概率图模型理论及应用

Theory and Applications of Probabilistic Graphical Models

(Lesson 7 – Maximum Likelihood)

欧智坚

清华大学电子工程系

Addr: 罗姆楼 6-104

Tel: 62796193

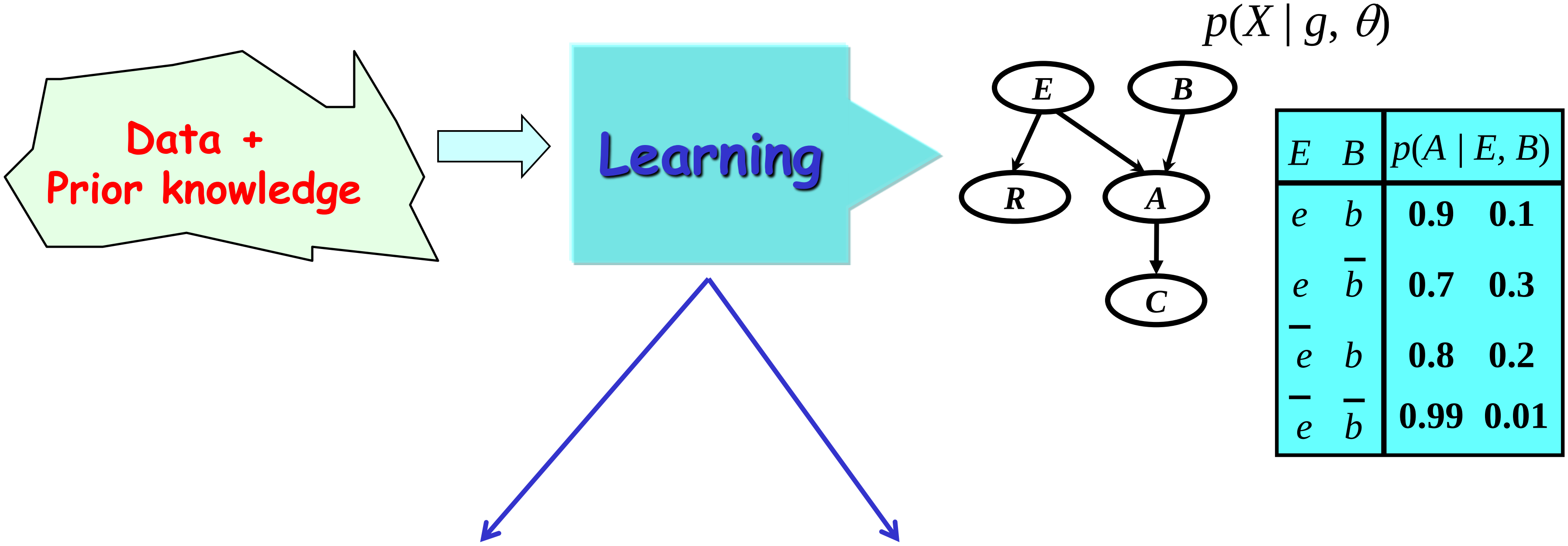
Email: ozj@tsinghua.edu.cn

课程章节

- 第一章 图模型的表示理论（2）
  - **Semantics (DGM, UGM)**
  - **HMM, CRF**
- 第二章 图模型的推理理论（4）
  - 精确推理：**variable-elimination**，**cluster-tree**，**triangulate**
  - 连续变量：**Kalman**
  - 采样近似：**sampling**
  - 变分近似：**variational**
- 第三章 图模型的学习理论（2）
  - 参数学习：**maxlikelihoodEstimate**，**RFLearning, BayesEstimate**
  - 结构学习：**StructureLearning**

	pgm-1 semantics √		pgm-2 hmm-crf √ pgm-3 exact √	pgm-4 kalman √ pgm-5 sampling √
pgm-6 variational √ pgm-7 ML	pgm-8 Bayesian			

Learning



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:

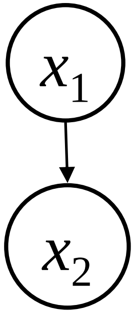
learning θ from data

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 $\sim p(x \mid g, \theta)$.
- 目标：从 $D = (x[1], \ldots, x[m], \ldots, x[M])$ 中估计出 g, θ

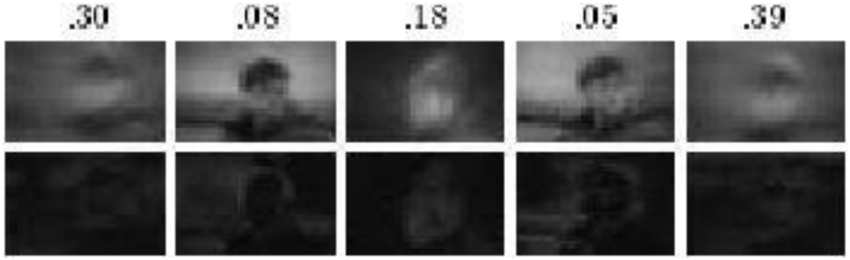
总体分布：$p(x_1, x_2 \mid g, \theta)$，其中 $w_k = p(x_1 = k)$ $p(x_2 \mid x_1 = k) = N(x \mid \mu_k, \Sigma_k)$

