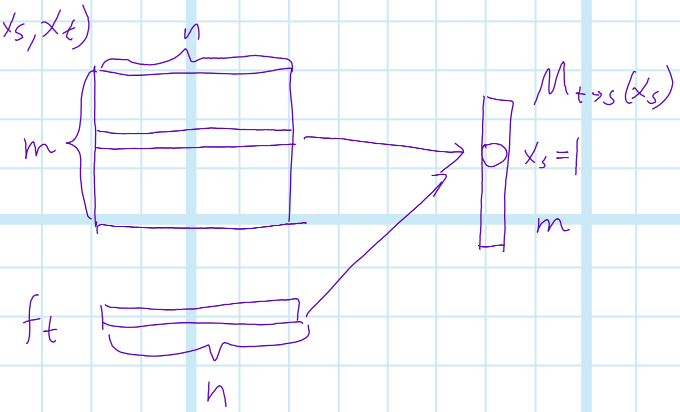
f_{t_1} 就是个n维向量.

$$M_{t_1\to s}(x_s) = \sum_{x_{t_1}} \phi_{st}(x_s, x_{t_1}) f_t(x_{t_1})$$

$$f_s(x_s) = \psi_s(x_s) \cdot \prod_{t\in Ch(s)} M_{t\to s}(x_s)$$

下面看 $M_{t\to s}(x_s)$

$\phi_{st}(x_s, x_t)$



2018.10.19.

How to read a paper

看 key graph

Basic idea 是什么. 在模块中的作用.

例如 corelated topic model 就是用了个 softmax adapter

key advantage your model can bring:

Boundary of the model.

How you should formulate the model

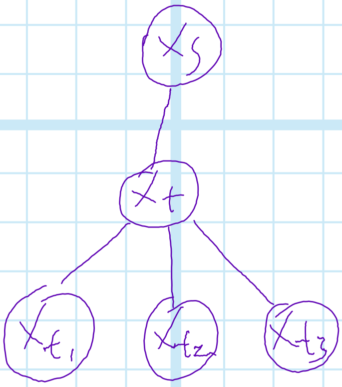
Identify the key problem.

Understand the problem, why existing model not good.

数学分布之类的只是工具.

2018.10.22

Belif propagation 回顾.



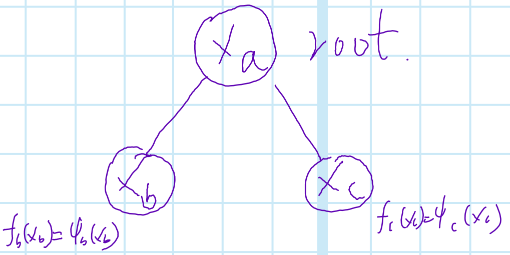
Tree based

send message from children to parent.

$M_{t\to s}(x_s)$

$$M_{t\to s}(x_s) = \sum_{x_t} \phi(x_s, x_t)\,\psi_t(x_t) \prod_{u\in ch(t)} M_{u\to t}(x_t)$$

上式只计算 root 的边缘分布, 如何计算所有结点的分布.



$$M_{b\to a}(x_a) = \sum_{x_b} \phi_{ab}(x_a, x_b)\,\psi_b(x_b)$$

$$M_{c\to a}(x_a) = \sum_{x_c} \phi_{ac}(x_a, x_c)\,\psi_c(x_c)$$

$$f_a(x_a) = \psi_a(x_a)\cdot M_{b\to a}(x_a)\cdot M_{c\to a}(x_a)$$

$$= \psi_a(x_a) \sum_{x_b} \phi_{ab}(x_a, x_b)\,\psi_b(x_b) \sum_{x_c} \phi_{ac}(x_a, x_c)\cdot\psi_c(x_c)$$

$$= \sum_{x_b}\sum_{x_c} \psi_a(x_a)\psi_b(x_b)\psi_c(x_c)\phi_{ab}(x_a,x_b)\phi_{ac}(x_a,x_c)$$

得出了最后边缘的定义.

那么如何计算 b 和 c 的边缘分布?

$$P(x_b) = \sum_{x_a}\sum_{x_c} \psi_a(x_a)\cdot\psi_b(x_b)\cdot\psi_c(x_c)\phi_{ab}(x_a,x_b)\phi_{ac}(x_a,x_c)$$

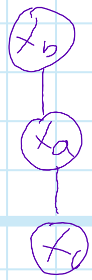
$$= \psi_b(x_b)\underbrace{\sum_{x_a}\psi_a(x_a)\cdot\phi_{ab}(x_a,x_b)\sum_{x_c}\psi_c(x_c)\phi_{ac}(x_a,x_c)}_{\text{定义为 } M_{a\to b}(x_b)}$$

rewrite

$$=\sum_{x_a}\phi_{ab}(x_a,x_b)\psi_a(x_a)M_{c\to a}(x_a)$$

$$=\sum_{x_a}\phi_{ab}(x_a,x_b)\psi_a(x_a)\sum_{x_c}\phi_{ac}(x_a,x_c)\cdot\psi_c(x_c)$$

$\therefore P(x_b)=\psi_b(x_b)M_{a\to b}(x_b)$ 其实就是把 b 当 root



求谁的概率就把谁当 root.

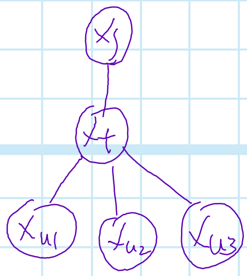
这里计算需要考虑谁是 parent, 谁是 children, 下面给出不依赖于 parent-children 选择的方法.

General Form of Message Passing

$$M_{t\to s}(x_s)=\sum_{x_t}\phi(x_s,x_t)\psi_t(x_t)\prod_{u\in N(t)\setminus\{s\}} M_{u\to t}(x_u)$$

只需要把邻接 $N(t)$ 中的那个 target node s 删掉.

$$M_s(x_s) \propto \psi_s(x_s) \prod_{t\in N(s)} M_{t\to s}(x_s)$$



$N(t)=\{s, u_1, u_2, u_3\}$

$N(t)\setminus s=\{u_1, u_2, u_3\}$

下面分析复杂度.

Complexity Analysis.

each edge (s,t) $\quad M_{s\to t}(x_t)$ $\quad M_{t\to s}(x_s)$

$|x_t| \quad |x_s|$

多少个 message.

- $\sum_{s\in V} \deg(s)\cdot|x_s|$ space-complexity to store all messages

$|x|$

$2 n_{edges}|x| = 2(|V|-1)\cdot|X|$ 总共这么多个 message.

for a tree $|E|=|V|-1$

- time complexity 每个 message 的复杂度.

$M_{t\to s}(x_s)$

compute $|x_s|$ values

$|x_t|\cdot$ terms/values

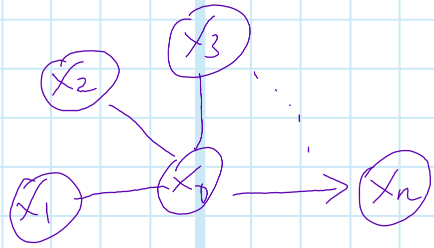
$\therefore O(|x_s||x_t|) = O(m^2)$

$\therefore$ 总复杂度为 $O(|V|\cdot|x|^2)$

下面看例子.

Star Graph

$\forall\ t=1,\cdots,n$



$$m_{t\to 0}(x_0)=\sum_{x_t}\phi_t(x_0,x_t)\psi_t(x_t)$$

$$m_{0\to t}(x_t)=\sum_{x_0}\phi_t(x_0,x_t)\underbrace{\prod_{u\in N_0/\{t\}} m_{u\to 0}(x_0)}_{\text{计算量大，下面来简化.}}$$

令 $m'(x_0)=\prod_{u\in N_0} m_{u\to 0}(x_0)$

$$\prod_{u\in N_0/\{t\}} m = \frac{m'(x_0)}{m_{0\to u}(x_0)}$$

$$m_0(x_0)\propto \psi_0(x_0)\cdot\prod_{u\in N_0} m_{u\to 0}(x_0)$$

$$m_t(x_t)\propto \psi_t(x_t)\cdot m_{0\to t}(x_t)$$

Chain

x_0 — x_1 — $\cdots$ — x_T

$$m_{i_0\to i_1}=\sum_{x}\phi(x_{i_0},x_{i_1})\psi_{i_0}(x_{i_0})m_{i_0-1,i_0}(x_{i_0})$$

$i_0+1=i_1$

$$m(x_i)=\psi_i(x_i)m_{i-1\to i}(x_i)m_{i+1\to i}(x_i)$$

下周讲带cycle的.

2018.10.26

Bethe Interpretation

考虑 MRF:

$$P_\theta(s) = \frac{1}{Z(\theta)} \prod_{s \in V} \psi_s(x_s) \cdot \prod_{(s,t) \in E} \phi_{st}(x_s, x_t)$$

$$\forall s, t, \in \mathcal{X}_s \leftarrow \text{finit space.}$$

下面用 exp family 来表示

— Index trick

$$\psi_s(x_s) \quad x_s \in \{0, \cdots, m_s - 1\}$$

$$= \exp(\log \psi_s(x_s))$$

$$= \exp\left(\sum_{k \in X_s} 1(x_s = k) \log \psi_s(k)\right)$$

$$= \exp\left(\sum_{i \in X_s} \theta_s^i \, 1(x_s = k)\right)$$

同理:

$$\phi_{st}(x_s, x_t) = \exp\left(\sum_{i \in X_s} \sum_{j \in X_t} \theta_{st}^{ij} 1(x_s = i) 1(x_t = j)\right)$$

∴ 用 index trick 来把 概率分布写成指数和的形式:

$$P_\theta(x) = \frac{1}{Z(\theta)} \exp\Big(\sum_{s \in V} \sum_{i \in X_s} \theta_s^i \, 1(x_s = i) +$$
$$\sum_{(s,t) \in E} \sum_{i \in X_s} \sum_{j \in X_t} \theta_{st}^{ij} 1(x_s = i) 1(x_t = j)\Big)$$

其中 $\Theta = \{(\theta_s)(\theta_{st})\}$ 是 canonical parameters

$\mu_s \in R^{|X_s|}$ is mean parameter.

$\mu_{st} \in R^{|X_s| \cdot |X_t|}$

Inference Problem:

Θ are given

mean parameters should be get.

μ_s and μ_{st}.

— Global consist 记作 $M(G)$

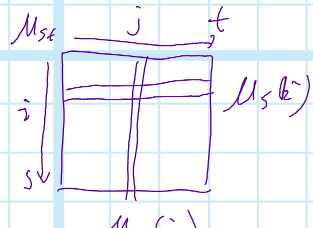
μ_s, μ_{st} are consist with some drawn distr. It's hard to calc,

Instead we use relaxed local const.

— Local consistency 记作 $L(G)$

$\mu_s(x_s) = p(x_s = i)$

$\mu_{st}(i,j) = p(x_s = i, x_t = j)$



观察: ① $\sum_{i \in X_s} \mu_s(i) = 1$

② $\sum_{j \in X_t} \mu_{st}(i,j) = \mu_s(i) \quad \forall i \in X_s$

③ $\sum_{i\in X_s} \mu_{st}(i,j) = \mu_t(j) \quad \forall j \in X_t$

$\sum_{i,j} \mu_{st}(i,j) = 1$ (这个已经被第1个表示了，∴不用写)

$\mu = \{(\mu_s)_{s\in V}, (\mu_{st})_{st\in E}\}$ μ表示所有solution

那么global 和 local constancy 的关系是：

$$\mu \in M(G) \Rightarrow \mu \in L(G)$$

$$M(G) \subseteq L(G)$$

$$\boxed{\begin{array}{l}\text{if } G \text{ is a tree} \\ \text{then } M(G) = L(G)\end{array}}$$

如何把 μ 和 tree-structured model 联系起来.

Tree-structured

$$P(x) = P_r(x_r) \prod_{v\in V/r} P_v(x_v \mid x_{\pi(v)}) \rightarrow \frac{P(x_v, x_{\pi(v)})}{P(x_{\pi(v)})}$$

$$= \mu_r(x_r) \prod_{s\in V/r} \frac{\mu_{\pi(s),s}(x_{\pi(s)}, s)}{\mu_{\pi(s)}(x_{\pi(s)})}$$

$$= \mu_r(x_r) \prod_{s\in V/r} \frac{\mu_{\pi(s),s}(x_{\pi(s)}, s)}{\mu_{\pi(s)}(x_{\pi(s)})\mu_s(x_s)} \underbrace{\mu_s(x_s)}_{\text{放到前面去}}$$

$$= \prod_{v\in V} \mu_v(x_v) \cdot \prod_{(s,t)\in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\cdot \mu_t(x_t)} \qquad (*)$$

以上说明了如何用 mean parameter 来表示 tree-structured model

只要我们有一个 locally consistency μ, tree-structured model 就能写成这种形式

下面看 exp family

$$\hat{\mu} = \arg\max_{\mu \in \underbrace{M(G)}_{\text{all set of realizable mean}}} \{\theta^T \mu - A^*(\mu)\}$$

A^* ← convex conjugate of log partial function

$$\boxed{\hat{\mu} = \arg\max_{\mu \in M(G)} \{\theta^T \mu + H(\mu)\}}$$

两个位置:

$M(G)$: realizable, global constancy are the same thing.

