概率图模型理论及应用

Theory and Applications of Probabilistic Graphical Models

(Lesson 4 - mlEstimate)

欧智坚

清华大学电子工程系

Addr: 罗姆楼 6-104

Tel: 62796193

Email: ozj@tsinghua.edu.cn

- 希望课程讲述不要太粗略就好。由于课时有限，希望尽可能学习到概率图模型的精髓。
- 希望在课堂上重点讲解图模型的基本概念和背后思想，对于公式繁冗的推导过程可以只在课上进行一些提示，并推荐一些参考资料供自学，不要占用过多的课堂时间。
- 最好理论部分能在直接有step by step 的推导指导，应用部分有hand by hand 的工程或作业指导。
- 求老师讲的详细一点，因为发现上课真的听不懂呀。很多概念稍微解释一下吧。
- 感觉头两次课听不太明白，老师讲的有点像期末复习过一遍的感觉。希望老师能降低听课的先修知识门槛，讲的具体一点，多举一些例子。谢谢

- 数学推导可以简略，偏工程应用。
- 可以重点讲一下系统和应用的东西。

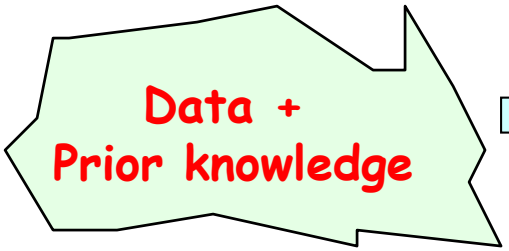
- 希望老师能够提前将期末 project 的选题发布，这样同学们能够较为充分地准备。

- 建议：给个5分钟让做笔记的同学自己说说总结或者问题，下面的同学可以提问之类的。

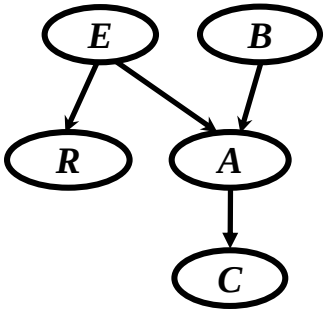
课程章节

- 第一章 引言（1）
- 第二章 图模型的表示理论（2）
  - **Semantics (DGM, UGM)**
  - **HMM, CRF**
- 第三章 图模型的推理理论（6）
  - 精确推理：variable-elimination，cluster-tree，triangulate
  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第四章 图模型的学习理论（3）
  - 参数学习：**maxlikelihoodEstimate**，RFLearning, BayesEstimate
  - 结构学习：StructureLearning
- 第五章 一个综合例子（1）

Learning



$p(X \mid g, \theta)$



E	B	$p(A \mid E, B)$	
e	b	**0.9**	**0.1**
e	$\bar{b}$	**0.7**	**0.3**
$\bar{e}$	b	**0.8**	**0.2**
$\bar{e}$	$\bar{b}$	**0.99**	**0.01**

Structure learning /model selection:

Learning g from data

Parameter learning /parameter estimation:

learning θ from data

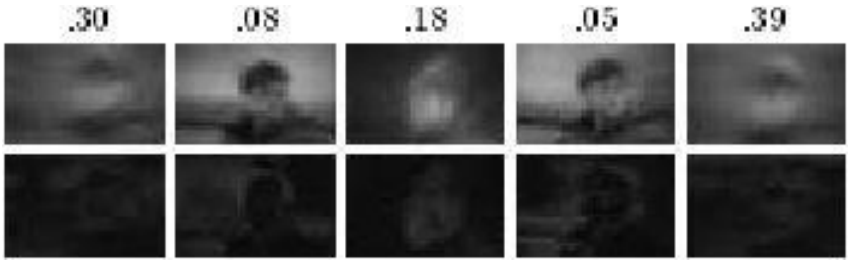
Data

- 总体分布 $p(x_{1:N} \mid g, \theta)$ 的一个个样本构成样本集/观测数据 $D = (x[1], \ldots, x[M])$
  - 一个样本 $x_{1:N}[m] = (x_1[m], x_2[m], \ldots, x_N[m])$
  - 独立同分布采样（IID）：assume $x[1], \ldots, x[M]$ are Independent and Identically Distributed ~ $p(x \mid g, \theta)$.
- 目标：从 $D = (x[1], \ldots, x[m], \ldots, x[M])$ 中估计出 g, θ

总体分布：$p(x_1, x_2 \mid g, \theta)$，其中 $w_k = p(x_1 = k)$ $p(x_2 \mid x_1 = k) = N(x \mid \mu_k, \Sigma_k)$

头姿类别 $x_1 \in 1{:}K$

观测图像 $x_2 \in R^{44*28}$



参数：$\theta = \{w_k, \mu_k, \Sigma_k, k = 1, \cdots, K\}$

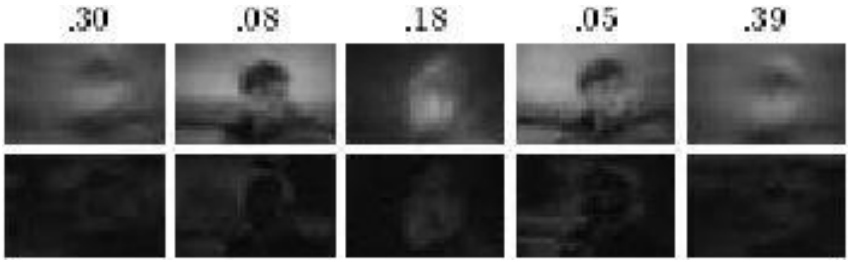
	x_1	x_2
$x[1]$	斜脸	
	侧脸	
	斜脸	
	斜脸	
$x[5]$	正脸	

Data—complete, incomplete

总体分布：$p(x_1, x_2 \mid g, \theta)$，其中 $w_k = p(x_1 = k)$ $p(x_2 \mid x_1 = k) = N(x \mid \mu_k, \Sigma_k)$

头姿类别 $x_1 \in 1{:}K$

观测图像 $x_2 \in R^{44*28}$



参数：$\theta = \{w_k, \mu_k, \Sigma_k, k = 1, \cdots, K\}$

	x_1	x_2
$x[1]$	斜脸	
	侧脸	
	斜脸	
	斜脸	
$x[5]$	正脸	

数据：400 幅图像及其头姿类别，D = (x[1],…, x[400])

- 完备数据：样本中各变量都赋值
- 不完备数据：

  latent/hidden variables：样本中某些变量的取值未知

The learning problem

参数学习　　结构学习

	Known structure		Unknown structure
Complete data	ML	Bayesian	
Incomplete data	ML	Bayesian	

DGMs, UGMs

Parameter learning

— ML (Known structure, complete data)

对单个分布的参数进行估计？

对一个贝叶斯网络的全体参数进行估计？

最大似然参数估计（MLE）

给定一个概率分布的参数表达式（parametric form）

记为 $p_\theta(x)$ 或 $p(x \mid \theta)$

从独立同分布样本集 $D = (x[1], \ldots, x[M])$ 中估计出参数 θ？

- $x \in \{1, 2, \ldots, K\}$ is discrete r.v.
- $\theta_k = p(x=k),\ 1 \le k \le K$, is the parameters, $\theta = \{\theta_k \mid 1 \le k \le K\}$

- x is Gauss r.v. $X \sim N(\mu, \Sigma)$
- $\theta = (\mu, \Sigma)$ is the parameters