头姿类别



$\in 1{:}K$

观测图像 $\in R^{44*28}$

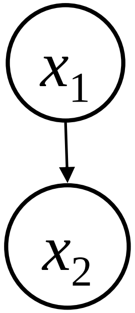


	x_1	x_2
$x[1]$	斜脸	
	侧脸	
	斜脸	
	斜脸	
$x[5]$	正脸	

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

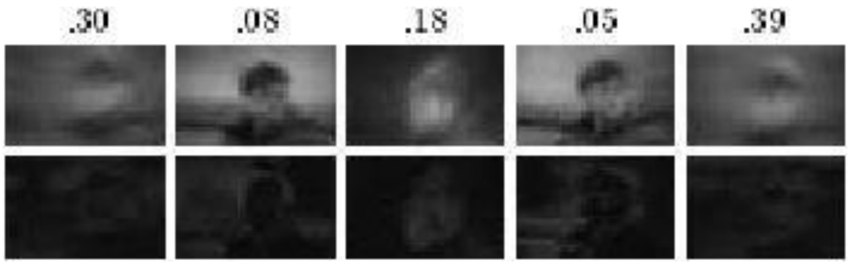
Data—complete, incomplete

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



头姿类别 $x_1 \in 1{:}K$

观测图像 $x_2 \in R^{44*28}$



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

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

数据：400 幅图像及其头姿类别，$D = (x[1], \ldots, x[400])$

- 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, $\theta = \left\{\theta_k \mid 1 \le k \le K\right\}$
- 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 $p\left(x[1:M] \mid \theta\right) = \underline{\prod_{m=1}^{M} p\left(x[m] \mid \theta\right) \ = \ \prod_{k=1}^{K} \theta_k^{N_k}}$

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<u>函数</u>
- 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：

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

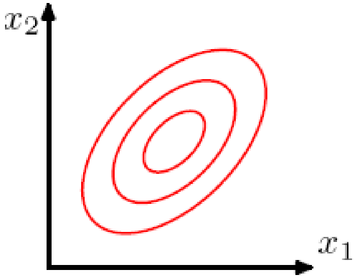
参数与样本的关联 完全通过充分统计量来体现

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

N_k : 在样本集中 $x[m]=k$ 出现的次数

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

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

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

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

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 $p(x_1 | pa_1),\ p(x_2 | pa_2), \ldots,\ p(x_N | pa_N)$

  有各自表征参数 $\{\theta_1, \theta_2, \ldots, \theta_N\} = \theta$

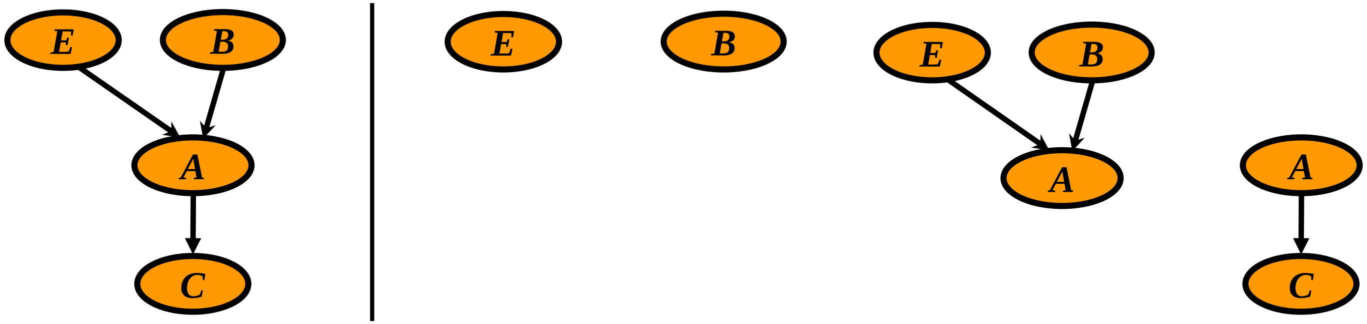
- IID样本集 $D = (x[1], \ldots, x[M])$ 下似然函数

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

❖ 对每个条件分布 $p(x_n | pa_n)$ 分别估计其参数 θ_n

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

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

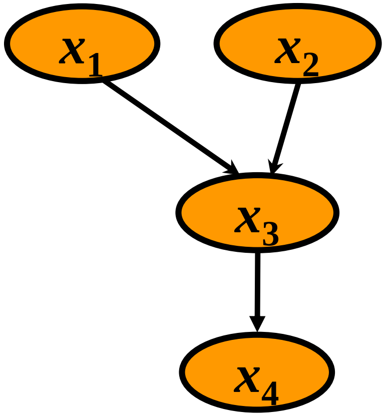


$$\max_{\theta} p(D|\theta) = \prod_{m=1}^{M} p(E[m]|\theta_1)\, p(B[m]|\theta_2)\, p(A[m]|E[m],B[m],\theta_3)\, p(C[m]|A[m],\theta_4)$$

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

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 $p(x_n | pa_n)$ 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 $p(x_n | pa_n{=}i)$
- $\theta = \{\theta_n \,|\, n = 1, \cdots, N\} \quad \theta_n = \{\theta_{n,\,i} \,|\, i = 1, \cdots\} \quad \theta_{n,\,i} = \{\theta_{n,\,i,\,k} \,|\, k = 1, \cdots, K_n\}$



$$p\left(x_3 \,|\, pa_3 = (0,0)\right)$$
$$p\left(x_3 \,|\, pa_3 = (0,1)\right)$$
$$p\left(x_3 \,|\, pa_3 = (1,0)\right)$$
$$p\left(x_3 \,|\, pa_3 = (1,1)\right)$$

x_n 的第k个可能取值

$$1 \le k \le K_n$$

$$\theta_{n,\,i,\,k} \;□\; p\left(x_n = k \,|\, pa_n = i\right)$$

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

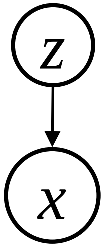
- 似然函数 $p\left(x_n[1:M] \mid \theta_n\right) \square \prod_{m=1}^{M} p\left(x_n[m] \mid pa_n[m], \theta_n\right)$