非常难计算，甚至验证是都很难，但幸运的是，tree-structure: $M(G) = L(G)$

$H(\mu)$: entropy

由信息论:

代入 $(*)$ 得:

$$H(\mu) = -\sum_x p \log p$$

$$= -\sum_x p(x) \cdot \left(\sum_{v\in V} \log \mu_v(x_v) + \sum_{(s,t)\in E} \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}\right)$$

$$\boxed{H(\mu) = \sum_{v\in V} H_v(\mu_v) - \sum_{(s,t)\in E} I_{st}(\mu_{st})}$$ only applies to tree-structured models.

entropy   mutual information

$$H_s(\mu_s) = -\sum_{x\in X_s} \mu_s(x)\log \mu_s(x)$$

$$I_{st}(\mu_{st}) = \sum_{(x_s,x_t)\in X_s\times X_t} \mu_{st}(x_s,x_t)\log\frac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\,\mu_t(x_t)}$$

由上, decomposed the entropy along the tree

为什么要研究 tree-structured model?

好处:

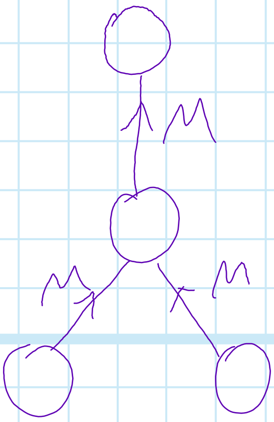
① $M(G) = L(G)$

② Entropy can be factorize into several terms which only contains one or two terms

这样, tree-structured model 同时解决了 $M(G)$ 和 H 的问题.

那么能否推广到所有的图?

可以用 approximation.



Belief propagation

但实际上写程序的时候，直接把这个算法扔到任意图上去跑就行了。效果还可以。后来人们开始找理论支持。

For Loopy-structured

$$M(G) \approx L(G)$$

$$H_{Be}(\mu) \approx \sum_{s} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st})$$

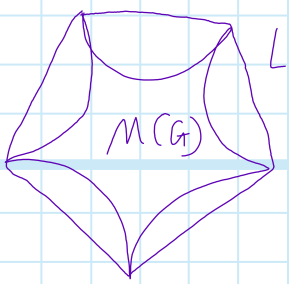
Bethe Entropy

$$\hat{\mu} = \arg\max_{\mu \in L(G)} \{\theta^T \mu + H_{Be}(\mu)\}$$

Bethe Variational

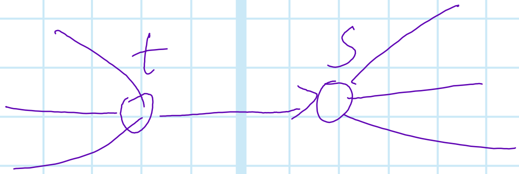
$\Downarrow$ optimization

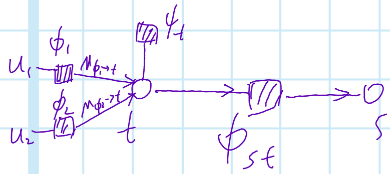
Loopy Belief Propagation (LBP)



L(G) : convex bound.

Loop Belief Propagation 也可以扩展到不止两个factor





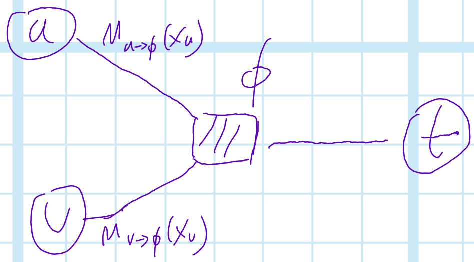
$$M_{t\to s}(x_s)=\sum_{x_t}\phi(x_s,x_t)\psi_t(x_t)\prod_{u\in N_t/\{s\}} M_{u\to t}(x_t)$$

$$=\sum_{x_t}\phi(x_s,x_t)M_{t\to\phi}(x_t)$$

$$M_{t\to\phi}(x_t)=M_{\psi_t\to t}(x_t)\cdot M_{\phi_1\to t}(x_t)M_{\phi_2\to t}(x_t)$$

$$M_{\phi\to s}(x_s)=\sum_{x_t}\phi(x_s,x_t)M_{t\to\phi}(x_t)$$

以上是 binary factor, 下面给 3 factors



$$M_{\phi\to t}(x_t)=\sum_{x_u}\sum_{x_v}\phi(x_u,x_v,x_t)M_{u\to\phi}(x_u)M_{v\to\phi}(x_v)$$

下面继续 Inference 的问题.

已知 θ 求 μ.

① $L(G)\approx M(G)$ 和之前相同.

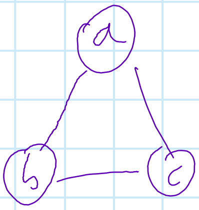
② New approximation of the Entropy.

原为: $\hat{\mu}=\arg\max_{\mu\in M(G)}\{\theta^T\mu-A^*(\mu)\}$

$$A^*(\mu) = \sup_{\theta}\{\theta^T\mu - \underbrace{A(\theta)}_{\text{also intractable.}}\}.$$

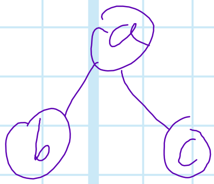
$A(\theta)$ is easy to compute for tree-structured models.

对任意图，可以分解为多个树的组合.



$$\exp(\theta_a \cdot 1(x_a) + \theta_b 1(x_b) + \theta_c \cdot 1(x_c) + \theta_{ab} 1(x_a, x_b) + \theta_{ac} 1(x_a, x_c) + \theta_{bc} 1(x_b, x_c))$$

set $\theta_{bc} = 0 \Rightarrow$



project the parameters into some subspace, it will becomes to tree.

① projection
$$\begin{cases} \theta \xrightarrow{\theta_{bc}=0} \theta' \\ \theta \xrightarrow{\theta_{ac}=0} \theta'' \end{cases}$$

② combination: $\alpha\theta' + (1-\alpha)\theta'' = \theta$

$$A(\theta) = A(\alpha \cdot \theta' + (1-\alpha)\theta'')$$
$$\le \alpha \underbrace{A(\theta')}_{\text{easy}} + (1-\alpha)\underbrace{A(\theta'')}_{\text{easy.}} \quad \because A \text{ is convex.}$$

这样由 θ 的凸组合的形式来估计 $A(\theta)$

上节课: 2018.10.29

Inference on exp family

$$\exp(\theta^T \phi(x) - A(\theta))$$

目标: $\theta \to \mu$.

法① $\mu = E_\theta[\phi(x)]$

法② $A(\theta) = \sup_{\mu} \{\theta^T\mu - A^*(\mu)\}$

$$= \sup_{\mu} \{\theta^T\mu + H(\mu)\}$$

estimate $A(\theta)$, is related to

$$\theta^T\mu + H(\mu)$$

$A(\theta)$ for tree-structure is tractable,

we can decompose non-tree to tree.

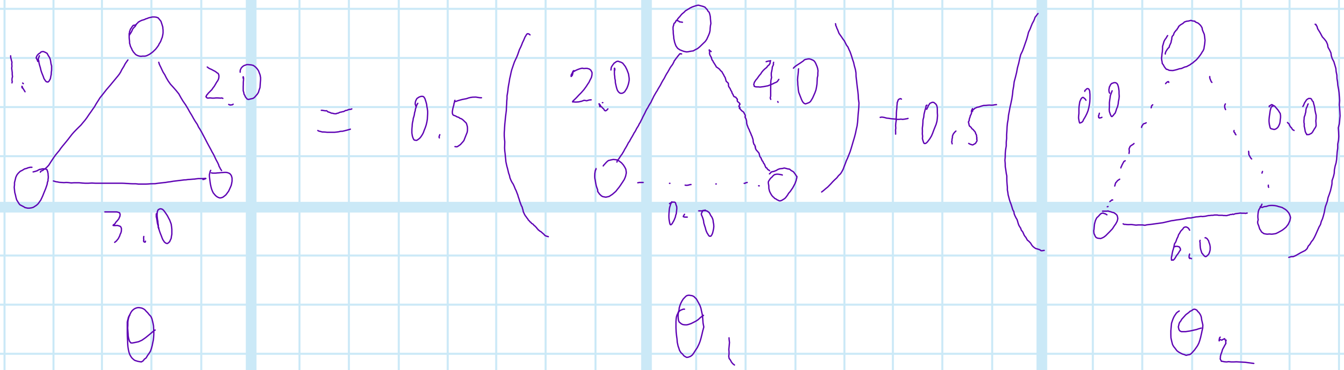
A is convex.

$$\theta = \alpha_1\theta_1 + \alpha_2\theta_2 \quad (\alpha_1 + \alpha_2 = 1)$$

$$A(\theta) \le \alpha_1 A(\theta_1) + \alpha_2 A(\theta_2)$$

通过分解 θ, 让其变成 tree-structured.

例：

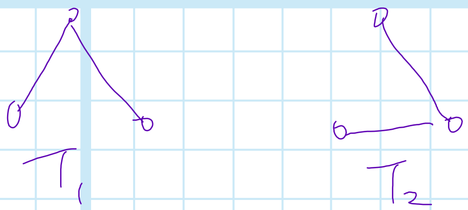


$\theta = 0.5\,\theta_1 + 0.5\,\theta_2$.

通过不同的分解能得到不同的 upper bound

能否找到一个最小的上界？

Find the best upper bound.



$\theta(T_1) = \begin{pmatrix} 0.2 \\ 0.4 \\ 0.0 \end{pmatrix} \qquad \theta(T_2) = \begin{pmatrix} 0.0 \\ 6.0 \\ 6.0 \end{pmatrix}$

$$\sum \rho(T)\,\theta(T) = E_\rho[\theta(T)]$$

$$\min \sum_{T\in\Sigma} \rho(T)\,A(\theta(T)) \to E_\rho[A(\theta(T))]$$

$$\text{s.t. } \sum_T \rho(T)\,\theta(T) = \bar{\theta}$$

target parameter

给定的θ，各个T的θ组合起来应等于θ̄

consistent constraint 等价 $E_\rho(\theta(T))$

用拉格朗日式解:

$$L(\theta,\mu) = E_\rho[A(\theta(T))] + \langle \mu, \bar{\theta} - E_\rho(\theta(T)) \rangle$$

μ和$\bar{\theta}$维度相同

$$= \mu^T \bar{\theta} + E_\rho[A(\theta(T)) - \mu^T \theta(T)]$$

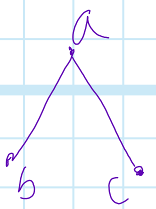
$$= \mu^T \theta + \sum \rho(T)\left(A(\theta(T)) - \mu^T \theta(T)\right)$$

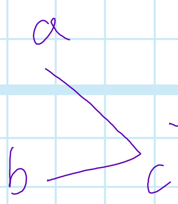
$$\frac{\partial L(\theta,\mu)}{\partial \theta(T)} = \rho(T)[\nabla A(\theta(T)) - \mu] = 0$$

$$E_{\theta(T)}[\phi_\alpha] = \mu_\alpha$$

α表示和某个树相关的参数。对于某个树,并不是所有参数都要求导.

例:

$\theta(T_1)$  $\rightarrow \alpha_1 = \begin{pmatrix} ab \\ ac \end{pmatrix}$

$\theta(T_2)$  $\rightarrow \alpha_2 = \begin{pmatrix} ac \\ bc \end{pmatrix}$

$$\begin{bmatrix} \mu_{ab} \\ \mu_{ac} \\ \mu_{bc} \end{bmatrix}$$

$\mu_{ab}, \mu_{ac} \rightarrow = E_{\theta(T_1)}[\phi_{\alpha_1}(x)]$

$\mu_{ac}, \mu_{bc} \rightarrow = E_{\theta(T_2)}[\phi_{\alpha_2}(x)]$

In order to calc μ, we have to move to the dual problem.

$$A^*(\Pi_T(\hat{\mu})) = \langle \hat{\theta}(T), \hat{\mu} \rangle - A(\hat{\theta}(T))$$

$$L(\hat{\theta}, \hat{\mu}) = \hat{\mu}^T \bar{\theta} + E_\rho[-A^*(\Pi_T(\hat{\mu}))]$$

$$= \hat{\mu} \bar{\theta} - E_\rho[A^*(\Pi_T(\hat{\mu}))]$$

$$= \hat{\mu} \bar{\theta} + E_\rho[H_T(\Pi_T(\hat{\mu}))]$$

projection of entropy to each tree.

$$\sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st})$$

等于每个 tree 的 entropy 之和减去互信息.

$$E_\rho[H_T(\Pi_T(\hat{\mu}))] = E_\rho[\underbrace{\sum_{s \in V} H_s(\mu_s)}_{\text{对所有 tree 都相同, 移出去.}} - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st})]$$

$$= \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\mu_{st})$$

here don't use $E(T)$, it mean for all pairs in the graph.

edge appearance probability

$$= H_{Taw}(\mu)$$

$$H_{se}(\mu) = \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})$$

例：edge appearance probability

$$\triangle_{abc} = 0.4\,\wedge + 0.2\,\angle + 0.4\,>$$

$\rho_{ab} = 0.6$

$\rho_{bc} = 0.6$

$\rho_{ac} = 0.8$

总结：

$\max\ \mu^T\theta + H_{approx}(\mu)$ is a common way to do inference.