最大似然参数估计（MLE）

给定一个概率分布的参数表达式（parametric form）

记为 $p_\theta(x)$ 或 $p(x \mid \theta)$

从独立同分布样本集 $D = (x[1], \ldots, x[M])$ 中估计出参数 θ？

- 将 θ 视为一个未知常数
- θ 的一个估计/猜测 $\hat{\theta}$ ：样本集的一个函数 $(x[1], \ldots, x[M])$

样本集 $x[1{:}M]$ 下，θ 的似然函数 $p\left(x[1:M] \mid \theta\right) = \prod_{m=1}^{M} p\left(x[m] \mid \theta\right)$

概率分布函数　似然函数

http://en.wikipedia.org/wiki/Likelihood_function

给定参数λ下，随机变量Y特定取值y的概率（密度）值 $p(y \mid \lambda)$ 视为

给定随机变量Y特定取值y下，参数λ的似然值

最大似然估计：使 似然函数 取最大 $\theta^{ML}\left(x[1:M]\right) = \arg\max_{\theta} p\left(x[1:M] \mid \theta\right)$

Multinomial distribution

- $x \in \{1, 2, \ldots, K\}$ is discrete r.v.
- $\theta_k = p(x=k),\ 1 \le k \le K$, is the parameters, $\theta = \{\theta_k \mid 1 \le k \le K\}$
- 观测到独立同分布样本集 $D = (x[1], \ldots, x[M])$
- 希望估计 θ ？

似然函数 $p(x[1:M] \mid \theta) = \prod_{m=1}^{M} p(x[m] \mid \theta) = \prod_{k=1}^{K} \theta_k^{N_k}$

N_k : 在样本集中 $x[m]=k$ 出现的次数

最大似然估计 $\theta_k^{ML} = \dfrac{N_k}{\sum_{l=1}^{K} N_l}$

$(N_1, \ldots, N_K)$ are sufficient statistics

Sufficient statistics

- 统计量：样本集 $D = (x[1], \ldots, x[M])$ 的某函数
- Neyman Factorization theorem

  一个统计量 $s(D)$, i.e., $s(x[1], \ldots, x[M])$ 是充分统计量 当且仅当 似然函数可以如下分解：

$$p(D|\theta) = g(\theta, s(D)) \cdot h(D)$$

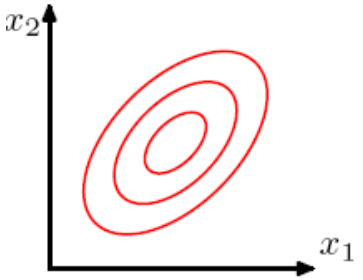
参数与样本的关联 完全通过充分统计量来体现

$$p(D|\theta) = \prod_{m=1}^{M} p(x[m]|\theta) = \prod_{k=1}^{K} \theta_k^{N_k}$$

N_k : 在样本集中 $x[m]=k$ 出现的次数

Gauss distribution

- x is Gauss r.v. $X \sim N(\mu, \Sigma)$
- $\theta = (\mu,\Sigma)$ is the parameters
- 观测到独立同分布样本集 $D = (x[1],\ldots, x[M])$
- 希望估计 θ ？



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

对数似然函数

$$\log p(x[1:M]\,|\,\mu,\Sigma) = \sum_{m=1}^{M} \log p(x[m]\,|\,\mu,\Sigma)$$

$$\max_{\mu,\Sigma} = -\frac{Md}{2}\log(2\pi) - \frac{M}{2}\left\{\log|\Sigma| + tr\left(\Sigma^{-1}\bar{\Sigma}\right) + (\bar{\mu}-\mu)\Sigma^{-1}(\bar{\mu}-\mu)^T\right\}$$

$$\begin{cases} \dfrac{\partial L}{\partial \mu} = M\cdot\Sigma^{-1}(\bar{\mu}-\mu) = 0 \\ \dfrac{\partial L}{\partial \Sigma^{-1}} = -\dfrac{M}{2}\cdot\left\{-\Sigma + \bar{\Sigma} + (\bar{\mu}-\mu)(\bar{\mu}-\mu)^T\right\} = 0 \end{cases} \Rightarrow \begin{cases} \bar{\mu} = \dfrac{1}{M}\sum_{m=1}^{M} x[m] \\ \bar{\Sigma} = \dfrac{1}{M}\sum_{m=1}^{M}(x[m]-\bar{\mu})(x[m]-\bar{\mu})^T \end{cases}$$

Learning parameters for BNs (complete data)

- 考虑贝叶斯网络 $X = \{X_1, X_2, \ldots, X_N\}$
  假设：各个条件分布 $p(x_1 \mid pa_1),\ p(x_2 \mid pa_2),\ \ldots,\ p(x_N \mid pa_N)$
  有各自表征参数 $\{\theta_1, \theta_2, \ldots, \theta_N\} = \theta$

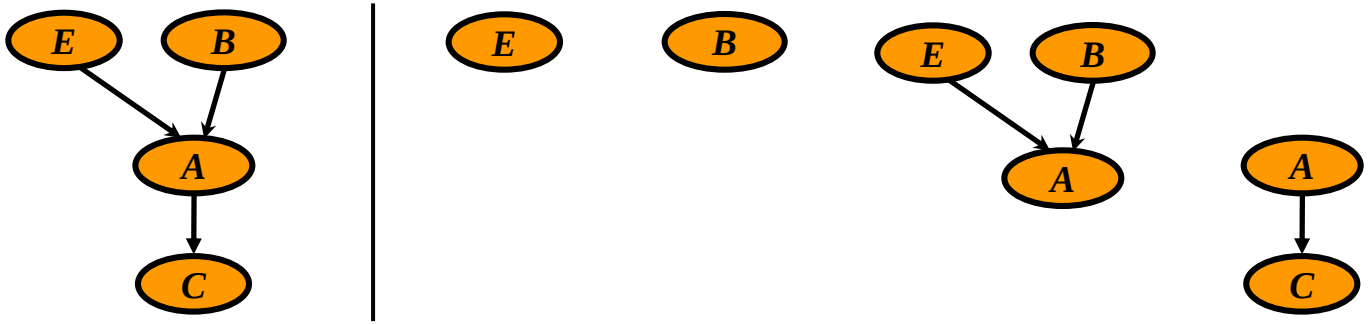
- IID样本集 $D = (x[1], \ldots, x[M])$ 下似然函数

$$\max_{\theta} p(D \mid \theta) = \prod_{m=1}^{M} p(x[m] \mid \theta) = \prod_{m=1}^{M}\prod_{n=1}^{N} p(x_n[m] \mid pa_n[m], \theta_n) = \prod_{n=1}^{N}\max_{\theta_n}\prod_{m=1}^{M} p(x_n[m] \mid pa_n[m], \theta_n)$$

  - 对每个条件分布 $p(x_n \mid pa_n)$ 分别估计其参数 θ_n

$$\hat{\theta}_n = \arg\max_{\theta_n} \prod_{m=1}^{M} p(x_n[m] \mid pa_n[m], \theta_n)$$

E, B, A, C
<1, 0, 0, 0>
<1, 1, 1, 1>
…
<1, 0, 1, 1>



$$\max_{\theta} p(D \mid \theta) = \prod_{m=1}^{M} p(E[m] \mid \theta_1)\, p(B[m] \mid \theta_2)\, p(A[m] \mid E[m], B[m], \theta_3)\, p(C[m] \mid A[m], \theta_4)$$

$$= \left\{\max_{\theta_1}\prod_{m=1}^{M} p(E[m] \mid \theta_1)\right\}\left\{\max_{\theta_2}\prod_{m=1}^{M} p(B[m] \mid \theta_2)\right\}\left\{\max_{\theta_3}\prod_{m=1}^{M} p(A[m] \mid E[m], B[m], \theta_3)\right\}\left\{\max_{\theta_4}\prod_{m=1}^{M} p(C[m] \mid A[m], \theta_4)\right\}$$

