

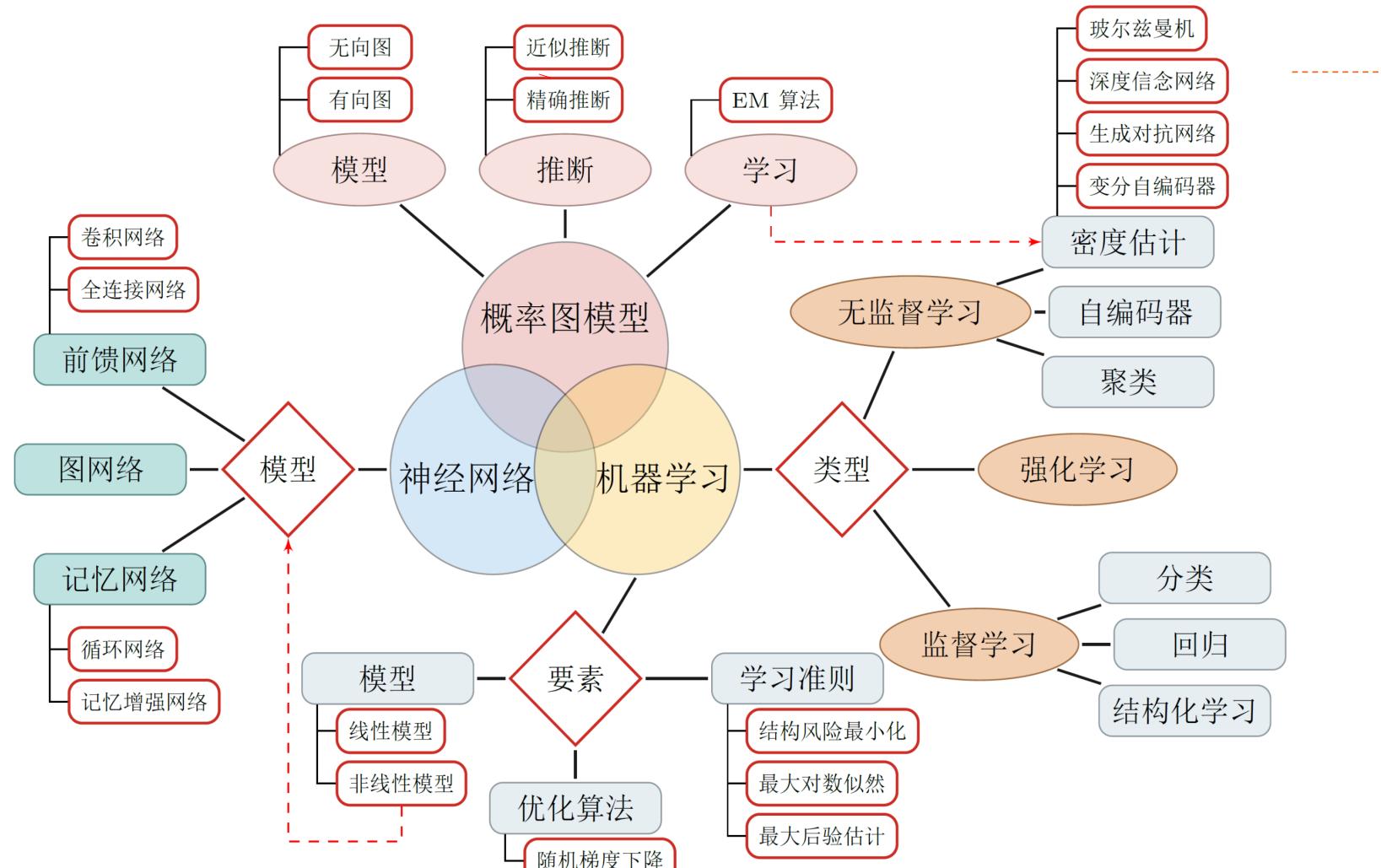
《神经网络与深度学习》



概率图模型

<https://nndl.github.io/>

课程概括

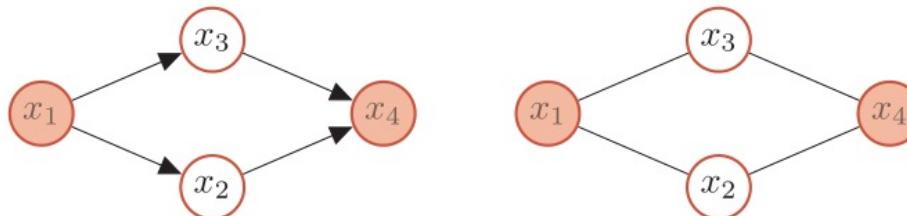


如何表示高维随机向量的概率密度？

- ▶ $P(X_1, X_2, \dots, X_K)$
 - ▶ 全概率公式
-
- ▶ 例子： $P(X_1, X_2, X_3, X_4)$
 - ▶ 如果已知 X_1 时， X_2 和 X_3 条件独立；已知 X_2 和 X_3 时， X_1 和 X_4 条件独立， 如何化简？
 - ▶
$$P(X_1, X_2, X_3, X_4) = P(X_1)P(X_2|X_1)P(X_3|X_1, X_2)P(X_4|X_1, X_2, X_3)$$