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

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

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

高斯混合模型—完备数据



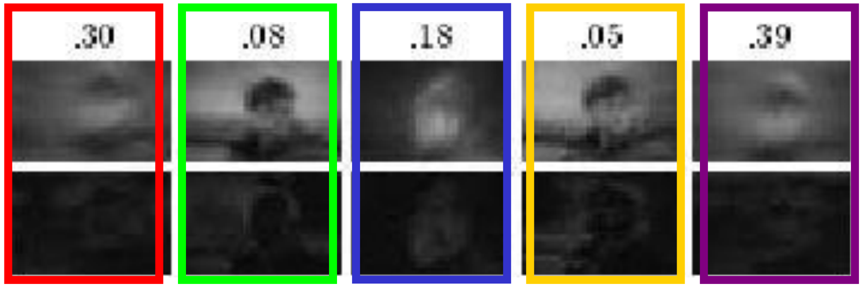
头姿类别

观测图像

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

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

参数估计

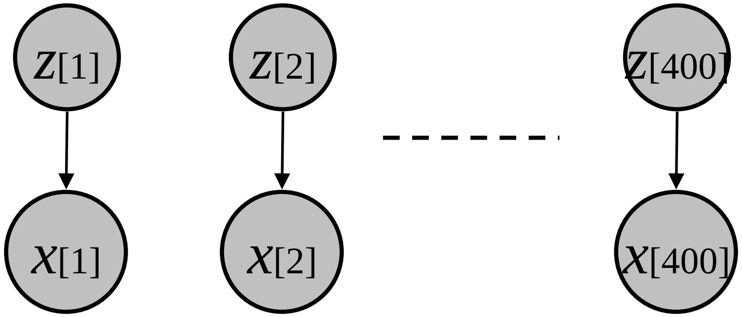


400 幅图像及其头姿类别标注 $D=\left(\begin{pmatrix}x[1]\\z[1]\end{pmatrix},\cdots,\begin{pmatrix}x[400]\\z[400]\end{pmatrix}\right)$

完备数据

高斯混合模型—完备数据

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



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

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

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

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

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

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

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，

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

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

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

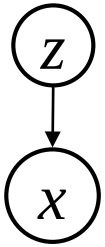
Parameter learning

— ML (Known structure, incomplete data)