高斯混合模型—完备数据

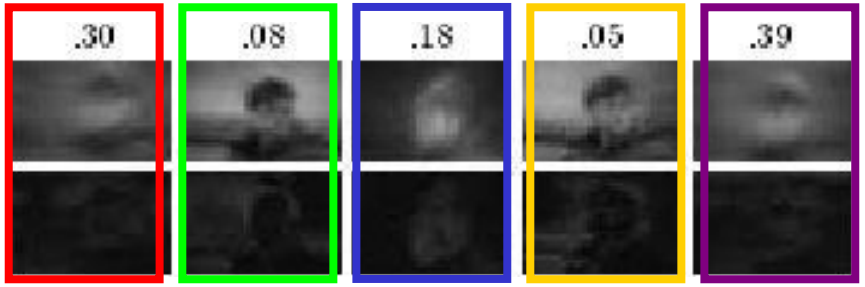
头姿类别 z

观测图像 x

总体分布 $p(z,x)=p(z|\theta_1)p(x|z,\theta_2)$

参数 $\theta=\{w_k,\mu_k,\sum_k,k=1,\cdots,K\}$

参数估计



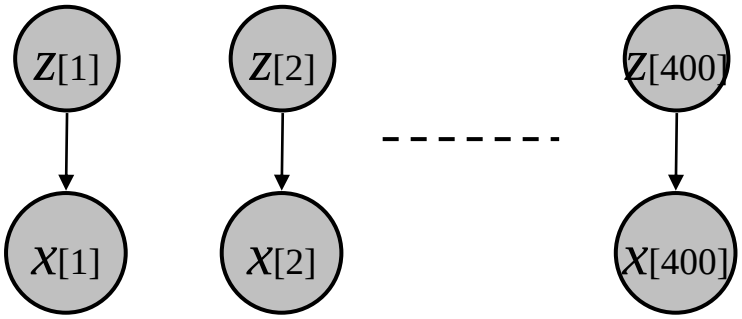
400 幅图像及其头姿类别标注

$$D=\left(\begin{pmatrix}x[1]\\z[1]\end{pmatrix},\cdots,\begin{pmatrix}x[400]\\z[400]\end{pmatrix}\right)$$

完备数据

高斯混合模型—完备数据

$$\theta = \{w_k, \mu_k, \Sigma_k, k = 1, \cdots, K\}$$



$$\log p\left(x[1:M], z[1:M] \mid \theta\right)$$

$$= \sum_m \log p\left(x[m], z[m] \mid \theta\right)$$

$$= \max_m \sum_m \left[\log p\left(z[m] \mid \theta\right) + \log p\left(x[m] \middle| z[m], \theta\right)\right]$$

$$= \max_m \sum_m \log p\left(z[m] \mid w_{1:K}\right) + \max_m \sum_m \log N\left(x[m] \mid \mu_{z[m]}, \Sigma_{z[m]}\right)$$

$$\sum_{k=1}^{K} \sum_{m:\, z[m]=k} \log w_k$$

$$\sum_{k=1}^{K} \max_m \sum_{m:\, z[m]=k} \log N\left(x[m] \mid \mu_k, \Sigma_k\right)$$

离散变量 z 的400个样本下的对数似然值，

对第k类，高斯变量 x 的 N_k 个样本下的对数似然值，

$$w_k^{ML} = \frac{N_k}{M} = \frac{1}{M}\sum_{m=1}^{M} 1\left(z[m] = k\right)$$

$$\begin{cases} \bar{\mu}_k^{ML} = \dfrac{1}{N_k} \displaystyle\sum_{m:\, z[m]=k} x[m] \\ \bar{\Sigma}_k^{ML} = \dfrac{1}{N_k} \displaystyle\sum_{m:\, z[m]=k} \left(x[m] - \bar{\mu}_k^{ML}\right)\left(x[m] - \bar{\mu}_k^{ML}\right)^T \end{cases}$$

$$\mu_k^{ML}, \Sigma_k^{ML}$$

Parameter learning

— ML (Known structure, incomplete data)

Expectation-Maximization 算法

高斯混合模型—不完备数据

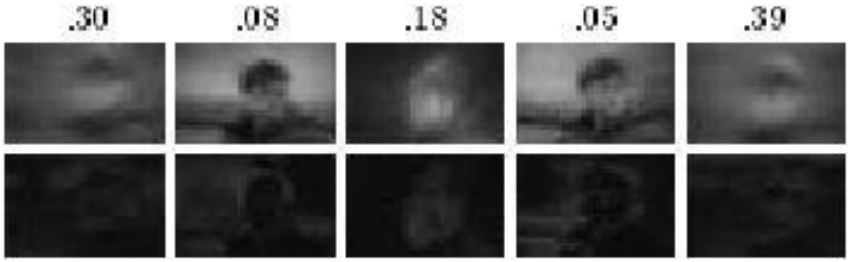
头姿类别 z

观测图像 x

总体分布 $p(z,x)=p(z)\,p(x\mid z)$

参数 $\theta=\{w_k,\mu_k,\Sigma_k,k=1,\cdots,K\}$

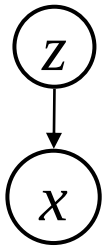
参数估计

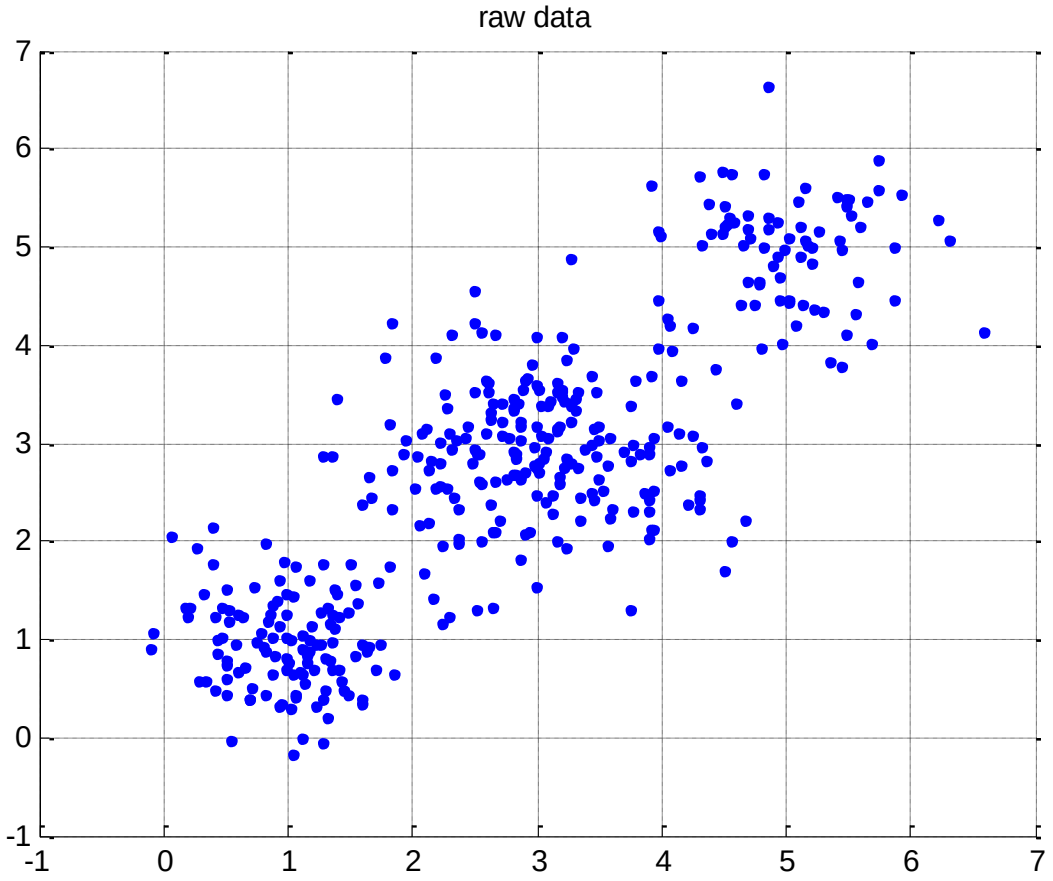


数据：400 幅图像 $D=(x[1],\cdots,x[400])$

不完备数据

Homework3_em






```
function [weight, meanvec, stdvec] = EmEstimate(x, iternum)
% x is the input observation, D-dim vectors * N
% iternum is the given number for EM iterations
```