另一种方法就是 Variational Inference.

2018.11.2

Learning

- Parameter estimation

  Data → Parameter

  说法: varintional estimation

  varintional inference

Space of distributions

in typical space, we have distance.

how to measure the distance bettwen two distribution?

Use K-L Divergence

two distribution p q.

$$D_{KL}(p\|q) = E_p\left(\log\frac{p(x)}{q(x)}\right)$$

some basic property:

① not symmetric

$$D_{KL}(p\|q) \neq D_{KL}(q\|p)$$

② non-negative

$$D_{KL}(p\|q) \geq 0$$

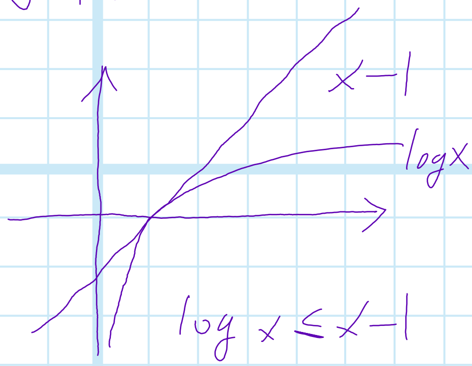
proof:

$$-D_{KL}(p\|q) = E_p\left[\log \frac{q(x)}{p(x)}\right]$$

$$= \int p(x) \cdot \log \frac{q(x)}{p(x)} \mu(dx)$$

$$\leq \int p(x)\left(\frac{q(x)}{p(x)} - 1\right)\mu(dx)$$

$$= \int q(x)\mu(dx) - \int p(x)\mu(dx) = 0$$



$\log x \leq x-1$

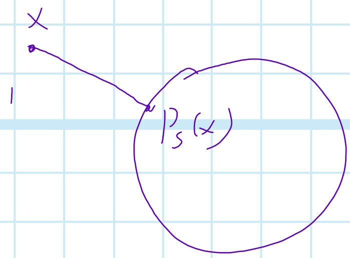
$\log x = x-1$ iff $x=1$

$D(p\|q)=0$ only when $\frac{p(x)}{q(x)}=1$ i.e. $p(x)=q(x)$

有了这个Distance后，我们可以引入projection

projection

$$P_S(x) = \arg\min_{x' \in S} d(x', x)$$



S: convex set

可以对distribution作同样的定义，把某个分布投影到某个分布族上.

∵ KL Divergence 非对称，∴有两种投影方法

设有一个分布族 P

$$I\text{proj}_Q(p) = \arg\min_{q \in Q} D_{KL}(q\|p)$$ information projection

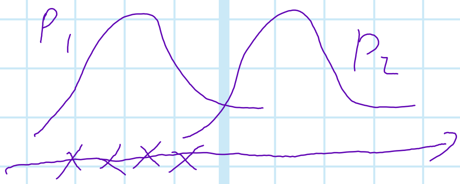
$$M\text{proj}_Q(p) = \arg\min_{q \in Q} D_{K}(p\|q)$$ moment projection

下面回到 Model Estimation

$P_\theta(x)$

已知 $D=\{x_1, \dots, x_n\} \xrightarrow{\text{estimate}} \theta$

basic idea: max likelihood Estimation



P_1 is good.

Likelihood

likelihood vs. density

$$p(x;\theta) \begin{cases} \to P_\theta(x) & \text{density} \\ \to L_x(\theta)=P_\theta(x) & \text{likelihood} \end{cases}$$

when we talk about density, we assume we already know θ

$$L_D(\theta)=\prod_{i=1}^{n} L_{x_i}(\theta) = \prod_{i=1}^{n} P_\theta(x_i)$$

This formula implicity assume samples are independent.

∵ 用乘法会 overflow, ∴ 用 log likelihood

log-likelihood

$$\log L_D(\theta)=\sum_{i=1}^{n} \log P_\theta(x_i)$$

Max $\log l_D(\theta)$ is MLE.

下面给出几何解释

Emprical Distribution

Given $D=\{x_1, \cdots, x_n\}$

$$\tilde{P}_D(x) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(x)$$

$$E_{\tilde{P}_D}(f) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

$$D_{KL}(\tilde{P}_D \| P_\theta) = E_{\tilde{P}_D}\left[\log \frac{\tilde{P}_D}{P_\theta}\right]$$

$$= \underbrace{E_{\tilde{P}_D}[\log \tilde{P}_D]}_{\text{与参数无关}} - E_{\tilde{P}_D}[\log P_\theta]$$

$$= \text{constant} - \underbrace{\frac{1}{n}\sum_{i=1}^{n} \log P_\theta(x_i)}_{\text{object of MLE}}$$

$\therefore$ maximise the likelihood $\Longleftrightarrow$ minimize the KL divergence.

$$P_{\hat{\theta}} = \underset{P_\theta \in P}{\operatorname{argmin}} D_{KL}(\hat{P}_\theta \| P_\theta) = \mathrm{Mproj}_P(\tilde{P}_D)$$

MLE 的几何意义就是一个经验分布

M-projection 距离下，向分布族 P 投影

当这个分布是指数分布族时：

exp family:

$$p_\theta(x) = h(x)\exp(\theta^T\phi(x) - A(\theta))$$

maximize

$$E_{\tilde{p}_D}[\log p_\theta(x)]$$

$$\log L_D(\theta) = E_{\tilde{p}_D}(\theta^T\phi(x) - A(\theta))$$

$$= \theta^T\left(\frac{1}{n}\sum_{i=1}^{n}\phi(x_i)\right) - A(\theta)$$

$$= \theta^T\tilde{\mu}_D - A(\theta) \quad \text{with } \tilde{\mu}_D = E_{\tilde{p}_D}[\phi(x)]$$

$$\nabla_\theta \log L_D(\theta) = \tilde{\mu}_D - \nabla_\theta A(\theta) = 0$$

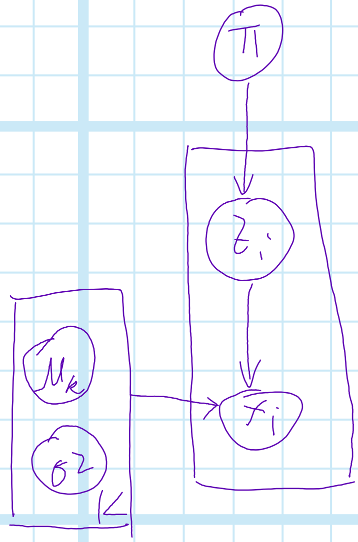
$$\nabla_{\hat{\theta}} A(\theta) = E_{\tilde{p}_D}[\phi(x)]$$

$$\nabla_{\hat{\theta}} A(\theta) = E_\theta[\phi(x)] = E_{\tilde{p}_D}[\phi(x)]$$

M projection try to align the moment of the data and parameter.

下面用高斯混合来做例子.

GMM: σ 已知.



$$p(x_i, z_i \mid \pi, \{\mu_k\})$$

$$= \pi(z_i) \cdot \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu_{z_i})^2}{2\sigma^2}\right)$$

$$= \exp\left(\log \pi(z_i) - \frac{(x-\mu_{z_i})^2}{2\sigma^2}\right)$$

$$= \exp\left(\sum_{k=1}^{K} \delta_k(z_i) \cdot \log \pi(k) - \sum_{k=1}^{K} \delta_k(z_i) \frac{(x-\mu_k)^2}{2\sigma^2}\right) \quad \text{(index trick)}$$

$$D: \sum_{i=1}^{n} \left(\sum_{k=1}^{K} \delta_k(z_i) \log \pi_k - \sum_{k=1}^{K} \delta_k(z_i) \frac{(x-\mu_k)^2}{2\sigma^2} \right)$$

求 π_k 和 μ_k, 分成两个问题.

1. Solve π_k.

$$J(\pi) = \sum_{i=1}^{n} \sum_{k=1}^{K} \delta_k(z_i) \log \pi_k$$

$$= \sum_{k=1}^{K} n_k \log \pi_k \qquad n_k = \#\{i : z_i = k\}$$

$$\text{s.t.} \sum_k \pi_k = 1 \qquad \pi_k \geq 0 \quad \forall k = 1, \dots, K$$

$$\Rightarrow \pi_k \propto n_k = \frac{n_k}{\sum_{l=1}^{K} n_l} = \frac{n_k}{n}$$ 拉格朗日乘子.

2. Solve μ_k

$J(\mu_k) \leftarrow$ minimize

$$= \sum_{i \in S_k} \frac{(x_i - \mu_k)^2}{2\sigma^2}$$

$$\mu_k = \frac{\sum_{i \in S_k} x_i}{|S_k|}$$

找一个点到其它的距离之和最短，就是 mean.

这里我们假设了 z_i 是已知的，z_i 未知的话不能这么解，要用EM，这就是 partial observed model.

$$p_\theta(\underset{\text{observed}}{x}, \underset{\text{latent}}{z})$$

$$= g(x)\, h_x(z) \exp\left(\theta^T \phi(x, z) - A(\theta)\right)$$

x is observed z is unknown.

$$p(z|x) = \frac{p(x,z)}{p(x)} = \frac{h_x(z) \exp(\theta^T \phi(x,z) - A(\theta))}{\int_z h_x(z) \exp(\theta^T \phi(x,z) - A(\theta))\, dz}.$$

这个条件分布仍然是 exp family，可以写成：

$h_x(z) \exp(\theta^T \phi(x) - A(\theta|x))$ 的形式.

$$p(z|z) = \frac{h_x(z)\exp(\theta^T\phi(x,z))}{\int_z h_x(z)\exp(\theta^T\phi(x,z))\,dz}$$

$$\therefore A(\theta|x) = \log \int_z \exp(\theta^T\phi(x,z))\,h_x(z)\,dz.$$

called conditional log-partition function

$$p(x) = \int h_x(z)\exp(\theta^T\phi(x,z) - A(\theta))\,dz$$

$$= \frac{1}{\exp(A(\theta))}\int h_x(z)\exp(\theta^T\phi(x,z))\,dz$$

$$= \frac{\exp(A(\theta|x))}{\exp(A(\theta))} = \exp(A(\theta|x) - A(\theta))$$

$$\log p(x) = A(\theta|x) - A(\theta)$$

这样就写成了和隐变量 z 无关的形式.

2018.11.5

$$P_\theta(x,z) = g(x) h_\theta(z) \exp(\theta^T \phi(x,z) - A(\theta))$$

$$\log \underset{\substack{\Downarrow \\ p(x|\theta)}}{L(\theta|x)} = A(\theta|x) - A(\theta)$$

目标：$\max \sum_i \log p(x_i|\theta)$

但是 $A(\theta|x) = \log \int_z \exp(\theta^T \phi(x,z)) h_x(z) dz$

可能 z 空间很大，可以用EM来求解

今天用EM求解这项。

EM: find lower bound of $\log L(\theta|x)$

$\log(\theta|x)$ is difficult to compute, we can max its lower bound. which is eaiser to compute.

$A(\theta|x)$: conditional log partion function

$A(\theta) = \sup_\mu \{\theta^T \mu - A^*(\mu)\}$ duality between parameter domain and canonical mean domain

$$A(\theta|x) = \sup_\mu \{\theta^T \mu - A^*(\underset{\substack{\Downarrow \\ E_{p(z|x)}[\phi(x,z)]}}{\mu|x})\}$$

$\Downarrow$ lower bound

$A(\theta|x) \geq \theta^T \mu - A^*(\mu|x) \quad \forall \mu$. $A^*(\mu|x)$ is conditional entropy

$\log L(\theta|x) \geq \theta^T\mu - A^*(\mu|x) - A(\theta) = Q_x(\theta, \mu)$

得到了 $\log L(\theta|x)$ 的下界，当 μ 取到 arg max 时，等号成立

左边只有一个参数，右边有两个参数 μ, θ，

但是可以保证任取 θ 都使不等式成立。

用 coodinate asend 来优化 $Q_x(\theta, \mu)$

Coodinate ascent

$\max f(x, y)$

fix $x^{(t)}$

$y^{(t+1)} \leftarrow \arg\max_y f(x^{(t)}, y)$

$x^{(t+1)} \leftarrow \arg\max_x f(x, y^{(t+1)})$

回到之前那个 $Q_x(\theta, \mu)$，EM 算法，

分为 E-step and M-step.