Expectation-Maximization 算法

高斯混合模型—不完备数据

头姿类别

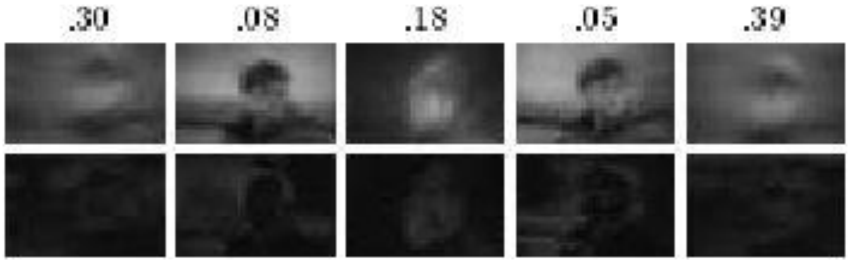


观测图像

总体分布 $p(z,x)=p(z)\,p(x\,|\,z)$

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

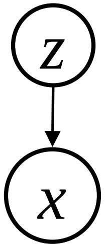
参数估计

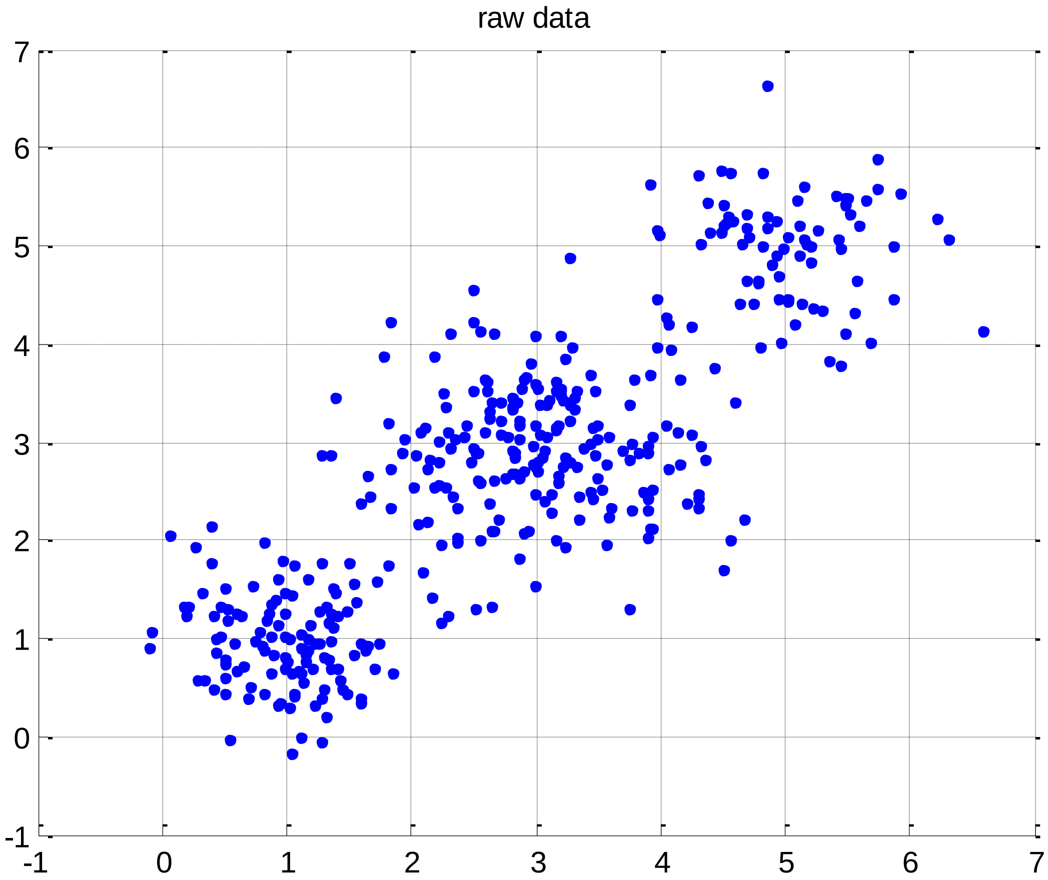


数据：400 幅图像 $D=\left(x[1],\cdots,x[400]\right)$

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
```

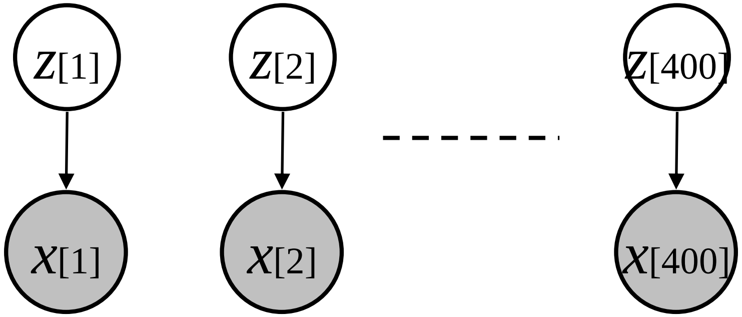
高斯混合模型—不完备数据

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

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

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

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



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

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

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

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

需要求解

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

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： $p(x,z\mid\theta)$

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

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

- $\log p(x\mid\theta)$ 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

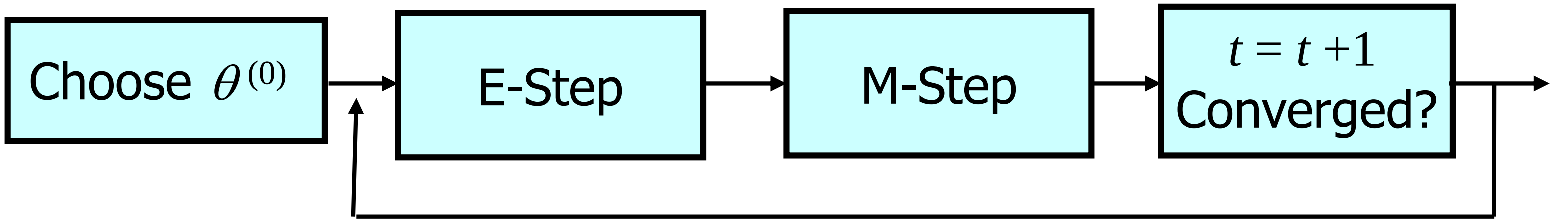
- z 是隐变量，完备对数似然函数 $\log p(x, z \mid \theta)$ 是 z 的一个函数
- 使用 完备对数似然函数的期望？
- 完备对数似然函数 $\log p(x, z \mid \theta)$ 在条件分布 $p(z|\theta^{old}, x)$ 下的期望

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

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

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

- EM算法是一个迭代过程



Choose $\theta^{(0)}$

E-Step

M-Step

$t = t + 1$
Converged?

Jensen Inequality : for convex ∩ function f

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

EM算法证明

$$\log p\left(x,z\mid\theta\right)=\log p\left(x\mid\theta\right)+\log p\left(z\mid\theta,x\right),\ \forall\theta$$

applying $E[\ldots\mid\theta^{(old)}, x]$

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

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

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

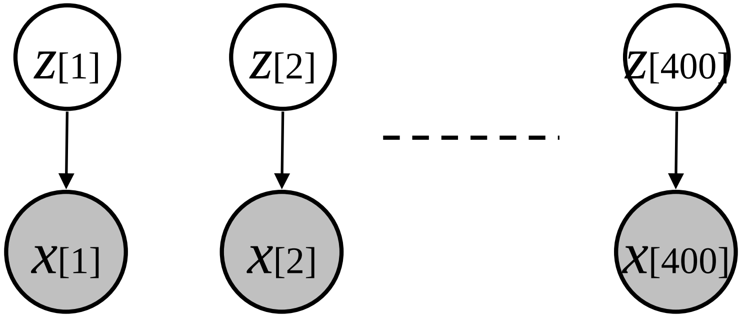
$$Q\left(\theta\mid\theta^{(old)}\right)=E\left[\log p\left(x,z\mid\theta\right)\mid\theta^{(old)},x\right]$$

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

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

EM Example: Learning with GMM

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