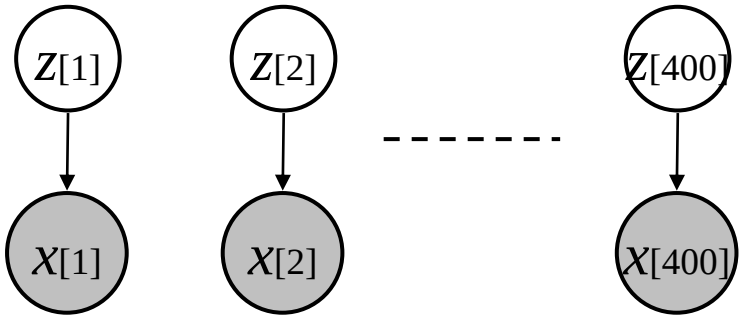
高斯混合模型—不完备数据

$$\theta = \{w_k, \mu_k, \Sigma_k, k = 1, \cdots, K\}$$

$$\log p(x[1:M]|\theta)$$

$$= \sum_m \log p(x[m]|\theta)$$

$$= \sum_m \log\left(\sum_k w_k N(x[m]|\mu_k, \Sigma_k)\right)$$



$$\frac{\partial}{\partial w_k}\log p(x[1:M]|\theta) = \sum_m \frac{\partial}{\partial w_k}\log\left(\sum_k w_k N(x[m]|\mu_k, \Sigma_k)\right)$$

$$= \sum_m \frac{1}{\sum_k w_k N(x[m]|\mu_k, \Sigma_k)} \frac{\partial}{\partial w_k}\left(\sum_k w_k N(x[m]|\mu_k, \Sigma_k)\right)$$

$$\frac{\partial}{\partial \mu_k}\log p(x[1:M]|\theta) =$$

$$\frac{\partial}{\partial \Sigma_k}\log p(x[1:M]|\theta) =$$

需要求解

$$\theta = \{w_k, \mu_k, \Sigma_k, k = 1, \cdots, K\}$$

联立非线性方程！

EM一般讨论

- 记 x 为全体观测值，记 z 为全体隐变量
  - 联合分布： $p(x, z \mid \theta)$

$$\theta^{ML} = \arg\max_{\theta} \log p\left(x \mid \theta\right)$$

$$= \log \sum_{z} p\left(x, z \mid \theta\right)$$

- $\log p(x \mid \theta)$ is called the incomplete log-likelihood.
  - 联合分布 $p(x, z \mid \theta)$ 的分解表示得不到利用

- $\log p(x, z \mid \theta)$ is called the complete log-likelihood.
  - 利用分解

EM算法描述

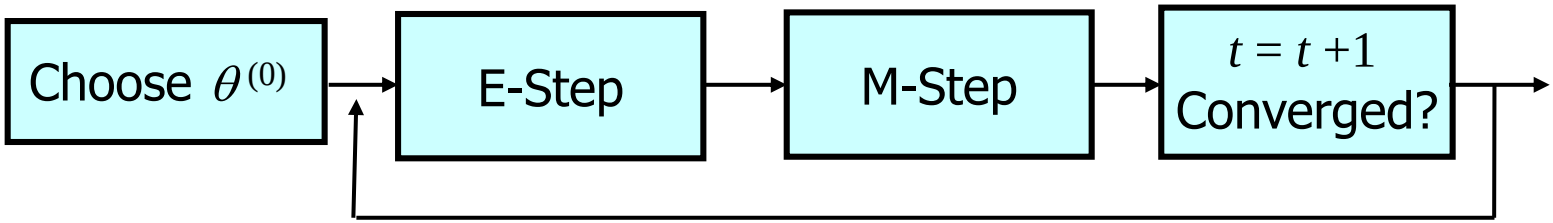
- z 是隐变量，完备对数似然函数 $\log p(x, z \mid \theta)$ 是 z 的一个函数
- 使用 完备对数似然函数的期望？
- 完备对数似然函数 $\log p(x, z \mid \theta)$ 在条件分布 $p(z \mid \theta^{(old)}, x)$ 下的期望

$$Q\left(\theta \mid \theta^{(old)}\right) = E\left[\log p\left(x, z \mid \theta\right) \mid \theta^{(old)}, x\right] = \sum_{z} p\left(z \mid \theta^{(old)}, x\right) \log p\left(x, z \mid \theta\right)$$

- 求解 $\theta^{*} = \arg\max_{\theta} Q\left(\theta \mid \theta^{(old)}\right)$

  成立 $\log p\left(x \mid \theta^{(old)}\right) \leq \log p\left(x \mid \theta^{*}\right)$

- EM算法是一个迭代过程



Jensen Inequality : for convex $\cap$ function f

$$E\left[f(U)\right] \le f\left(E[U]\right)$$

EM算法证明

$$\log p(x,z\mid\theta)=\log p(x\mid\theta)+\log p(z\mid\theta,x),\ \forall\theta$$ applying $E[\ldots \mid \theta^{(old)}, x]$

$$E\left[\log p(x,z\mid\theta)\mid\theta^{(old)},x\right]=\log p(x\mid\theta)+E\left[\log p(z\mid\theta,x)\mid\theta^{(old)},x\right],\ \forall\theta$$

$$E\left[\log p\left(x,z\mid\theta^{(old)}\right)\mid\theta^{(old)},x\right]=\log p\left(x\mid\theta^{(old)}\right)+E\left[\log p\left(z\mid\theta^{(old)},x\right)\mid\theta^{(old)},x\right]$$

$$\begin{pmatrix} E\left[\log p(x,z\mid\theta)\mid\theta^{(old)},x\right] \\ -E\left[\log p\left(x,z\mid\theta^{(old)}\right)\mid\theta^{(old)},x\right]\end{pmatrix}=\begin{pmatrix}\log p(x\mid\theta)\\ -\log p\left(x\mid\theta^{(old)}\right)\end{pmatrix}+E\left[\log\frac{p(z\mid\theta,x)}{p\left(z\mid\theta^{(old)},x\right)}\mid\theta^{(old)},x\right]$$

$$Q\left(\theta\mid\theta^{(old)}\right)=E\left[\log p(x,z\mid\theta)\mid\theta^{(old)},x\right]$$