$\hat{\mu} \leftarrow \arg\max_\mu Q(\theta; \mu)$

$= \arg\max_\mu \theta^T\mu - A^*(\mu|x)$

$\hat{\mu} = E_\theta[\phi(x, z)]$

实质上就是在计算expection

M-tep:

$\hat{\theta} \leftarrow \arg\max_\theta Q(\theta; \mu)$

$= \arg\max_\theta \theta^T\mu - A(\theta)$

这里 E 和 M 的形式都一样，但是计算

可以解释为计算 mean，然后 do the model estimation

下面看为什么 EM 能 work.

$$\log L_x(\theta^{(t)})$$

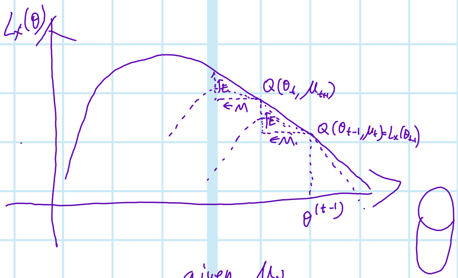
$\leftarrow x$ is constant

$$\mu^{(t+1)} = \arg\max_{\mu} Q(\theta^{(t)}, \mu)$$

$$Q(\theta^{(t)}, \mu^{(t+1)}) = L_x(\theta^{(t)})$$ μ取到最优，等号成立.

$$\geq Q(\theta^{(t)}, \mu^{(t)})$$

$$\geq Q(\theta^{(t-1)}, \mu^{(t)}) = L_x(\theta^{(t-1)})$$



given μ_t,
$\theta_t \leftarrow \arg\max_{\theta} Q(\theta, \mu_t)$

E: close the gap (μ, θ are coupled)

M: move to another procedure.

EM in distribution space.

$$KL_x(\mu \| \theta)$$

try to get the KL divergence between two distr. in which one is parameterize by mean parameter μ

another is paramter by canonical parameter θ

$$KL_x(\mu \| \theta)$$
$$= A(\theta | x) + A^*(\mu | x) - \theta^T \mu$$

$$Q(\theta; \mu)$$
$$= \mu^T \theta - A^*(\mu | x) - A(\theta)$$

$$L(\theta|x) - Q(\theta, \mu)$$

$$= (A(\theta|x) - A(\theta)] - [\mu^T\theta - A^*(\mu|x) - A(\theta)]$$

$$= A(\theta|x) + A^*(\mu|x) - \mu^T\theta = KL_x(\mu \| \theta)$$

这个KL距离就是那个gap.

以上只考虑了一个sample x，对于一个sample集合 D

$$L_D(\theta) = \sum_i (A(\theta|x_i) - A(\theta)]$$

$$= \sum_i A(\theta|x_i) - nA(\theta)$$

$$= \sum_i (\mu_i^T\theta - A^*(\mu_i|x_i)) - nA(\theta)$$

$$= \sum_i (\mu_i^T\theta - A^*(\mu_i|x_i) - A(\theta))$$

現对不同的sample有相同的parameter θ 和各自的 mean μ.

⋮

E-Step:

$$\hat{\mu}_i \leftarrow \arg\max_{\mu} \mu^T\theta - A^*(\mu|x_i)$$

respectively for $i=1,\cdots,n$.

M-step

$$\hat{\theta} = \arg\max_{\theta} \{\theta^T \bar{\mu} - A(\theta)\} \qquad \bar{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mu_i$$

与单个的EM的区别：E各自算，M取平均.

2018.11.9

Variational Inference

very closely related to partially observed model, so us EM.

$$P(x,z) = g(x)h_z(z)\exp(\theta^T\phi(x,z) - A(\theta))$$

E-M:

origional: $L_x(\theta)$

alternative: $\geq Q(\theta,\mu) = \mu^T\theta - A^*(\mu|x) - A(\theta)$

E-step: update μ: $\hat{\mu} \leftarrow \underset{\mu}{\operatorname{argmax}}\{\mu^T\theta - A^*(\mu|x)\}$

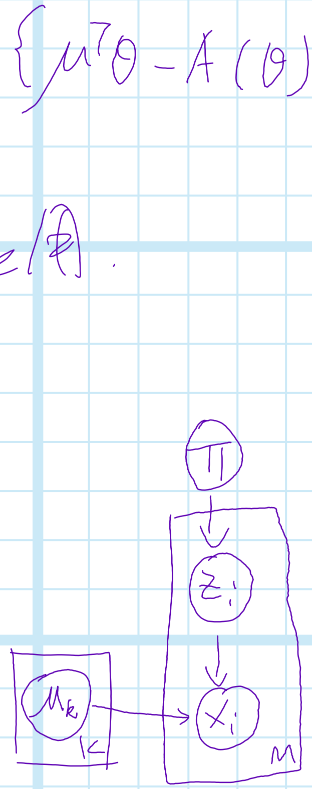
M-step: update θ: $\hat{\theta} \leftarrow \underset{\theta}{\operatorname{argmax}}\{\mu^T\theta - A(\theta)\}$

下面看EM如何在图模型上应用.

Gaussian Mixture Model

$$p(x_i, z_i | \{\mu_k\}, \pi) \propto \pi(z_i)\exp\left(-\frac{(x_i-\mu_{z_i})^2}{2\sigma^2}\right)$$

$$= \exp\left(\sum_{k=1}^{K} 1_k(z_i)\log\pi_k - \sum_{k=1}^{K} 1_k(z_i)\frac{(x-\mu_k)^2}{2\sigma^2}\right)$$



x_i: observed σ^2: known

z_i: latent.

$$= \exp\left(\sum_{k=1}^{K} 1_k(z_i)\log\pi_k + \sum_{k=1}^{K} 1_k(z_i)\frac{x_i\mu_k}{\sigma^2} - \sum_{k=1}^{K} 1_k(z_i)\frac{x_i^2}{2\sigma^2} + \cdots\right)$$

写成标准形式后，充分统计量有 $1_k(z_i)$, $1_k(z_i)x_i$, $1_k(z_i)x_i^2$

E-step: infer the expection

$$E[1_k(z_i)] = q_i^k = P_r(z_i = k)$$

$$E[1_k(z_i)x_i] = q_i^k \cdot x_i$$

$$E[1_k(z_i)x_i^2] = q_i^k \cdot x_i^2 \quad \cdots$$

$\therefore$ 只需计算 q_i^k

$$q_i^k = P_r(z_i = k) \propto \pi_k \cdot \exp\left(-\frac{(x_i-\mu_k)^2}{2\sigma^2}\right)$$

E-step

M-step

q_i^k 类似于一个类型的 soft-assignment.

$$\mu^T\theta - A(\theta)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} q_i^k \log\pi_k - A(\pi)$$

$$\text{s.t. } \pi \in S_K$$

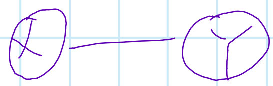
$$\pi_k \propto \sum_{i=1}^{n} q_i^k$$

$$\mu_k = \frac{\sum_{i=1}^{n} q_i^k x_i}{\sum_{i=1}^{n} q_i^k}$$

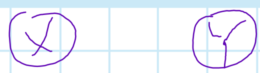
In variational inference, $A^*(\mu|x)$ is very hard to compute. we relay on Entroy Approximation of it.

$$A^*(\mu|x)$$
$$\downarrow$$
$$A_f^*(\mu|x)$$

f: factorized variational distribution.



$P(x,y)$ $\Uparrow$ approximate.



$$p(x,y) \approx q_x(x)\, q_y(y)$$

$q \leftarrow$ parameter: λ q_λ 由λ控制的分布.

$\hat{\lambda} \leftarrow \arg\min D_{KL}(q_\lambda \underset{\text{估计}}{\|} p)$

回忆 MLE 中，在 $\arg\min D_{KL}(\tilde{p}_P \underset{\text{估计}}{\|} p_\theta)$

用右边估计左边，叫做 M-projection

而这里用去也估计分布，不是 I-projection

call mean field approximate.

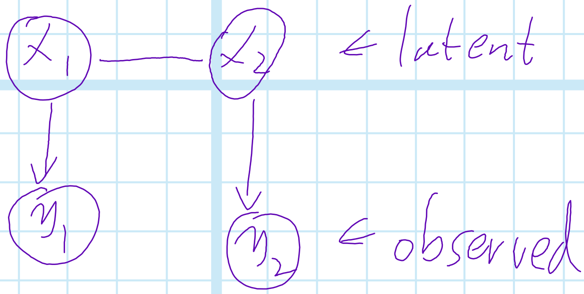
下面看一个例子。

Hidden MRF

会用到很多 tricks.

$p(x_1, x_2)$ prior on latent space

$x_1, x_2 \in \{0, 1\}$



$$p(x_1, x_2)\, p(y_1 | x_1)\, p(y_2 | x_2)$$

$$\propto \exp\left(\psi_1(x_1) + \psi_2(x_2) + \phi(x_1, x_2)\right) \exp\left(f(x_1, y_1)\right) \exp\left(f(x_2, y_2)\right)$$

$$= \exp\Big(\sum_{i \in \{0,1\}} \delta_i(x_i) \cdot \theta_1^i + \sum_{j \in \{0,1\}} \delta_j(x_j)\, \theta_2^j + \sum_{i,j} \theta_{12}^{ij} \delta_i(x_1) \delta_j(x_2)$$

$$+ \sum_i f_1^i \delta_i(x_1) + \sum_j f_2^j \delta_j(x_2)\Big)$$

$\theta_1^i = \psi_1(i) \quad i \in \{0, 1\}$ $\qquad \theta_2^j = \psi_2(j) \quad j \in \{0, 1\}$

$f_1^i = f(i, y_1) \quad i \in \{0, 1\}$ $\qquad \theta_{12}^{ij} = \phi(i, j)$ $\qquad$ 国

canonical	suff. stats
$\delta_i(x_1)$	$\theta_1^i + f_1^i$
$\delta_j(x_2)$	$\theta_2^j + f_2^j$
$\delta_i(x_1)\delta_j(x_2)$	θ_{12}^{ij}

$E[\delta_i(x_1)] = P_r(x_1 = i)$

$E[\delta_j(x_2)] = P_r(x_2 = j)$

$E[\delta_i(x_1)\delta_j(x_2)] = P_r(x_1 = i, x_2 = j)$

∴ 可以合并成三项.

$$= \exp\left(\sum_{i\in\{0,1\}} \delta_i(x_i)\cdot\theta_1^i + \sum_{j\in\{0,1\}} \delta_j(x_j)\theta_2^j + \sum_{i,j} \theta_{12}^{ij}\delta_i(x_1)\delta_j(x_2)\right)$$

就是 Ising Model.

在 E-step 中需要计算 $E[\delta_i(x_1)]$ $E[\delta_j(x_2)]$ $E[\delta_i(x_1)\delta_j(x_2)]$ 很麻烦. 可以用 variational inference 来计算.

问题

$$p(x_1,x_2) = \exp\left(\sum_i \theta_1^i \delta_i(x_1) + \sum_j \theta_2^j \delta_j(x_2) + \sum_{i,j} \theta_{12}^{ij}\delta_i(x_1)\delta_j(x_2)\right)$$

用 Mean field Approx.

$$q(x_1,x_2) = q_1(x_1)q_2(x_2)$$

approx.

$$\hat{q} = \arg\min_{q} D_{KL}(q\|p)$$

$$D_{KL}(q\|p)$$

$$= E_q\left[\log \frac{q}{p}\right]$$

$$= E_q[\log q - \log p)$$

$$= E_{q_1,q_2}[\log q_1 + \log q_2 - \log p]$$

用 coordinate descent

fix q_2, update q_1. 先去掉和 q_1 无关的项

$$E_{q_1}[\log q_1 - \log p)$$

注意到:

$$E_{q_1 q_2}[\log p]$$

$$= E_{q_1 q_2}\Big[\sum_i \theta_1^i \delta_i(x_1) + \sum_j \theta_2^j \delta_j(x_2) +$$

$$\sum_{ij} \theta_{12}^{ij} \delta_i(x_1)\delta_j(x_2)\Big]$$

this is the key to simplify.

$$= \sum_i \theta_1^i q_1^i + \sum_j \theta_2^j q_2^j + \sum_{ij} \theta_{12}^{ij} q_1^i q_2^j$$

化成了一个很简单的优化问题.

回顾:

① E-M:

$$L_x(\theta) \geq Q(\theta,\mu)$$

$$\Downarrow$$

$KL(\mu\|\theta) \leftarrow$ close minimize gap $\rightarrow$ variational inference.

② Ising Model

$$G=(V,E)$$

uniary term, binary term

$$p(x) \propto \exp\Big(\sum_{v\in V}\theta_v x_v + \sum_{(u,v)\in E}\theta_{uv}x_u x_v - A(\theta)\Big)$$

if it's loopy graph.

use Mean-Field Approximation

$$q_\lambda(x) = \exp\Big(\sum_{v\in V}\lambda_v x_v - B_v(\lambda)\Big)$$

use $q_\lambda(x)$ to approximate $p(x)$.

$$D_{KL}(q_\lambda\|p_\theta) = E_q\Big[\log q - \sum_{v\in V}\theta_v x_v - \sum_{(u,v)\in E}\theta_{uv}x_u x_v + A(\theta)\Big]$$

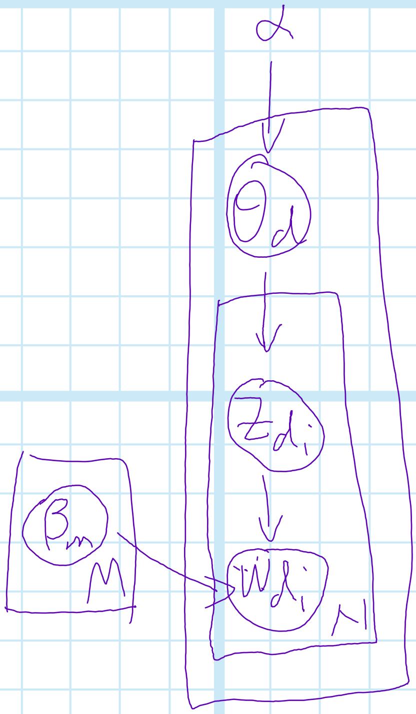
$\leftarrow$ not related to optimization

$$= -\sum_{v\in V}H_v(q_v) - \sum_{v\in V}\theta_v E_{q_v}[x_v] - \sum_{(u,v)\in E}\theta_{uv}E_{q_uq_v}[x_u, x_v]$$

$$= -\sum_{v\in V} H_v(q_v) - \sum_{v\in V}\theta_v q_v - \sum_{(u,v)\in E}\theta_{u,v} q_u q_v$$

下面看LDA的例子.