- 给定不完全数据 $(x[1],\ldots, x[M])$

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

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

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

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

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

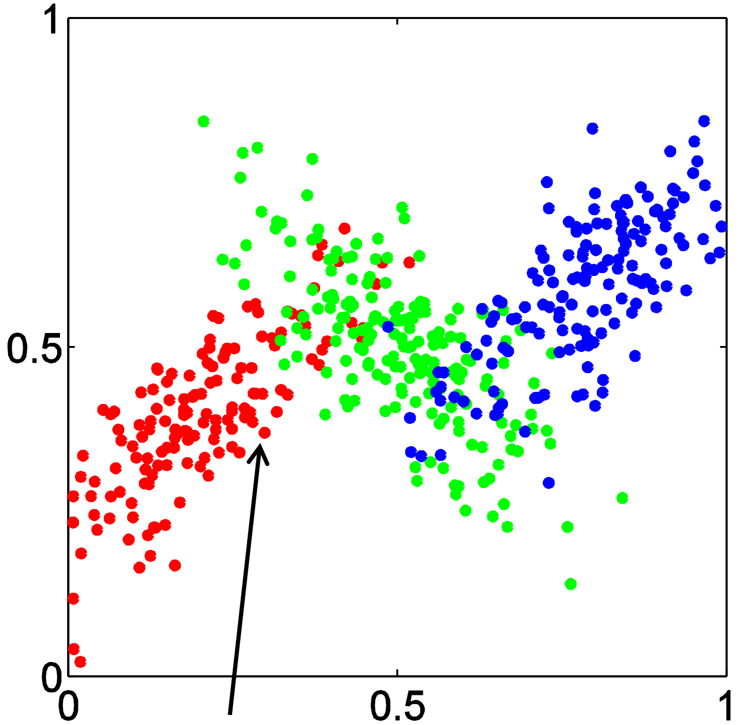
$$\max_{\{w_k, \mu_k, \Sigma_k, k=1:K\}} \left\{ \sum_k \overset{\max\limits_{\mu_k, \Sigma_k}}{\boxed{\sum_m \gamma_m(k) \log N\left(x[m] \mid \mu_k, \Sigma_k\right)}} + \overset{\max\limits_{\{w_k, k=1:K\}}}{\boxed{\sum_k \sum_m \gamma_m(k) \log w_k}} \right\}$$

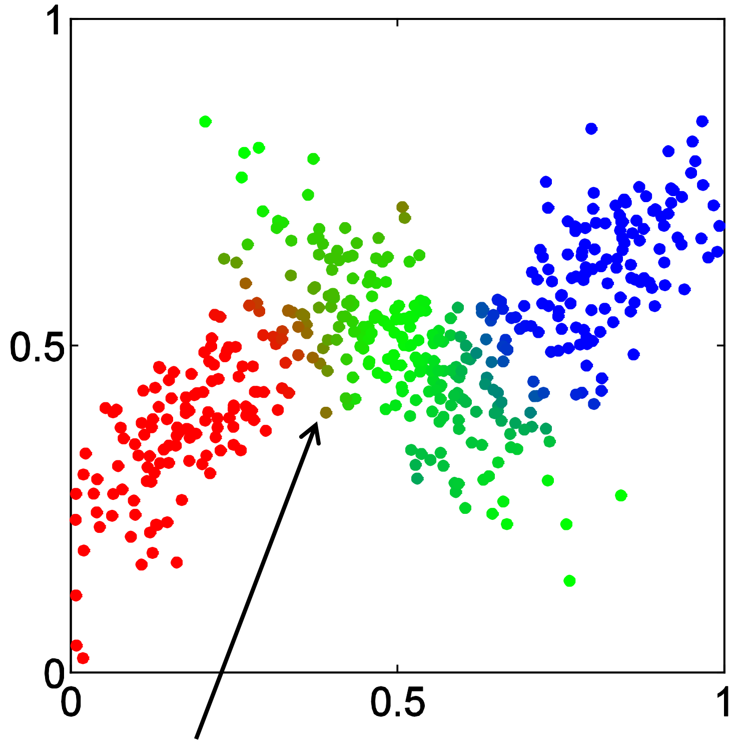
subject to: $\sum_k w_k = 1$

Posterior Probabilities

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

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





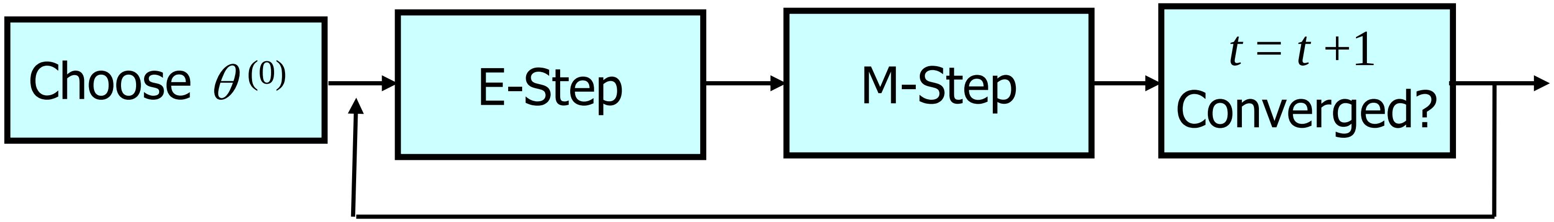
Hard assignment

🟥	1
□	0
□	0

Soft assignment

🟥	0.6
🟩	0.39999
🟦	0.00001

EM Example: Learning with GMM



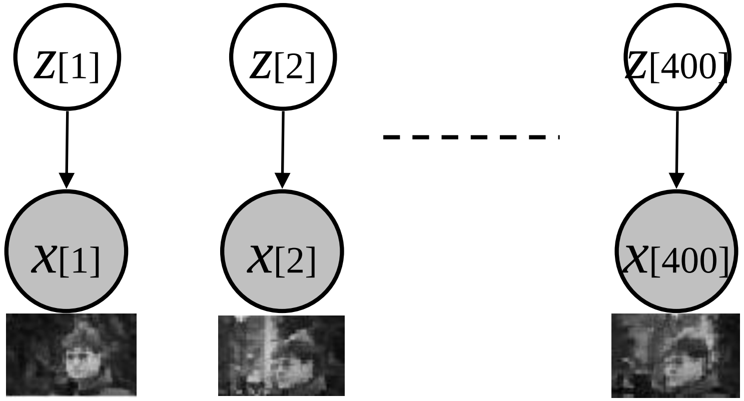
Choose $\theta^{(0)}$

E-Step

M-Step

$t = t+1$

Converged?



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

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

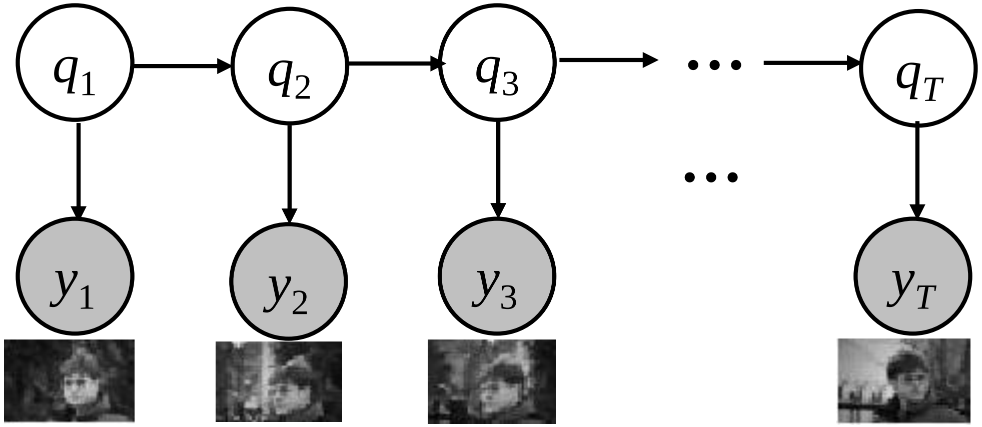
$$\mu_k^* = \frac{\sum_m \gamma_m(k)\, x[m]}{\sum_m \gamma_m(k)} \qquad \bar{\mu}_k^{ML} = \frac{\sum_{m=1}^{M} 1\left(z[m]=k\right) x[m]}{\sum_{m=1}^{M} 1\left(z[m]=k\right)}$$

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

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

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

EM Example: Learning with HMM



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

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

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

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

From EM to SA

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

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

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

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

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

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 $H(Y;\theta) \in R^d$.

- Delyon, Lavielle, and Moulines. "Convergence of a stochastic approximation version of the EM algorithm." Annals of statistics, 1999.
- Zhijian Ou. “A Review of Learning with Deep Generative Models from Perspective of Graphical Modeling.” arXiv:1808.01630.