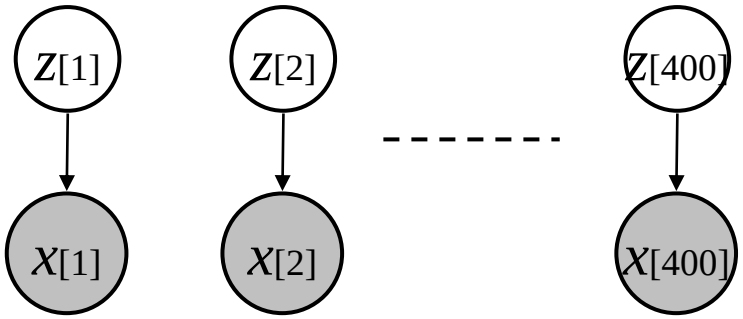
$$\le\log E\left[\frac{p(z\mid\theta,x)}{p\left(z\mid\theta^{(old)},x\right)}\mid\theta^{(old)},x\right]$$

$$=\log\sum_z\frac{p(z\mid\theta,x)}{p\left(z\mid\theta^{(old)},x\right)}p\left(z\mid\theta^{(old)},x\right)$$

EM Example: Learning with GMM

$$Q\left(\theta \mid \theta^{(old)}\right) = E\left[\log p\left(x, z \mid \theta\right) \mid \theta^{(old)}, x\right]$$

- 给定不完备数据 $(x[1], \ldots, x[M])$



$$E\left[\log p\left(x[1:M], z[1:M] \mid \theta\right) \mid \theta^{(old)}, x[1:M]\right]$$

$$= \sum_m E\left[\log p\left(x[m], z[m] \mid \theta\right) \mid \theta^{(old)}, x[1:M]\right]$$

$$= \sum_m \sum_{z[m]} p\left(z[m] \mid \theta^{(old)}, x[m]\right) \log p\left(x[m], z[m] \mid \theta\right)$$

$$= \sum_m \sum_{z[m]} p\left(z[m] \mid \theta^{(old)}, x[m]\right) \left\{\log p\left(x[m] \mid \theta, z[m]\right) + \log p\left(z[m] \mid \theta\right)\right\}$$

$$= \sum_m \sum_k p\left(z[m] = k \mid \theta^{(old)}, x[m]\right) \left\{\log p\left(x[m] \mid \theta, z[m] = k\right) + \log p\left(z[m] = k \mid \theta\right)\right\}$$

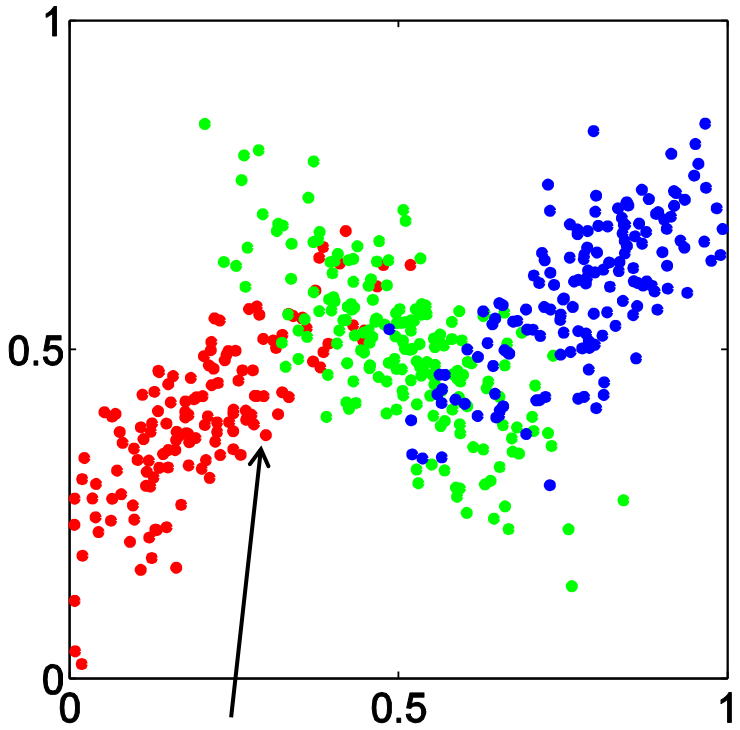
$$\max_{\{w_k, \mu_k, \Sigma_k, k=1:K\}} \left\{ \max_{\mu_k, \Sigma_k} \sum_k \sum_m \gamma_m(k) \log N\left(x[m] \mid \mu_k, \Sigma_k\right) + \max_{\{w_k, k=1:K\}} \sum_k \sum_m \gamma_m(k) \log w_k \right\}$$

subject to: $\sum_k w_k = 1$

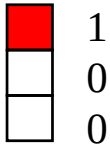
Posterior Probabilities

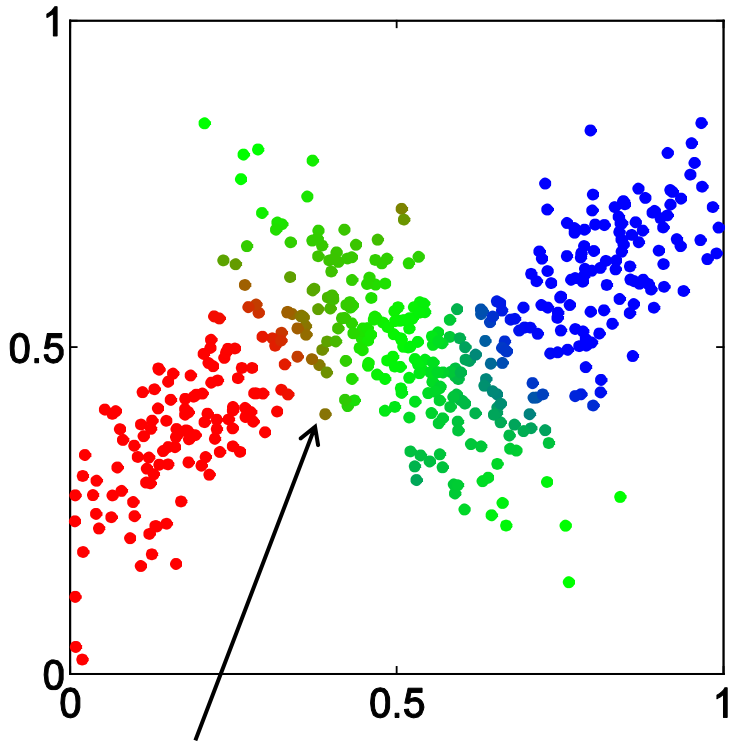
$$\max \text{不完备数据下目标函数} = \max \sum_{k=1}^{K}\sum_{m=1}^{M} \gamma_m(k) \log N\left(x[m] \mid \mu_k, \Sigma_k\right) + \max \sum_{k=1}^{K}\sum_{m=1}^{M} \gamma_m(k) \log w_k$$

$$\max \text{完备数据下目标函数} = \max \sum_{k=1}^{K}\sum_{m=1}^{M} 1\left(z[m]=k\right) \log N\left(x[m] \mid \mu_k, \Sigma_k\right) + \max \sum_{k=1}^{K}\sum_{m=1}^{M} 1\left(z[m]=k\right) \log w_k$$