Latent Direhlet Allocation



$$p(\underbrace{\theta_d, \{z_{d_i}}_{\text{latten}}, \underset{\text{observe}}{w_{d_i}}\} \mid \alpha, \beta)$$

$$\text{Dir}: p(\theta|\alpha) = \frac{1}{B(\alpha)}\prod_{m=1}^{m}\theta_m^{\alpha_m - 1}$$

$$= \exp\left(\sum_{m=1}^{m}(\alpha_m - 1)\log\theta_m - \log B(\alpha)\right)$$

$$\therefore p \propto \exp\Big(\sum_{k=1}^{K}(\alpha_k - 1)\log\theta_d^k \quad \text{①} \quad \leftarrow \text{Dir}$$

$$+ \sum_{i=1}^{n_d}\sum_{k=1}^{m} 1_k(z_{d_i})\cdot\log\theta_d^k \quad \text{②}$$

$$+ \sum_{i=1}^{n_d}\sum_{k=1}^{M} 1_k(z_{d_i})\log\beta_k(w_{d_i})\Big) \quad \text{③}$$

Follow Mean Field Approximation.

$$q(\theta_d, \{z_{d_i}\}) = \underbrace{q_\gamma(\theta_d)}_{\text{Dir}(\gamma)}\prod_{i=1}^{n_d}\underbrace{\rho_{d_i}(z_{d_i})}_{\text{categorical distr.}}$$

① $E_q[\log \theta_d^k] = E_{q_\gamma}[\log \theta_d^k] = \psi(\gamma_d^k) - \psi(\mathbb{1}^T \gamma_d)$

$\left(E[\phi(x)] = \nabla_\theta A(\theta) = \psi(\alpha^k) - \psi(\sum_k \alpha^k) \right)$ ← digamma function

② $E_q[\mathbb{1}_k(z_{di}) \cdot \log \theta_d^k]$

$= E_q[\mathbb{1}_k(z_{di})] \cdot E_{q_\lambda}[\log \theta_{di}^k]$

②是①和③乘起来.

③ $E_q[\mathbb{1}_k(z_{di})] = P_\gamma(z_{di} = k) \leftarrow$ w.r.t. ρ_{di}

$= \rho_{di}(k)$

2018.11.12

回顾

Inference.

$$\theta \longrightarrow E_\theta[f(x)]$$

基本方法：$E_\theta[f(x)] = \int_X f(x)p(x)\mu(dx)$ 　 写完！

对于 exp family：

$$p(x) = h(x)\exp(\theta^T\phi(x) - A(\theta))$$

可以利用它的最重要的式子：

$$A(\theta) = \sup_\mu \{\theta^T\mu - A^*(\mu)\}$$

① 变成了一个优化问题.

$$\hat{\mu} \Leftarrow \arg\max_\mu \ \theta^T\mu - A^*(\mu)$$
$$\qquad\qquad\qquad\qquad \| $$
$$\qquad\qquad\qquad\qquad \theta^T\mu + H(\mu)$$

剩下的就是对 $H(\mu)$ entropy 来估计.

② $\hat{\mu} = E_\theta[\phi(x)] = \nabla_\theta A(\theta)$

EM 算法：

E-step：inference

M-step：$\hat{\theta} = \arg\max_\theta \{\theta^T\mu - A(\theta)\}$

if have complete observe.

$$\mu = \frac{1}{n}\sum_{i=1}^{n}\phi(x_i)$$

has unobservable

利用 ①和④ optimization.

本节开始讲蒙特卡罗

目标：计算 $E_\theta[f(x)] = \int_{\mathcal{X}} f(x)p(x)\mu(dx)$

$$= \begin{cases} \sum_{x\in\mathcal{X}} f(x)\,p(x) & \text{discrete} \\ \int_{\mathcal{X}} f(x)p(x)d(x) & \text{contious} \end{cases}$$

计算量很大.

蒙特卡罗利用 the law of large number.

Law of Large Number (L.L.N)

$$x_1, \cdots, x_n \overset{iid}{\sim} d$$

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \xrightarrow{a.s.} \underbrace{E[f(x)]}_{\text{expection}}$$

almost surely converge.

given a distr., an expection is also fixed no mattere how to calculate.

Monte Carlo sample mean

$$\frac{1}{n}\sum_{i=1}^{n} f(x_i) \quad x_i \sim d.$$

How many sample is enough to give such approximation? use CLT.

Central Limit Theorem

denote $I_n(f) = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$

- $I_n(f)$ is also random variable.

- $E[I_n(f)] = E[\frac{1}{n}\sum_{i=1}^{n} f(x_i)]$

$$= \frac{1}{n}\sum_{i=1}^{n} E[f(x_i)] \quad (\because \text{i.i.d})$$

$$= E[f(x)]$$

$\therefore$ this estimation is unbiased.

- $\sqrt{n}\,(I_n(f) - E) \xrightarrow{d} N(0, \sigma_f^2)$

$$\sigma_f^2 = Var(f(x))$$

$$Var(I_n(f)) \sim \frac{\sigma_f^2}{n}$$

when $n \uparrow$ the var $\downarrow$, usually set a tolerance and choose n to satisfy the tolerance.

The most difficult part of MC is how to get the samples.

下面介绍 random sampling.

一般计算机语言都提供 rand 函数，目标是利用这个函数来产生新的分布.

如何评估 random number generator 的好坏?

run a generator long time, it will repeat.

Linear Congruential Generator (LCG)

1100101100 11 . . . . random bits

$$m = 2^{32} / 2^{64}$$

rand() function in C/C++/Java is in this way

Mersenne Twister (MT) 更好的 generator.

C++11 <random>

MATLAB/Numpy/Julia.

$$m = 2^{19937}$$

现在可以用 generator 做的:
- generate integer [a, b]
- uniform distribution [0, 1) real number
- normal distribution (randn)

下面看如何生成指定分布.

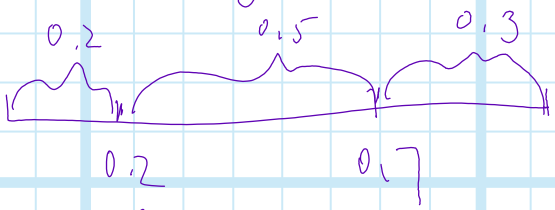
Discrete Distribution

categorical distribution:

$p = (p_1, \cdots, p_k) \quad p_1 + \cdots + p_k = 1$

例如 (0.2, 0.5, 0.3)

stick breaking



利用uniform 生成一个数看落在哪个区间.

① Linear search $O(k)$

遍历所有分段, 看到底落在哪里.

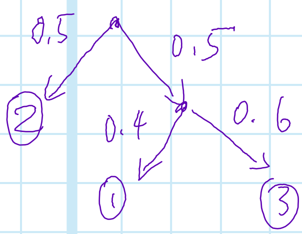
not efficient.

② Sorted search $O(k)$

加快速度: 把概率从大到小排序.

③ Binary Search.

按概率选择节点.



$O(\log_2 k)$

可以构造个 Hoffman Search Tree.
来让查找效率最优.

$$O(\text{Entropy}) \le O(\log_2 k)$$

④ Alias Table. 构造复杂，但使用很快
$O(1)$.

现在看如何从任意分布采样.
已知 $p(x)$.

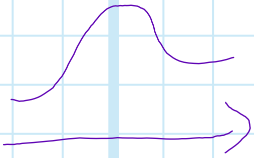
Transform Sampling

$u \sim U[0,1]$

$f(u) \sim P$ how can we get f?

2018.11.6

Sampling $p(x)$



Transform Sampling

$U \sim \text{Uniform}([0,1])$

$X \sim T(u)$ 通过变换T，从均匀分布生成任意分布.

$$X = F^{-1}(u)$$

F: cumulation distribution function (cdf)

$$F(x) = Pr(X \le x) = \int_{-\infty}^{x} p(v)\,dv.$$

以下证明X的分布就是我们需要的.

Proof:

$$X = F^{-1}(u)$$

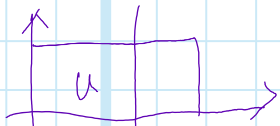
$$Pr(X \le x) = F(x)$$

$$\Updownarrow$$

$$Pr(F^{-1}(u) \le x) = F(x)$$

$$Pr(u \le F(x)) = F(x)$$

uniform 即 $Pr(u \le x) = x$



$Pr(u \le 0.6) = 0.6$

例：Exponential distribution $P(x)=\exp(-x)$

$F(x)=1-\exp(-x)$

$y=1-e^{-x} \Rightarrow x=-\log(1-y)$

$\therefore F^{-1}(x)=-\log(1-x)$

$u \sim U[0,1] \quad x=-\log(1-x)$

实际上 $u \sim U[0,1] \Leftrightarrow 1-u \sim U[0,1]$

$\therefore$ 也可以 $x=-\log u$.

上面实现了单一变量的采样，下面看多变量.

$N(\mu,\Sigma)$

$$P(x)=\frac{1}{\sqrt{(2\pi)^d|\Sigma|}}\exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$$

cdf 很难求. 可以从单变量的正态分布采样

$u \sim N(0,1)$

$$\begin{pmatrix} u_1 \\ \vdots \\ u_d \end{pmatrix} \sim N(0,I)$$

正态分布中，若 $x \sim N(\mu, \Sigma)$

$$x + b \sim N(\mu + b, \Sigma)$$

$$Ax \sim N(A\mu, A\Sigma A^T)$$

∴ 生成步骤：

① $\begin{pmatrix} u_1 \\ \vdots \\ u_d \end{pmatrix} \sim N(0, I)$

Cholesky Decomposition ✓

② Get A, s.t. $AA^T = \Sigma$

$$A \cdot u \sim N(0, AA^T) = N(0, \Sigma)$$

③ $Au + \mu \sim N(\mu, \Sigma)$

上面仅适于分布的性质.
下面看 Rejection Sampling

Rejection Sampling

target distribution $P(x)$,

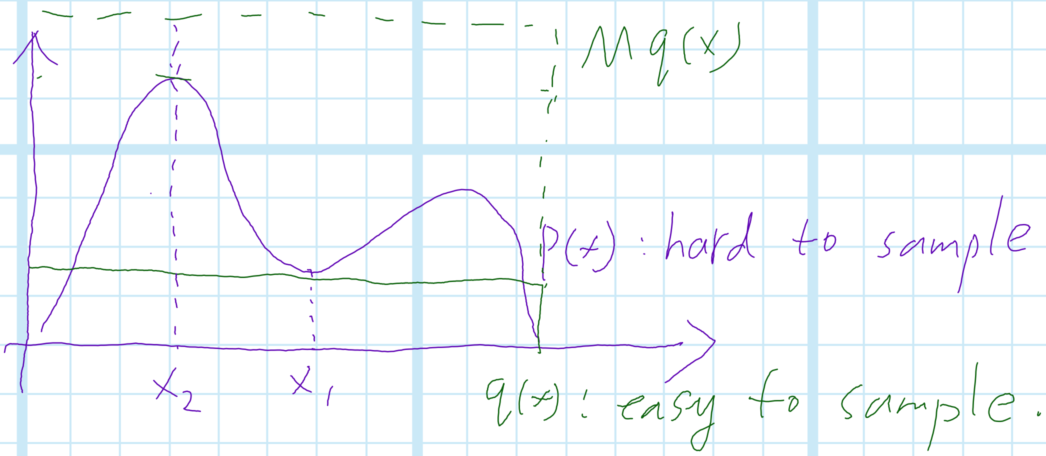
proposal distribution $q(x)$ (easy to sample)

Steps:

① $x \sim q$

② Accept x with $\dfrac{P(x)}{M \cdot q(x)}$

M: s.t. $Mq(x) \geq P(x)$



$a(x_1) = \dfrac{p(x_1)}{Mq(x_1)}$ 很小.

$a(x_2) = \dfrac{p(x_2)}{Mq(x_2)}$ 几乎等于 1

Auxiliary Variable.

① $x \sim q$

② $u|x \sim \text{Bernoulli}\left(\dfrac{p(x)}{Mq(x)}\right)$

$q(x) \cdot a(u|x)$

$p(x|u=1)$ 因为只有 $u=1$ 时，才接收，∴这个条件概率才是 x 的分布.