UGM Semantics - Factorization property (F)

- A probability distribution $p(x_V)$ is said to factorize according to g, if there exist non-negative functions (called potential functions) $\phi_C(x_C)$ for all cliques C such that

$$p(x_V) = \frac{1}{Z}\prod_{C\in\mathbf{C}}\phi_C(x_C) \quad \text{or} \quad p(x_V) \propto \prod_{C\in\mathbf{C}}\phi_C(x_C)$$

where Z is the normalizing constant (partition function)

$$Z = \sum_{x_V}\prod_{C\in\mathbf{C}}\phi(x_C)$$

  - Potential functions $\phi_C(x_C)$ are not uniquely determined.
  - Without loss of generality, define potentials over maximum cliques.

Hammersley-Clifford Theorem: If p is strictly positive, (F)⟺(G).

UGMs and log-linear models

- Let each clique potential be a log-linear function

$$log\phi_C(x_C) = \theta_C^T f_C(x_C)$$

where $f_C(x_C)$ is a feature vector derived from the values of the variables x_C, θ_C is the associated feature weight vector.

- The resulting joint has the form

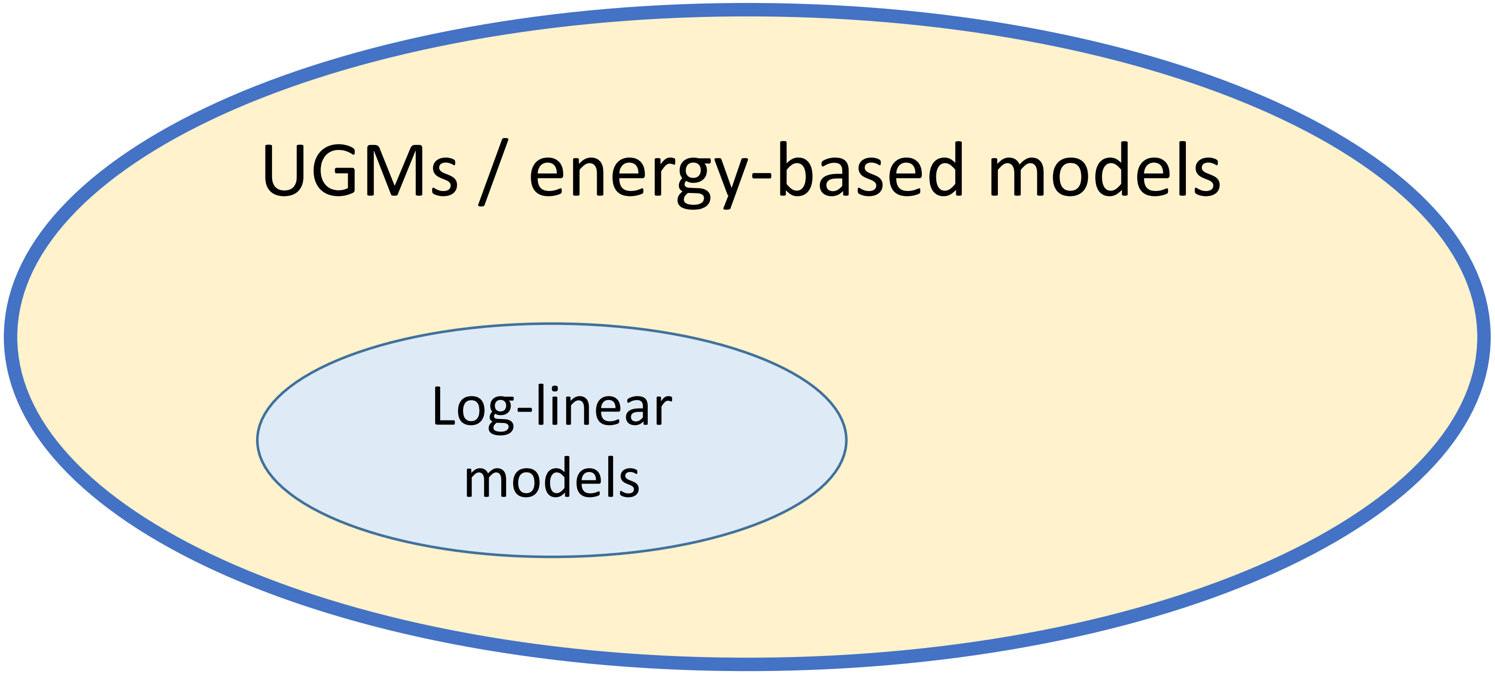
$$p(x_V) = \frac{1}{Z(\theta)} exp\left[\sum_C \theta_C^T f_C(x_C)\right]$$

This is known as a log-linear model or a Maximum Entropy model.

It can be proved that the maxent distribution is the same as
the maximum likelihood distribution from the closure of the set of log-linear RF distributions.

S. D. Pietra, V. D. Pietra, and J. Lafferty, “Inducing features of random fields”, IEEE PAMI, 1997.

Relationship between UGMs and other models


UGMs / energy-based models

Log-linear models

Feature-based potential representation in log-linear models

- Consider an edge potential $\phi_{s,t}(x_s, x_t)$ associated with two discrete variables x_s and x_t, both of which can take K values.
- Define a feature vector of length K^2 as follows:

$$f_{s,t}(x_s, x_t) = [\cdots, 1(x_s = j, x_t = k), \cdots]^T, \qquad j, k = 1, \cdots, K$$

with the associated weights:

$$\theta_{s,t} = \left[\cdots, log\left(\phi_{s,t}(x_s = j, x_t = k)\right), \cdots\right]^T, \qquad j, k = 1, \cdots, K$$

- Then the tabular potential $\phi_{s,t}(x_s, x_t)$ can be represented as the log-liner form

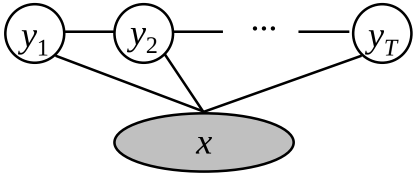
$$\phi_{s,t}(x_s, x_t) = exp\left[\theta_{s,t}^T f_{s,t}(x_s, x_t)\right]$$

- Note: the log-linear form is more general because we can choose (or learn) the features.

Linear-chain CRFs

for sequence tagging, e.g. POS tagging, shallow parser, Chinese word segmentation, …