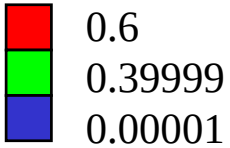


Hard assignment

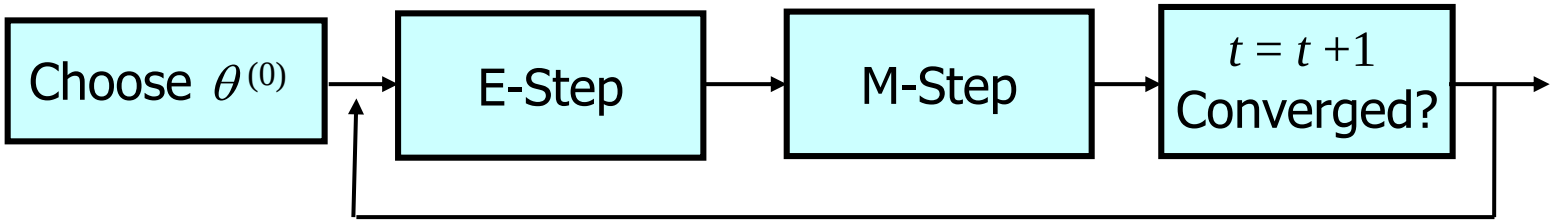


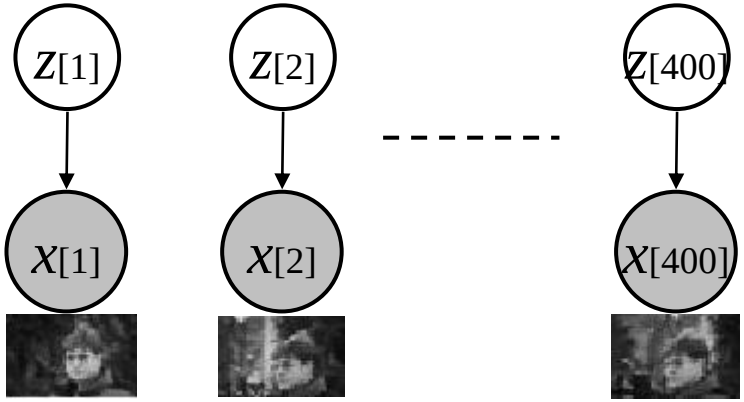


Soft assignment



EM Example: Learning with GMM





$$\gamma_m(k) = p\left(z[m]=k \mid \theta^{(old)}, x[m]\right)$$
$$= \frac{w_k^{(old)} N\left(x[m] \mid \mu_k^{(old)}, \Sigma_k^{(old)}\right)}{\sum_k w_k^{(old)} N\left(x[m] \mid \mu_k^{(old)}, \Sigma_k^{(old)}\right)}$$

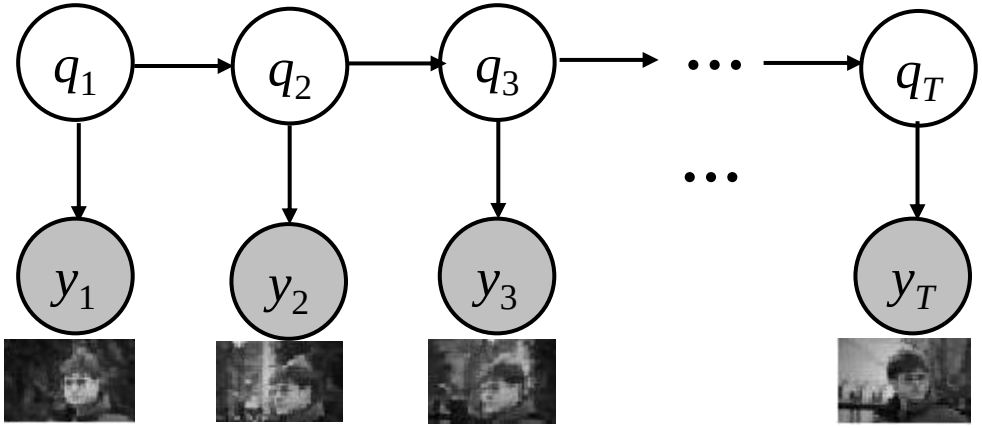
$$\mu_k^* = \frac{\sum_m \gamma_m(k)\, x[m]}{\sum_m \gamma_m(k)} \qquad \bar{\mu}_k^{ML} = \frac{\sum_{m=1}^{M} 1\left(z[m]=k\right) x[m]}{\sum_{m=1}^{M} 1\left(z[m]=k\right)}$$

$$\Sigma_k^* = \frac{\sum_m \gamma_m(k)\left(x[m]-\mu_k^*\right)\left(x[m]-\mu_k^*\right)^T}{\sum_m \gamma_m(k)}$$

$$= \frac{\sum_m \gamma_m(k) x[m] x[m]^T}{\sum_m \gamma_m(k)} - \mu_k^* \left(\mu_k^*\right)^T$$

$$w_k^* = \frac{\sum_m \gamma_m(k)}{M}$$

EM Example: Learning with HMM



$$\lambda = (\pi,\ A,\ B) \qquad \max_{\lambda} \log p(y_{1:T} \mid \lambda)\ ?$$

$$\lambda^* = \arg\max_{\lambda} E\left[\log p(y_{1:T}, q_{1:T} \mid \lambda) \mid \lambda^{(t)}, y_{1:T}\right]$$

$$= \arg\max_{\lambda} \left\{ \begin{array}{l} E\left[\log p(q_1 \mid \lambda) \mid \lambda^{(t)}, y_{1:T}\right] \\ +\sum_{t=2}^{T} E\left[\log p(q_t \mid \lambda, q_{t-1}) \mid \lambda^{(t)}, y_{1:T}\right] \\ +\sum_{t=1}^{T} E\left[\log p(y_t \mid \lambda, q_t) \mid \lambda^{(t)}, y_{1:T}\right] \end{array} \right\}$$

$$\log p(y_{1:T}, q_{1:T} \mid \lambda)$$
$$= \log p(q_1 \mid \lambda)$$
$$+\sum_{t=2}^{T} \log p(q_t \mid \lambda, q_{t-1})$$
$$+\sum_{t=1}^{T} \log p(y_t \mid \lambda, q_t)$$

From EM to SA

- 记 x 为全体观测值，记 z 为全体隐变量
  - 联合分布： $p(x, z \mid \theta)$

$$\theta^{ML} = \arg\max_{\theta} \log p(x \mid \theta)$$

$$Q\left(\theta \mid \theta^{(old)}\right) = E\left[\log p(x, z \mid \theta) \mid \theta^{(old)}, x\right] = \sum_{z} p\left(z \mid \theta^{(old)}, x\right) \log p(x, z \mid \theta)$$

Fisher Equality: $\frac{\partial \log p(x|\theta)}{\partial \theta} = E_{p(z|x,\theta)}\left[\frac{\partial \log p(x,z|\theta)}{\partial \theta}\right]$

$$\because \; E_{p(z|x,\theta)}\left[\frac{\partial \log p(z|x,\theta)}{\partial \theta}\right] = 0$$

Problem: The objective is to find a solution θ to $E_{Y \sim f(\cdot;\theta)}[H(Y;\theta)] = \alpha$, where $\theta \in R^d$, noisy observation $H(Y;\theta) \in R^d$.

- Gu & Kong, A stochastic approximation algorithm with Markov chain Monte-Carlo method for incomplete data estimation problems, PNAS 1998.
- Delyon, Lavielle, and Moulines. "Convergence of a stochastic approximation version of the EM algorithm." Annals of statistics, 1999.