下面证明 $p(x|u=1) = p(x)$.

$$p(x|u=1) \propto q(x)\cdot \alpha(u=1|x)$$
$$= q(x)\cdot \frac{p(x)}{M q(x)}$$
$$= \frac{p(x)}{M}$$

优点：

① simple，只要知道 $p(x)$ 就能采样.

缺点：

M 可能会特别大.

这样接收率就特别低

Overall acceptance rate

$$P_r(u=1) = \int q(x)\cdot\frac{p(x)}{M q(x)}\, d(x)$$
$$= \int \frac{p(x)}{M}\, d(x) = \frac{1}{M}$$

∴接收率为 $\frac{1}{M}$

为了减小 M，可以找一个和 $p(x)$ 很接近的分布 $q(x)$.

Importance Sampling

$$E[f(x)] \approx \frac{1}{n}\sum_{i=1}^{n} f(x_i) \quad \text{with } x_i \sim p$$

difficult

proposal distribution q

$$E_p[f(x)] = \int p(x)f(x)dx = \int f(x)\frac{p(x)}{q(x)}\cdot q(x)\,dx$$

$$= E_q\left[f(x)\frac{p(x)}{q(x)}\right]$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)w(x_i) \quad \text{with } x_i \sim q$$

$w(x_i) = \dfrac{p(x_i)}{q(x_i)}$ is called importance weight

$w(x_i)$ 有时也很难计算. 例如:

$$p(x) = \frac{1}{Z_p}\exp(A(x))$$

$$q(x) = \frac{1}{Z_q}\exp(B(x))$$

Z_p, Z_q 很难计算.

只能计算 $\frac{\tilde{p}(x)}{\tilde{q}(x)} = \exp(A(x) - B(x))$

只能算出一个 scale，不能算出具体值。

self-normalized　Z_p，Z_q 这些 normalize 项不好计算，可以直接对 w 进行 normalize.

计算 $\tilde{w}(x_i) = \frac{\tilde{p}(x_i)}{\tilde{q}(x_i)}$

$$w(x_i) = \frac{\tilde{w}_i}{\sum_{j=1}^{n} \tilde{w}_j}$$　可以证明是无偏估计

和 Rejection sampling 类似，也存在相应问题。

会采到很多 weight 很小的样本，会不稳定.

下面介绍更有效的方式 MCMC

(实际上 GAN 也是一种 sampling，generator 采样，discriminator 来拒绝.)

MCMC (Markov Chain Monte Carlo)

- Markov Chain

$x_0, x_1, \cdots, x_T, \cdots$

$$p(x_t \mid x_{t-1}) = p(x_t \mid x_{t-1}, x_{t-2}, \cdots)$$

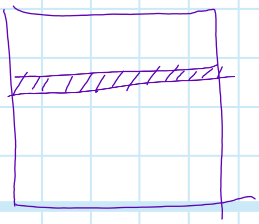
$$\mathrm{Markov}(\pi_0, P)$$

π_0: initial distribution

P: transition probability matrix

如果知道当前状态 x

$x \to x'$ $\quad p(x' \mid x)$



P

如果知道的是当前状态的概率分布 μ.

$$p(x') = \sum_x \mu(x)\, p(x' \mid x)$$

$$= \sum_x \mu(x) P(x, x')$$

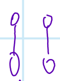
$$\mu' = \mu P$$

如果 $\mu = \mu P$ 则称 μ 为 invariant distribution

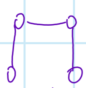
if P is irreducible & aperiodic

there exists a unique μ

s.t. $\mu = \mu P$



reducible



irreducible

irreducible & aperiodic 叫做 ergodic

MCMC idea:

target distribution μ hard to sample.

$\Downarrow$

construct a MC with P

s.t. $\mu = \mu P$

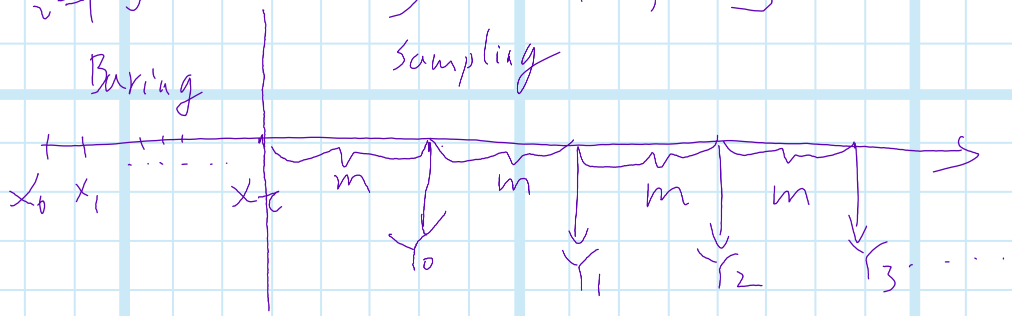
$$x_0 \to x_1 \to \cdots \to \underbrace{X_{100000}}_{} \to X \to X \cdots \quad \sim \mu$$

target distribution π.

ergodic Markov chain P s.t. $\pi = \pi P$

Ergodic Threom:

$$\frac{1}{n}\sum_{i=1}^{n} f(\tau + i \cdot m) \to E_\pi[f(x)]$$



τ: burning time

开始时的分布可能还不是μ，运行一段时间让分布接近μ.

m：一般采样都是假设的样本间相互独立，如果采了1个马上采1个，那么这样个样本相关性太强。

m used to reduce dependence.

下面看如何构造 P.

P: $\pi P = \pi$

$$\sum \pi(x) P(x,y) = \pi(y) \quad \forall y$$

Detail Balance $\Uparrow$ $\not\Downarrow$

$$\pi(x) P(x,y) = \pi(y) P(y,x) \quad \forall x,y$$

这个式子更简单，没有求和，积分（更严）

Proof: $\sum_x \pi(x) P(x,y) = \sum_y \pi(y) P(x,y)$

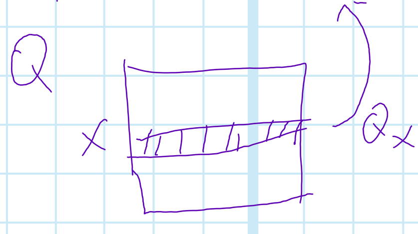
$$= \sum_x \pi(y) p(x|y) = \pi(y) \overbrace{\sum_x p(x|y)}^{=1} = \pi(y)$$

Metropolis – Hasting (M-H algorithm)

构造 MC 的算法。 类似于 reject sampling，先由 proposal kernal 产生一个样本，然后来拒绝。

1. $Q(x,y)$ given x, next step $y \sim Q(x,y)$

proposal kernal.



2. Every step:

1) $y \sim Q_x$ x是当前状态

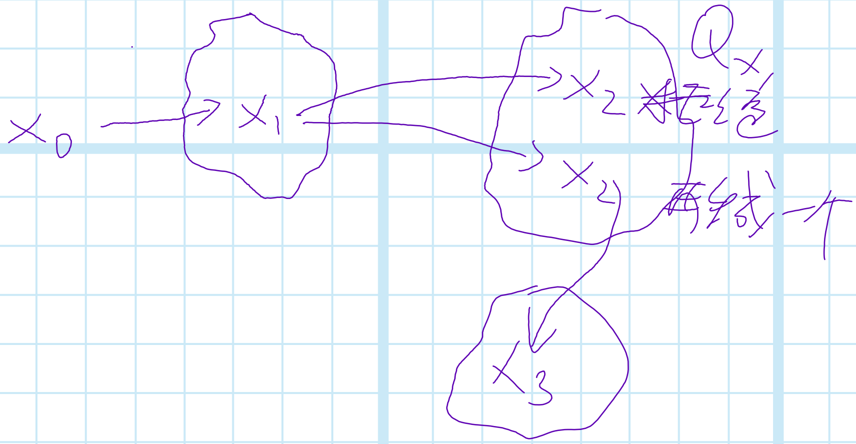
2) accept y with chance $a(x,y) = \min\{r(x,y), 1\}$

where $r(x,y) = \dfrac{\pi(y)\, Q_y(x)}{\pi(x)\, Q_x(y)}$

$$= \frac{\pi(y)\, Q(y \to x)}{\pi(x)\, Q(x \to y)}$$

$$= \frac{\frac{1}{Z} h(y)\, Q(y \to x)}{\frac{1}{Z} h(x)\, Q(x \to y)}$$

$$= \frac{h(y)\, Q(y \to x)}{h(x)\, Q(x \to y)}$$



下面分析 correctness 和 efficiency

— correctness

— efficiency $\begin{cases} \text{acceptance (high)} \\ \text{mixing rate} \\ \text{(convergence rate)}\ |\mu_t - \pi| \end{cases}$ trade off

如果样本都在一起，

conservative → { high acceptance
low mixing rate.

如果样本采得跳得远.

aggressive → { low acceptance
accelerate mixing rate.

回顾: 2018.11.19

Markov Chain Monte Carlo

target distribution π 很难采

构造 Markov chain P: $\pi P = \pi$

如何构造 P?

通过 Detailed Balance

$$\pi(x) P(x,y) = \pi(y) P(y,x)$$

Metropolis-Hasting Alg.

Q: proposal kernal

$$Q(\underset{\substack{\uparrow \\ \text{current}}}{x}, \underset{\substack{\uparrow \\ \text{next}}}{y})$$

give proposal for next based on current.

$$x \Rightarrow y \sim Q(x,y)$$
$$q_x(y)$$

acceptance rate $a(x,y) = \min\left\{ \dfrac{\pi(y) q_y(x)}{\pi(x) q_x(y)}, 1 \right\}$

下面证明这个 $a(x,y)$ 满足 Detailed Balance.

proof: proposal * acceptance ratio:

$$\pi(x)\,Q(x,y)\min\left\{\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)},1\right\}$$

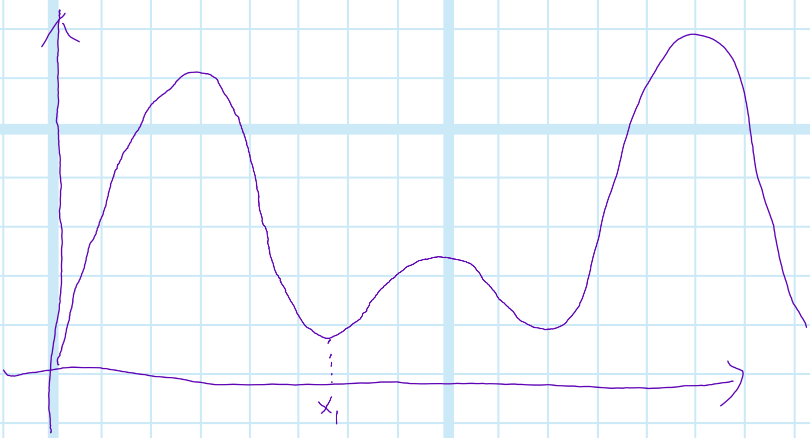
$$=\min\left\{\pi(x)Q(x,y)\cdot\frac{\pi(y)Q(y,x)}{\pi(x)Q(x,y)},\ \pi(x)Q(x,y)\right\}$$

$$=\min\left\{\pi(y)Q(y,x),\ \pi(x)Q(x,y)\right\}$$

$$\frac{\pi(y)\,q_y(x)}{\pi(x)\,q_x(y)}$$ 很难计算, ∵ $\pi(y)=\underset{\substack{\uparrow\\\text{normalize term 很难}}}{\frac{1}{Z}}h(y)$

$$=\frac{h(x)\cdot q_y(x)}{h(x)\,q_x(y)}$$ 去掉 normalize term.

例



$$X_{t+1}\leftarrow X_t+\varepsilon_t \qquad \varepsilon_t\sim N(0,\sigma^2)$$

也是 kernel Q

$$q_x(y)\sim N(x,\sigma^2)$$

例如，开始的 x，如图，概率很小。

用这个 kernel：$q_x(y) \sim N(x, \sigma^2)$

注意到这是一个对称的 kernel 即：

q 只与 x, y 的距离有关。

symmetric kernel

$$q_x(y) = q_y(x)$$

$$\therefore \frac{\pi(y)\, q_y(x)}{\pi(x)\, q_x(y)} = \frac{\pi(y)}{\pi(x)}$$

$\therefore$ 跳到概率高的地方，其接收率就高，

跳到概率低的地方接收率低。

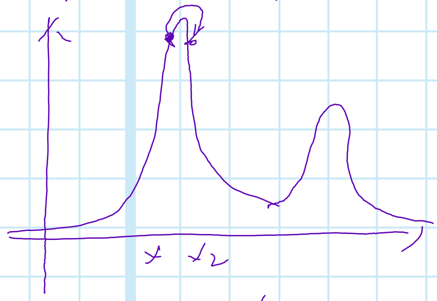
那么如何设计一个好的 proposal kernel.