$$p\left(y_{1:T} \mid x\right) \propto \exp\left\{\sum_{t=1}^{T-1} \psi_t\left(y_t, y_{t+1}, x\right) + \sum_{t=1}^{T} \psi_t\left(y_t, x\right)\right\}$$



Log-linear representation of tabular potentials

$$p\left(y_{1:T} \mid x\right) \propto \exp\left\{\sum_{t=1}^{T-1} \sum_i \lambda_i f_i\left(y_t, y_{t+1}, x, t\right) + \sum_{t=1}^{T} \sum_j \mu_j f_j\left(y_t, x, t\right)\right\}$$

Transition/edge features

$$\lambda_i f_i\left(y_t, y_{t+1}, x, t\right) = \lambda_i \cdot 1\left(y_t = prep, y_{t+1} = non\right)$$

State/node features

$$\mu_j f_j\left(y_t, x, t\right) = \mu_j \cdot 1\left(y_t = prep,\ x_t = \text{on}\right)$$

$$\mu_j f_j\left(y_t, x, t\right) = \mu_j \cdot 1\left(y_t = adv,\ x_t \text{ ends in } ly\right)$$

Training of UGMs in general

$$p(x;\theta) = \frac{1}{Z(\theta)} \exp[Q(x;\theta)]$$

Normalization constant:

$$Z(\theta) = \sum_{x} \exp[Q(x;\theta)]$$

- Maximum likelihood (ML) training

The scaled log-likelihood of observed $\{x_i, i = 1, \cdots, N\}$

$$l(\theta) \triangleq \frac{1}{N}\sum_{i=1}^{N} logp(x_i;\theta) = \left[\frac{1}{N}\sum_{i=1}^{N} Q(x_i;\theta)\right] - logZ(\theta)$$

$$\frac{\partial l(\theta)}{\partial \theta} = E_{\tilde{p}(x)}\left[\frac{\partial Q(x;\theta)}{\partial \theta}\right] - E_{p(x;\theta)}\left[\frac{\partial Q(x;\theta)}{\partial \theta}\right] = 0$$

Maximum Entropy

Expectation under empirical distribution $\tilde{p}(x) = \frac{1}{N}\sum_{i=1}^{N} 1(x = x_i)$

Expectation under model distribution $p(x;\theta)$

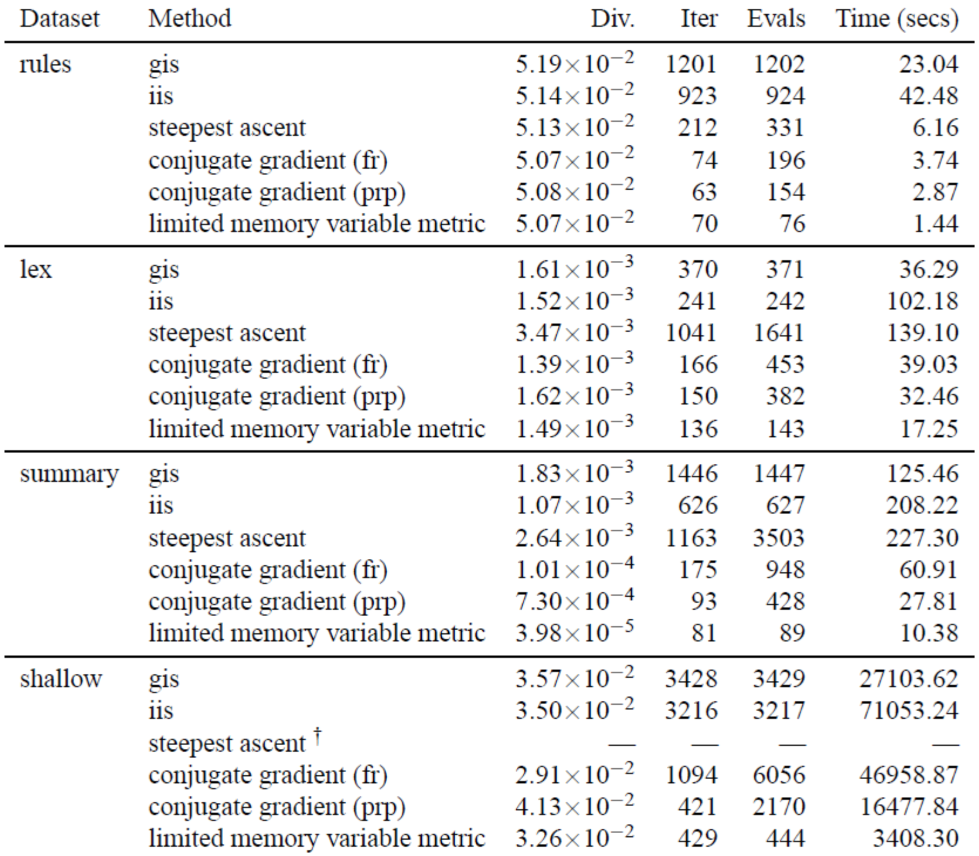
Training of UGMs - overview

- Roughly speaking, two types of approximate methods
- Gradient methods
  - Make explicit use of the gradient: Gradient descent, conjugate gradient, L-BFGS.
  - Stochastic approximation (SA)
  - Stochastic maximum likelihood (SML)
  - Persistent contrastive divergence (PCD)
- Lower bound methods
  - Generalized iterative scaling (GIS)
  - Improved iterative scaling (IIS)
  - Mostly studied in the context of maximum entropy (maxent) parameter estimation of log-linear models.
- In practice the gradient methods are shown to be much faster than the lower bound methods

Comparison on learning CRFs

- Div: the relative entropy between the fitted model and the training data
- Iter: Iteration number
- Evals: the number of calculating log-likelihood and gradient
- Time: the total time.