Parameter learning

— ML (Known structure, complete data)

无向图

Parameter learning for UGMs (complete data)

- 贝叶斯网络：

  全体参数的对数似然函数＝各个节点处参数的对数似然函数之和

- 无向图模型：

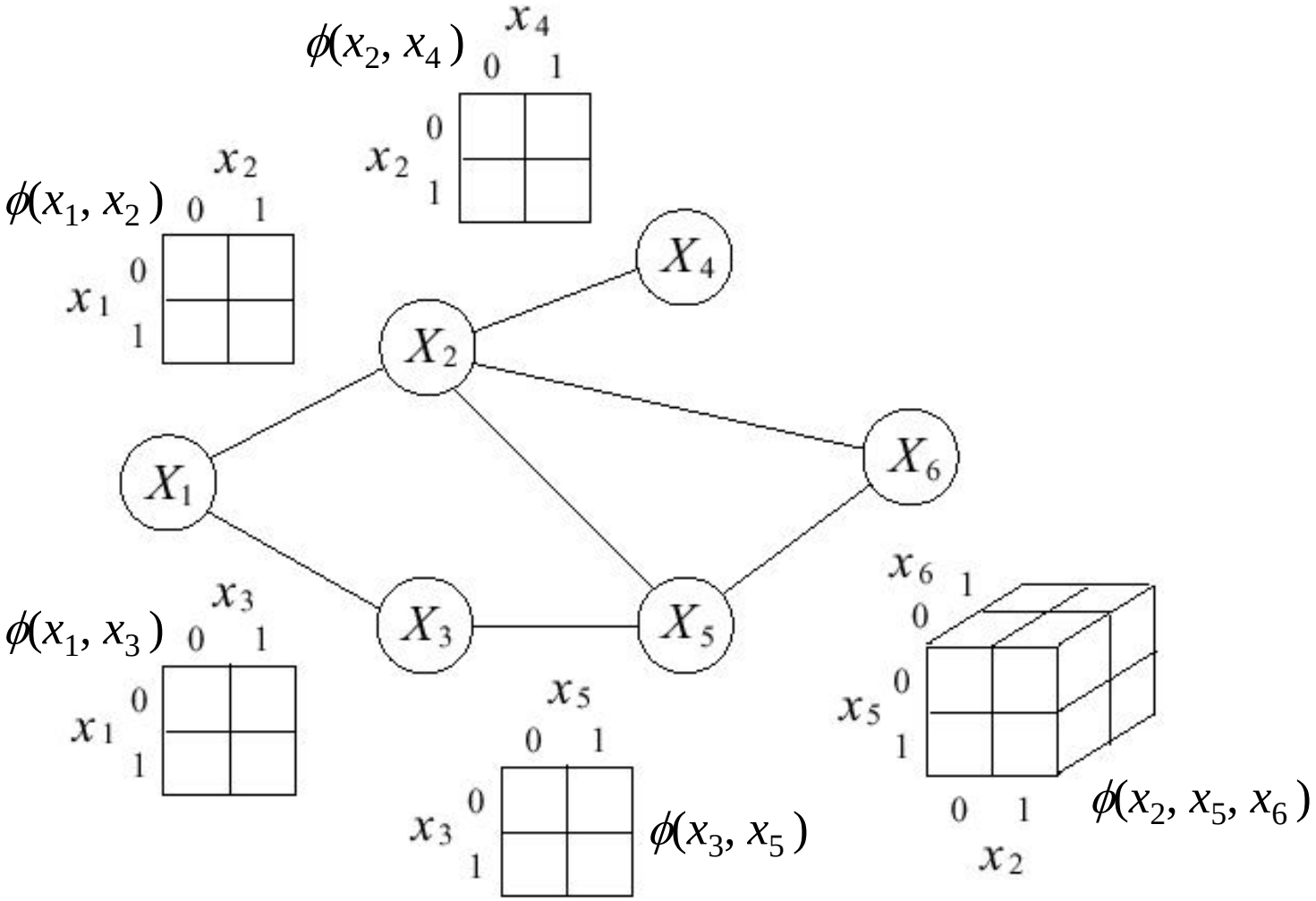
  归一化常数 Z 是所有参数的函数

$$\log p(x) = \sum_{c \in \mathcal{C}} \log \phi_c(x_c) - \log Z \qquad Z = \sum_x \prod_{c \in \mathcal{C}} \phi_c(x_c)$$

- 即使是完备数据情形下，无向图模型的参数估计需要推理计算（求边缘分布）
- 写出目标函数，直接运用梯度下降法等最优化方法

$$p(x_1, x_2, x_3, x_4, x_5, x_6) = \frac{1}{Z}\phi(x_1, x_2)\phi(x_1, x_3)\phi(x_2, x_4)\phi(x_3, x_5)\phi(x_2, x_5, x_6)$$

Example



无向图的对数似然函数：离散情形

- 给定IID样本集 $D = (x[1], \ldots, x[M])$

对数似然函数

$$\log p(D|\theta) = \sum_m \log p(x[m]|\theta)$$

$$= \sum_m \sum_x 1(x[m] = x) \log p(x|\theta)$$

$$= \sum_x \sum_m 1(x[m] = x) \log p(x|\theta)$$

$$= \sum_x Count(x) \log \left(\frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(x_c) \right)$$

$$= \sum_x \sum_c Count(x) \log \phi_c(x_c) - M \log Z$$

$$= \sum_c \sum_{x_c} Count(x_c) \log \phi_c(x_c) - M \log Z$$

样本集中特定 x 出现次数

$$count(x) = \sum_m 1(x[m] = x)$$

样本集中特定 x_c 出现次数

$$count(x_c) = \sum_{x \setminus x_c} count(x)$$

簇变量取特定值的次数(clique count)是(离散)无向图模型的充分统计量

对数似然函数求导

对数似然函数 $\log p(D|\theta) = \sum_c \sum_{x_c} Count(x_c) \log \phi_c(x_c) - M \log Z$

第一项求导 $\frac{\partial l_1}{\partial \phi_c(x_c)} = \frac{Count(x_c)}{\phi_c(x_c)}$

第二项求导 $\frac{\partial \log Z}{\partial \phi_c(x_c)} = \frac{1}{Z} \frac{\partial Z}{\partial \phi_c(x_c)} = \frac{1}{Z} \frac{\partial \sum_y \prod_d \phi_d(y_d)}{\partial \phi_c(x_c)}$ $\quad Z = \sum_x \prod_{c \in \mathcal{C}} \phi_c(x_c)$

$= \frac{1}{Z} \sum_y 1(y_d, x_c) \frac{\partial \prod_d \phi_d(y_d)}{\partial \phi_c(x_c)}$ 只留下$y_d = x_c$的项

$= \frac{1}{Z} \sum_y 1(y_d, x_c) \frac{1}{\phi_c(x_c)} \prod_d \phi_d(y_d)$

$= \frac{1}{\phi_c(x_c)} \sum_y 1(y_d, x_c) \frac{1}{Z} \prod_d \phi_d(y_d) = \frac{p(x_c)}{\phi_c(x_c)}$

必要条件：簇上边缘分布…

对数似然函数求导 $\frac{\partial \log p(D|\theta)}{\partial \phi_c(x_c)} = \frac{Count(x_c)}{\phi_c(x_c)} - M\frac{p(x_c)}{\phi_c(x_c)}$

在最大似然参数估计θ^{ML}下 $p^{ML}(x_c) = \frac{Count(x_c)}{M} = q_c(x_c)$

取θ^{ML} 时簇上的边缘分布＝簇上的经验分布

Model marginal = empirical marginal

只是指出：取得θ^{ML}时，簇上的边缘分布所需满足的必要条件

具体怎么求出θ^{ML}？

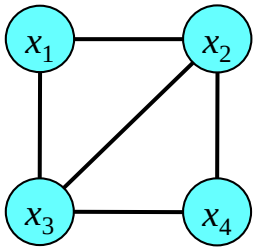
课程章节

- 第一章 引言（1）
- 第二章 图模型的表示理论（2）
  - **Semantics (DGM, UGM)**
  - **HMM, CRF**
- 第三章 图模型的推理理论（6）
  - 精确推理：variable-elimination，cluster-tree，triangulate
  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第四章 图模型的学习理论（3）
  - 参数学习：**maxlikelihoodEstimate**，RFLearning, BayesEstimate
  - 结构学习：StructureLearning
- 第五章 一个综合例子（1）