要求：

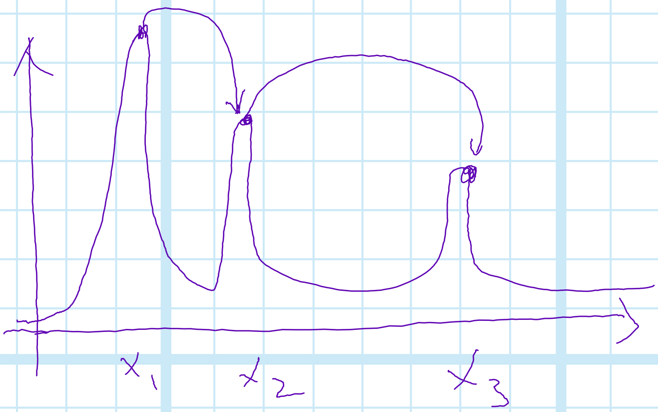
— high acceptance ratio

— explore the space efficient.

接收率低



不能很好地探索空间。



好的 proposal: Jump from peak to peak.

Gibbs Sampling

$p(x_1, \cdots, x_n)$ 多变量采样,

每次只改变一维.

$$s_1 = (x_1, x_2, x_3)$$

$$s_2 = (x_1, x_2', x_3) \qquad x_2' \sim p(x_2 | x_1, x_3)$$

下面来看正确性:

$p(x, y)$

$$s_1 = (x, y)$$

$$s_2 = (x', y) \qquad Q(s \to s') \propto p(x' | y)$$

$$r(s \to s') = \frac{p(x', y)}{p(x, y)} \cdot \frac{p(x | y)}{p(x' | y)}$$

$$= \frac{p(y)\, p(x' | y)}{p(y)\, p(x | y)} \frac{p(x | y)}{p(x' | y)} = 1$$

$$a(s \to s') = \min\{r(s \to s'), 1\}$$

∵ Gibbs sampling 的接收率恒等于1

效率很高。

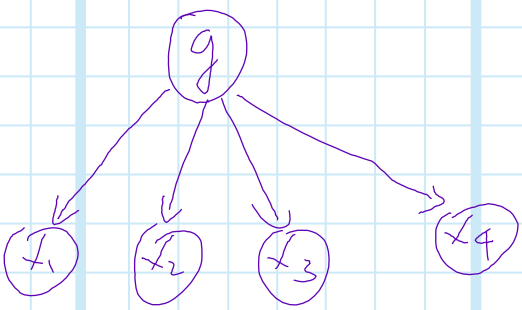
如何选变化的坐标。

cycling scheme

- fix scheme $1 \to 2 \to 3 \to 4 \to 1 \to 2 \to 3 \to \cdots$
- random cycle. $1 \to 3 \to 4 \to 2 \to 4 \to 3 \to 1 \cdots$

Gibbs sampling 在参数之间很大的时候不是很有效，很多算法发明出来提高 efficiency.

Collapsed Gibbs Sampling



$$g \sim N(\mu_0, \sigma_0^2)$$

$$x_i \sim N(g, \varepsilon^2) \qquad \varepsilon << \sigma$$

如果用 Gibbs Sampling

(x_i 间相互独立)

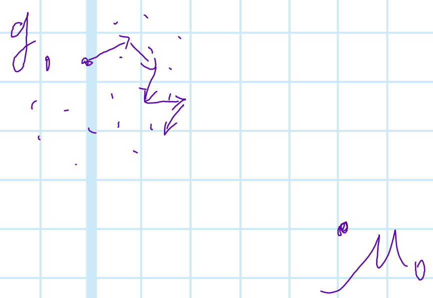
$$x_i \mid g \sim N(g, \varepsilon^2) \quad \forall i = 1, 2, 3, 4.$$

$$g \mid x_1, x_2, x_3, x_4 \sim N(\mu, \sigma'^2)$$

$$\mu' = \frac{\frac{1}{\sigma_0^2} \cdot \mu_0 + \sum_{i=1}^{n} \frac{1}{\varepsilon^2} x_i}{\frac{1}{\sigma_0^2} + \sum_{i=1}^{n} \frac{1}{\varepsilon^2}}$$

$$(\sigma'^2)^{-1} = \frac{1}{\sigma_0^2} + \sum_{i=1}^{h} \frac{1}{\varepsilon^2} \gg \frac{1}{\sigma_0^2}$$

$$\sigma'^2 \ll \sigma_0^2$$



mutual loching：当前结点 conditioned on other nodes

∵ mutual loching，g_0 到 μ_0 速度很慢.

解决方法：

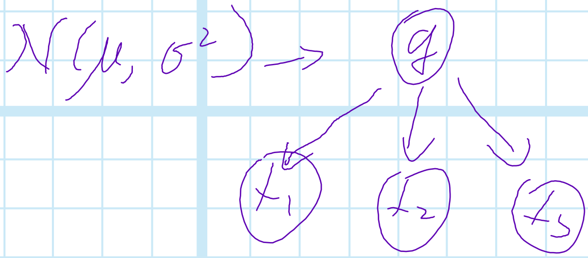
marginalize out

$$g_0 \to g_1 \to g_2 \cdots \to g_t \longrightarrow g_{t+1}$$

$x_1 \quad x_2 \quad x_3 \quad x_4$

2018.11.23

Collapsed Gibbs Sampling



迭代时，变量相互影响，move 很少

当采样 g 时，不受到 x_1, x_2, x_3 影响.

Rao-Blackwell Theorem

① $P(X, Y)$

$(x_1, y_1), (x_2, y_2) \cdots, (x_n, y_n) \sim P$

$E[h(x, y)] \approx \frac{1}{n}\sum_{i=1}^{n} h(x_i, y_i) \rightarrow \bar{h}_1$

以上是标准的 MC 过程.

② $P(X, Y)$

$\downarrow$ margional out Y

$P(X)$

$x_1, x_2, \cdots, x_n \sim P(X)$

$E[h(X, Y)] \approx \frac{1}{n}\sum_{i=1}^{n} E_{Y|x_i}[h(x_i, Y)]$

$\rightarrow \bar{h}_2$

正确性:

$$E[h(X,Y)] = \int_{X\times Y} h(x,y)\,p(x,y)\,\mu(dx,dy)$$

$$= \int_X\int_Y h(x,y)\,p(y|x)\,p(x)\,\mu(dx)\,\mu(dy)$$

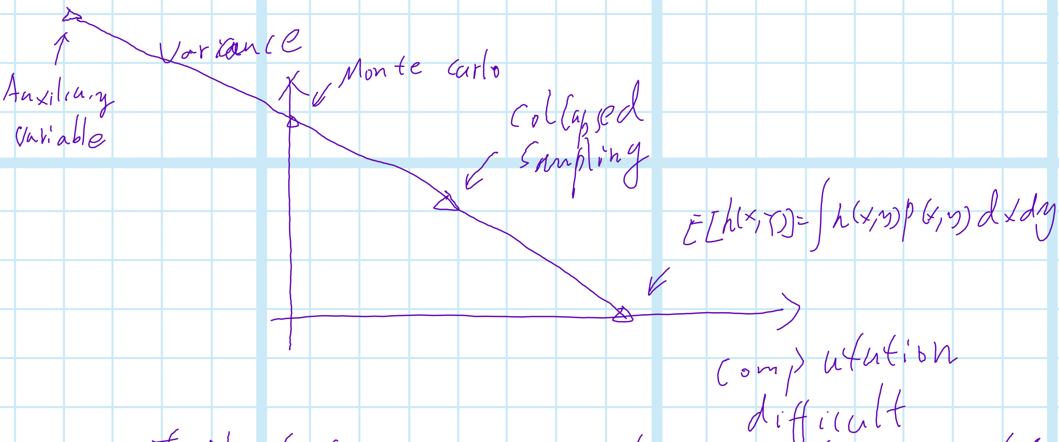
$$= \int_X \int_Y h(x,y)\,p(y|x)\,dy\;p(x)\,dx$$

$$= \int_X \underbrace{E_{Y|X}[h(x,Y)]}_{f(x)}\,p(x)\,dx$$

$$\approx \frac{1}{n}\sum_{i=1}^{n} f(x_i)$$

Rao-Blackwell Theoms

$$Var(\bar{h}_1) \geq Var(\bar{h}_2)$$



MC 直接采样，不用从复杂的积分计算.

直接计算没有 variance.

Collapsed Sampling 是一个折衷的方式.

Collapsed Sampling 先积分积掉一些变量，再来 Sampling，这种方式叫做 Rao-Blackdization

Collapsed Sampling 积掉了一些变量，一种相反的方法是引入更多的辅助变量。有些情况下，需要让 Sampling 比 MC 更简单

Sampling with Auxiliary variable.

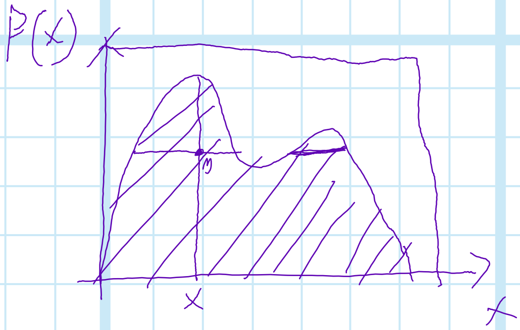
target: $P(x)$

$\Downarrow$

$P(x, u)$

$(x_1, u_1), \ldots, (x_n, u_n)$

① Slice Sampling



每次不从 x 采样，而是从二维的阴影部分采样

可以用 Gibbs Sampling 来从阴影部分采样

$y|x \sim U[0, f(x)]$

$x|y \sim U[\{x: f(x) \geq y\}]$

Property:

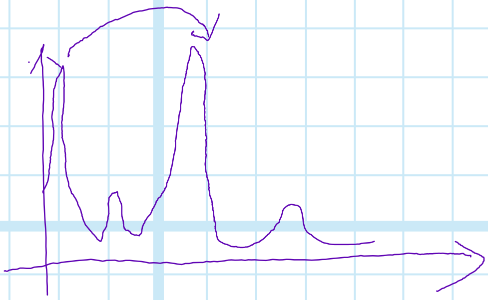
— very efficient

对于 Gibbs Sampling, 所有的样本都被接收.

— $\{x: f(x) \geq y\}$ 不好计算

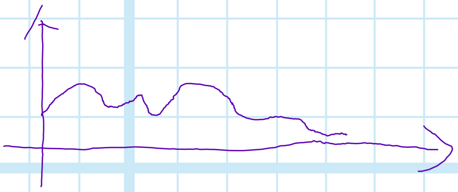
② Tempering

introduce temperature to certain distribution



目标: Jump from peak to peak.

很困难, 可以考虑一个 smooth 的分布



引入 temperage

$$p(x) = \frac{1}{Z} \exp(-E(x))$$

Gibbs Distribution

$$P_\alpha(x) = \frac{1}{Z_T} \exp\left(-\frac{E(x)}{T_\alpha}\right)$$ T_α — temperature.

温度越高，越 smooth

低 peaky

$T \gg 1$: much easier to sample

$T = 1$: Our target distribution

Basic Idea

introduce τ as an auxiliary varible.

Simulated Tempering

$$p(x, k) = \frac{\pi_k}{Z_k} \exp\left(-\frac{E(x)}{T_k}\right)$$

构造了一个不同温度 $\{T_1, \cdots, T_n\}$ 的分布

π_i 是比例

	π		
$P_1 = \frac{1}{Z_1} \exp(-E(x))$	0.3	π_1	a_1
$P_2 = \frac{1}{Z_2} \exp\left(-\frac{E(x)}{T_2}\right)$	0.4	π_2	a_2
$P_3 = \frac{1}{Z_3} \exp\left(-\frac{E(x)}{T_3}\right)$	0.3	π_3	a_3

MCMC

一、transition Proposal

$$(x,k) \longrightarrow (x',k) \quad x' \sim q_k(x \to x')$$

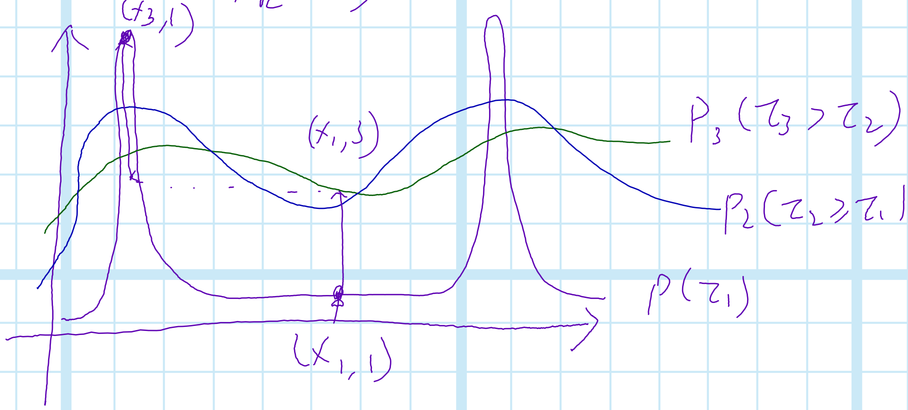
一、Temperature switch

$$q_T(k \to k')$$

$$a(k,k'|x) = \min\left\{1, \frac{p(x,k')}{p(x,k)}\right\}$$ 温度转移的接收率。

另一种方法

$$p(k|x) \propto \pi_k \cdot p_k(x)$$



$\pi_1 = \frac{1}{3}$

$\pi_2 = \frac{1}{3}$

$\pi_3 = \frac{1}{3}$

在温度高的地方，容易explore space，在某些另一个分布大的时候就切换分布进行采样。

二、高温分布的作用是 bridge for large move.