- T. Tieleman, “Training restricted boltzmann machines using approximations to the likelihood gradient”, ICML 2008.
- R. Malouf, “A comparison of algorithms for maximum entropy parameter estimation”, in Proc. Conference on Natural Language Learning (CoNLL), 2002.



Dataset	Method	Div.	Iter	Evals	Time (secs)
rules	gis	5.19×10^{-2}	1201	1202	23.04
	iis	5.14×10^{-2}	923	924	42.48
	steepest ascent	5.13×10^{-2}	212	331	6.16
	conjugate gradient (fr)	5.07×10^{-2}	74	196	3.74
	conjugate gradient (prp)	5.08×10^{-2}	63	154	2.87
	limited memory variable metric	5.07×10^{-2}	70	76	1.44
lex	gis	1.61×10^{-3}	370	371	36.29
	iis	1.52×10^{-3}	241	242	102.18
	steepest ascent	3.47×10^{-3}	1041	1641	139.10
	conjugate gradient (fr)	1.39×10^{-3}	166	453	39.03
	conjugate gradient (prp)	1.62×10^{-3}	150	382	32.46
	limited memory variable metric	1.49×10^{-3}	136	143	17.25
summary	gis	1.83×10^{-3}	1446	1447	125.46
	iis	1.07×10^{-3}	626	627	208.22
	steepest ascent	2.64×10^{-3}	1163	3503	227.30
	conjugate gradient (fr)	1.01×10^{-4}	175	948	60.91
	conjugate gradient (prp)	7.30×10^{-4}	93	428	27.81
	limited memory variable metric	3.98×10^{-5}	81	89	10.38
shallow	gis	3.57×10^{-2}	3428	3429	27103.62
	iis	3.50×10^{-2}	3216	3217	71053.24
	steepest ascent †	—	—	—	—
	conjugate gradient (fr)	2.91×10^{-2}	1094	6056	46958.87
	conjugate gradient (prp)	4.13×10^{-2}	421	2170	16477.84
	limited memory variable metric	3.26×10^{-2}	429	444	3408.30

Training of log-linear models

$$p(x;\theta) = \frac{1}{Z(\theta)} exp\left[\sum_{C} \theta_C^T f_C(x)\right]$$ where C indexes the cliques.

$$\frac{\partial l(\theta)}{\partial \theta_C} = E_{\tilde{p}(x)}[f_C(x)] - E_{p(x;\theta)}[f_C(x)] = 0$$ Statistics matching

Empirical statistics of features

Expected statistics of features

- $l(\theta)$ is convex in θ, so it has a unique global maximum which we can find using gradient-based optimizers. ☺
- The exact calculation of the gradient is intractable in general, involving high-dimensional integration. ☹

课程章节

- 第一章 图模型的表示理论（2）
  - Semantics (DGM, UGM)
  - HMM, CRF
- 第二章 图模型的推理理论（4）
  - 精确推理：variable-elimination，cluster-tree，triangulate
  - 连续变量：Kalman
  - 采样近似：sampling
  - 变分近似：variational
- 第三章 图模型的学习理论（2）
  - 参数学习：maxlikelihoodEstimate，RFLearning, BayesEstimate
  - 结构学习：StructureLearning

	pgm-1 semantics √		pgm-2 hmm-crf √ pgm-3 exact √	pgm-4 kalman √ pgm-5 sampling √
pgm-6 variational √ pgm-7 ML √	pgm-8 Bayesian			