MLE for undirected graph

Tabular?	Decomposable?	Max-clique?	Method
√	√	√	Direct
√	—	—	IPF Iterative Proportional Fitting
×	—	—	GIS Generalized Iterative Scaling
×	—	—	梯度下降法

局部势函数 是否 全部定义在最大簇上?



$$p(x_1,x_2,x_3,x_4)=\frac{1}{Z}\phi(x_1,x_2,x_3)\phi(x_2,x_3,x_4)$$

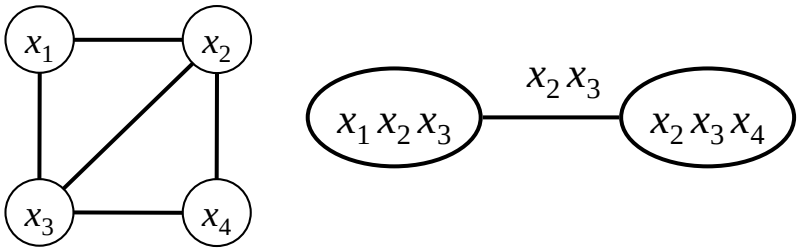
$$p(x_1,x_2,x_3,x_4)=\frac{1}{Z}\phi(x_1,x_2)\phi(x_1,x_3)\phi(x_2,x_3,x_4)$$

MLE for decomposable undirected graph

- 定理：一个可分解无向图 $\mathcal{G}$ 上的联合分布可以表示成：

$$p(x)=\frac{\prod_{c\in\mathcal{C}} p_c(x_c)}{\prod_{s\in\mathcal{S}} p_s(x_s)}$$

  - $\mathcal{C}$ 是最大簇集合，$\mathcal{S}$ 是图 $\mathcal{G}$ 的连接树上的隔离子集合



$$p(x_1,x_2,x_3,x_4)=\frac{1}{Z}\phi(x_1,x_2,x_3)\phi(x_2,x_3,x_4)$$
$$=\frac{p(x_1,x_2,x_3)\,p(x_2,x_3,x_4)}{p(x_2,x_3)}$$

在最大似然参数估计 θ^{ML} 下 $p^{ML}(x)=\frac{\prod_{c\in\mathcal{C}} q_c(x_c)}{\prod_{s\in\mathcal{S}} q_s(x_s)}=\frac{1}{Z}\prod_{c\in\mathcal{C}}\phi_c^{ML}(x_c)$

取 $\phi_c^{ML}(x_c)=\frac{q_c(x_c)}{\cdots}$

$$p^{ML}(x_1,x_2,x_3,x_4)=\frac{q(x_1,x_2,x_3)q(x_2,x_3,x_4)}{q(x_2,x_3)}=\frac{1}{Z}\phi^{ML}(x_1,x_2,x_3)\phi^{ML}(x_2,x_3,x_4)$$

$\phi^{ML}(x_1,x_2,x_3)=$　　$\phi^{ML}(x_2,x_3,x_4)=$

IPF (Iterative Proportional Fitting)

对数似然函数求导 $$\frac{\partial \log p(D|\theta)}{\partial \phi_c(x_c)} = M\frac{q_c(x_c)}{\phi_c(x_c)} - M\frac{p(x_c)}{\phi_c(x_c)}$$

最大似然参数估计θ^{ML}应满足 $$\frac{q(x_c)}{\phi_c(x_c)} = \frac{p(x_c)}{\phi_c(x_c)}$$

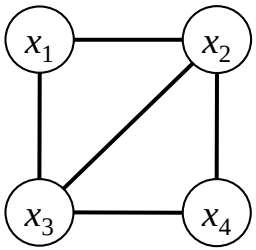
$p(x_c)$ 是 $\{\phi_c(x_c)\}$ 的函数

直接求解 $\{\phi_c(x_c)\}$ 很困难，$\{\phi_c(x_c)\}$ 是非线性方程的解

IPF: 将 $\phi_c(x_c)$ 视为下面不动点方程的不动点，迭代求解：

$$\phi_c(x_c) = \frac{q(x_c)}{p(x_c)}\phi_c(x_c) \qquad \phi_c^{(t+1)}(x_c) = \frac{q(x_c)}{p^{(t)}(x_c)}\phi_c^{(t)}(x_c)$$

推理计算

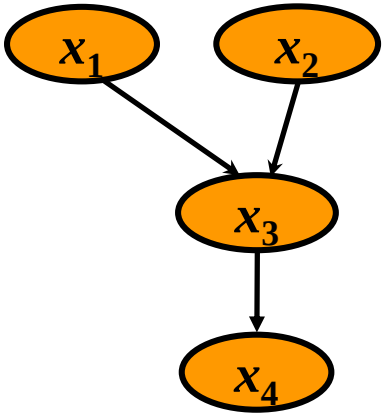


$$p(x_1,x_2,x_3,x_4) = \frac{1}{Z}\phi_{1,2}(x_1,x_2)\phi_{1,3}(x_1,x_3)\phi_{2,3,4}(x_2,x_3,x_4)$$

$$\phi_{1,2}^{(t+1)}(x_1,x_2) = \frac{q(x_1,x_2)}{\frac{1}{Z^{(t)}}\sum_{x_3,x_4}\phi_{1,2}^{(t)}(x_1,x_2)\phi_{1,2}^{(t)}(x_1,x_3)\phi_{2,3,4}^{(t)}(x_2,x_3,x_4)}\phi_{1,2}^{(t)}(x_1,x_2)$$

Example: Multinomial Bayes net

- 假设变量 X_n 有 K_n 个不同可能取值
- 结点 x_n 的条件分布 $p(x_n \mid pa_n)$ 含有一系列多元分布。对父结点集 pa_n 的每个可能取值组合 i，有一个多元分布 $p(x_n \mid pa_n = i)$
- $\theta = \{\theta_n \mid n = 1, \cdots, N\}$ $\theta_n = \{\theta_{n,i} \mid i = 1, \cdots\}$ $\theta_{n,i} = \{\theta_{n,i,k} \mid k = 1, \cdots, K_n\}$



$$p(x_3 \mid pa_3 = (0,0))$$
$$p(x_3 \mid pa_3 = (0,1))$$
$$p(x_3 \mid pa_3 = (1,0))$$
$$p(x_3 \mid pa_3 = (1,1))$$

x_n 的第k个可能取值

$$1 \le k \le K_n$$

$$\theta_{n,i,k} \triangleq p(x_n = k \mid pa_n = i)$$

结点

pa_n的第 i 个可能取值

$$1 \le i \le \prod_{x_l \in pa_n} K_l$$

Example: MLE for multinomial Bayes net

- 似然函数 $p\left(x_n[1:M]\,|\,\theta_n\right) \propto \prod_{m=1}^{M} p\left(x_n[m]\,|\,pa_n[m],\theta_n\right)$

$$= \prod_i \prod_k \left(\theta_{n,i,k}\right)^{N_{n,i,k}}$$

- 充分统计量 $N_{n,i,k} \propto \sum_{m=1}^{M} 1\left(pa_n[m]=i, x_n[m]=k\right)$

- 最大似然估计： $\hat{\theta}_{n,i,k} = \dfrac{N_{n,i,k}}{\sum_{l=1}^{K_n} N_{n,i,l}}$