低温分布适合采样，高温分布适合移动。

Parallel Tempering (适合并行计算)

Crop/ica switch

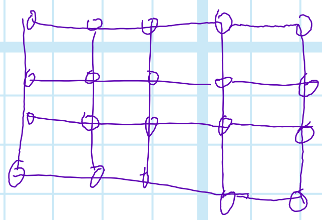
target $\rightarrow P_1 \rightarrow x_1^{(1)}, \cdots, x_n^{(1)}$

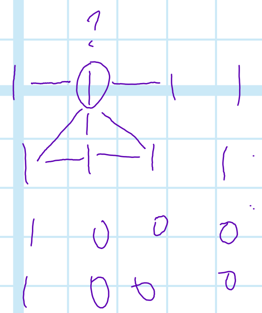
$P_2 \rightarrow x_1^{(2)}, \cdots x_n^{(2)}$

$P_3 \rightarrow x_1^{(3)}, \cdots, x_n^{(3)}$

$$(x_{t_1}^{(1)}, x_{t_2}^{(2)}, x_{t_3}^{(3)}) \rightarrow (x_t^{(1)}, x_t^{(2)}, t_t^{(3)})$$

Suendsen - Wang

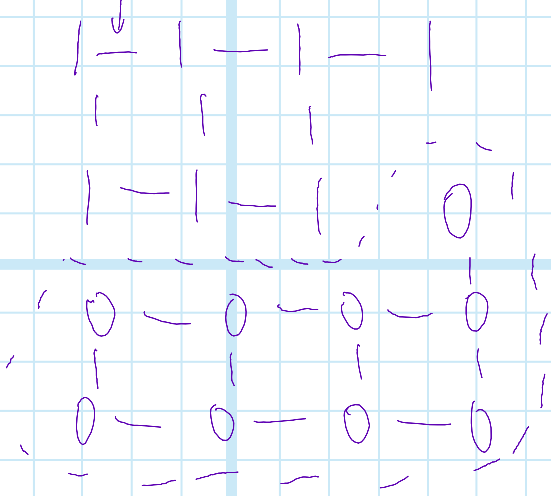




Gibbs Sampling

$$p(x) \propto \exp\left(-\sum_{i,j} w_{ij} \mathbb{1}(x_i = x_j)\right)$$

g_{ij}: =0时有边, =0时无边.



Basic: cast node into groups and update each group together

$$p(x|\theta) = \frac{1}{Z}\prod_{(i,j)\in E} f_{ij}(x_i, x_j)$$

$$= \frac{1}{Z}\exp\Big(\sum_{i,j\in E}\theta_{ij} x_i x_j\Big)$$

u_{ij}: for each edge turn on or off

$$p'(x,u) = \frac{1}{Z'}\prod_{(i,j)\in E} g_{ij}(x_i, x_j, u_{ij})$$

$$g(x_i, x_j, u_{ij}) = \begin{cases} \exp(-\theta_{ij}) & u_{ij}=0 \\ 1(x_i=x_j)(e^{\theta_{ij}}-e^{-\theta_{ij}}) & u_{ij}=1 \end{cases}$$

$u_{ij}=1 \Rightarrow x_i=x_j$ (all variable in the same component should have the same value)

$\Updownarrow$

$x_i\neq x_j \Rightarrow u_{ij}=0$

$$g(x_i, x_j, u_{ij})$$

$$= f_{ij}(x_i, x_j)\, q_{ij}(u_{ij}|x_i, x_j)$$

when $x_i \neq x_j$ $u_{ij}=0$

when $x_i = x_j$ $p(u_{ij}=0|x_i=x_j)=\exp(-\theta_{ij})$

算法步骤

1. Clustering step

$u|x \sim p(u_{ij}=0|x_i=x_j)=\exp(-\theta_{ij})$

2. mapping $x\sim p(x|u)$

$$p'(y, u) = \prod_{(i,j)\in E} g(x_i, x_j, u_{ij})$$

$$= \prod_{(i,j)\in E} f_{ij}(x_i, x_j)\, q(u_{ij} \mid x_i, x_j)$$

$$= \left(\prod_{(i,j)\in E} f_{ij}(x_i, y_j)\right) \cdot \prod_{(i,j)\in E} q(u_{ij} \mid x_i, y_j)$$

$$\xrightarrow{\text{marginalize}} p(x)$$

总结

Common ideas of auxiliary variables

$$p(x) \xrightarrow{\text{aux}} p(x, \mu)$$

$\downarrow$ Gibbs Sampling

Given μ: $p(x \mid \mu)$

Given x: $p(\mu \mid x)$.

	Slice Sampling	Simulated Tempering	Swendson-Wang
Aux var μ		T_k: Temperature	use u as connections u_{ij} $(i,j)\in E$
Gibbs Sampling: $x \sim p(x \mid \mu)$	$x \sim U[\{x : f(x) \geq y\}]$	$x \sim p(x \mid T)$	Given u, update x by components
$\mu \sim p(\mu \mid x)$	$u \sim U[0, f(x)]$	$p(T \mid x) \propto \pi(T)\, p(x \mid T)$	$u_{ij} \sim q(u_{ij} \mid x_i, x_j)$

可以看出 $\mu \sim p(\mu|x)$ 这一步的作用是 switch the environment

$x \sim p(x|\mu)$ 作用是 sampling

Variational Auto-Encoder

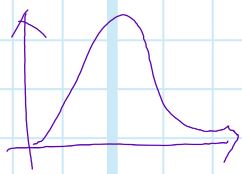
Generative Adversarial Learning.

是连接的方法.

Fundamental Problem: How to character a distribution

- Descriptive way

  density function $p(x)$



- Construtive way (Generative way)

  focus on how to get samples

  It's actually a sampler.

△ Transform sampler: 寻求分布式的采样

$$u \overset{T}{\sim} T(u) \sim p$$

△ Rejection Sampling

$$\begin{tikzcd} & u \arrow[r] & a \arrow[r] \arrow[d, leftrightarrow] & u' \\ q \arrow[ur] & & V & \end{tikzcd}$$

采一个 V ($v \sim [0,1]$) 如果比 a 大，就接受。

△ MCMC

EM

$$p(X, Y | \theta)$$

observed	latent	model parameter

$\tilde{\theta} = \underset{\theta}{\arg\max}\, p(X|\theta)$ Y被 marglize 掉

很难作积分，算不出!

$$p(X|\theta) = E_{q(Y)}[p(X,Y)|\theta)]$$

$\sim p(Y|X;\theta)$ ← difficult

Variational E-M

$q = q_{y_1} \cdot q_{y_2} \cdots$ 假设 q 可分解.

Mean field Approximation

这个估计可以直接用 NN 来做

$p(Y|X;\theta)$ 用一个神经网络

也可以用 law of large Number

GAN:

D → Descriptive way

G → Constructive way

2018.11.26

Revisit

$$P_\theta(x, z)$$

observe (x), latent (z)

$$L(\theta) = \frac{1}{n}\sum_{i=1}^{n} \log P_\theta(x_i)$$ ← marginal density hard to compute. consider lower bound

$$E_q[p(x,z)] = E_q[\log P(x) + \log P(z|x)]$$

$$= \log P_\theta(x) + E_q[\log P(z|x)]$$

① $\log P_\theta(x) = E_q[\log P(x,z)] - E_q[\log P(z|x)]$

② $L_x(\theta, q) = E_q[\log P(x,z)] - E_q[\log q(z)]$

∴ ① − ② :

$$\log P_\theta(x) - L_x(\theta, q)$$

$$= E_q[\log q(z) - \log P(z|x)]$$

$$= E_q\left[\frac{\log q(z)}{\log P(z|x)}\right] = D_{KL}[q(z) \| P(z|x)] \geq 0$$

那么 q(z) 如何建模.

— mean field approximation

$$q(z_1, z_2) = q_1(z_1)\, q_2(z_2)$$ 直接假设形式分解

— neural network

VAE

目标：建模 $L_x(\theta, q) = E_q[\log p(x,z)] - E_q[\log q(z)]$

假设 q 的参数为 ϕ

$$L(\theta, \phi) = \underbrace{E_\phi[\log p_\theta(x,z)]}_{\Downarrow \text{ monte carlo}} - E_{q_\phi}[\log q_\phi(z)]$$

sampler g $\quad \varepsilon \sim p_0$ ← 某个特殊分布 $\quad g(\underset{\text{randomness}}{\varepsilon}, \underset{\text{condition}}{x}) \sim q_x(\cdot)$

$$E_q[f(z)] \approx \frac{1}{n}\sum_{i=1}^{n} f(z_i) \qquad z_i \sim q_\phi(\cdot \mid x)$$

$$= \frac{1}{n}\sum_{i=1}^{n} f(g_\phi(\varepsilon_i, x))$$

代入目标得：

$$L(\theta, \phi)$$

$$= \frac{1}{L}\sum_{l=1}^{L}\left\{\log p_\theta(x, z^l) - \log q_\phi(z^l \mid x)\right\}$$

where $z^l = \underset{\text{neural network}}{g_\phi}(\varepsilon^l, x) \qquad \varepsilon^l \sim p_0$

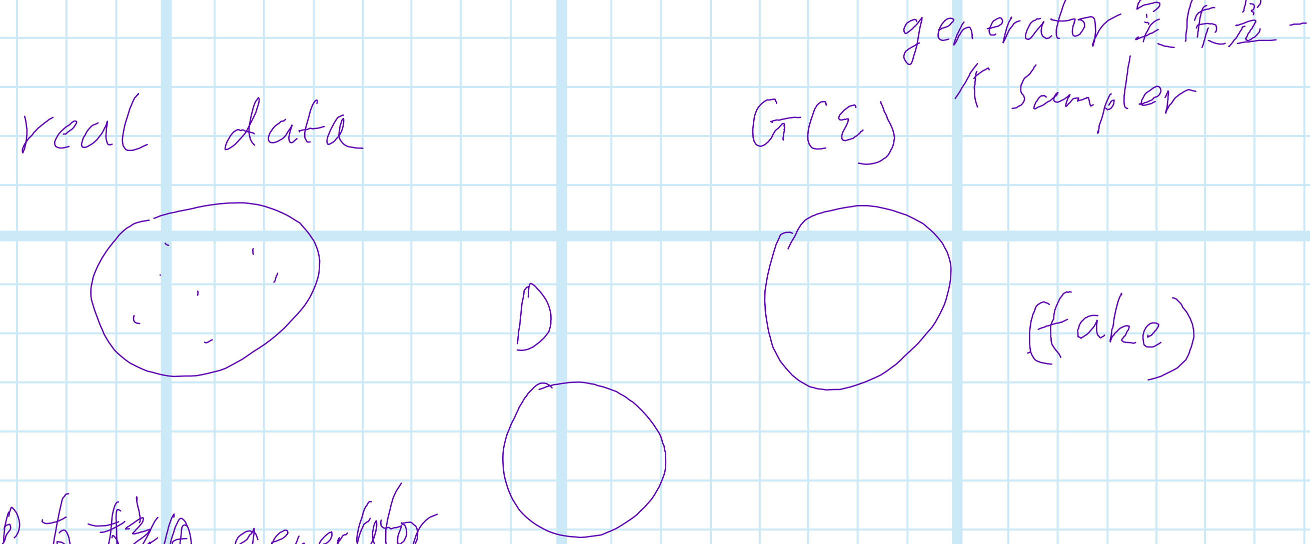
$P_\theta(x,z)$ 是对于分布的 descriptive 描述方式.

而 $q(z,x)$ 则是 generative way.

VAE 同时使用了 descriptive 和 generative way

GAN

GAN 的 generator 和 descriminator 是一种直观的解释.



现在想直接用 generator 来生成样本, 需要一个 descriptive way of real data, 这其实就是 Descriminator

$$\min_G \max_D E_{x\sim p_{data}}[\log D(\underset{x_{real}}{\underset{\uparrow}{x}})] + E_{x\sim p(z)}[\log(1-D(\underbrace{G(z)}_{x_{fake}}))]$$

$$D: \max \; \log D(x_{real}) + \log(1-D(x_{fake}))$$

$$G: \min \; \log(1-D(G(z)))$$

discriminative training of a generation

$$D^*(x) = \frac{P_{data}(x)}{P_{data}(x)+P_g(x)}$$

capture the density in a transformed way

$$= \frac{1}{1+\frac{P_g(x)}{P_{data}(x)}} = \frac{1}{1+\left(\frac{P_{data}(x)}{P_g(x)}\right)^{-1}}$$

real data is non-parametric distribution

$D^*(x)$ convert it to a parametric way.

因子:

$$\max P_\theta(G(\varepsilon_i))$$

$\log(1-D(G(\varepsilon)))$ transformed way

NCE → estimation 提供了GAN的理论基础.

NCE 没有 generator, 而是用 fake example.

G-M 时: inference

2018.11.30

estimator

两个指标很重要.

consistency：给足够多的样本，能收敛.

efficiency：收敛速度.