概率图模型

- ▶ 概率图模型是指一种用图结构来描述多元随机变量之间条件独立关系的概率模型。

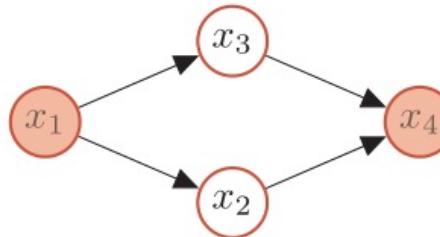


- ▶ 每个节点都对应一个随机变量，可以是观察变量，隐变量或是未知参数等；
- ▶ 每个连接表示两个随机变量之间具有依赖关系。

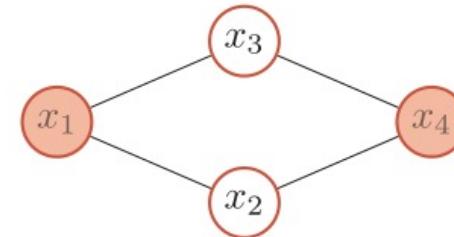
概率图模型

► 模型表示（图结构）

- 有向图
- 无向图



(a) 有向图：贝叶斯网络



(b) 无向图：马尔可夫随机场

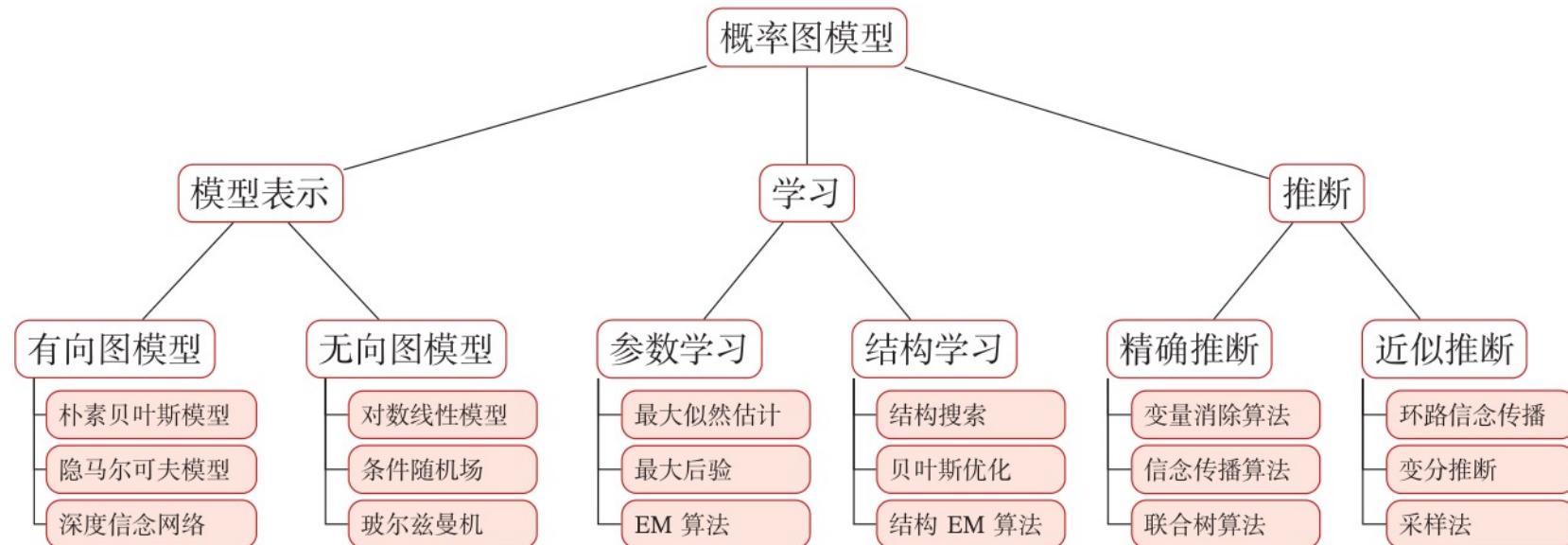
► 推断（Inference）

- 给定部分变量，推断另一部分变量的后验概率。

► 学习（Learning）

- 参数学习：给定一组训练样本，求解模型参数

概率图模型





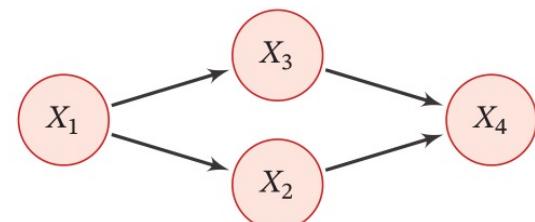
贝叶斯网络

- ▶ 有向图模型 (Directed Graphical model) , 也称为贝叶斯网络 (Bayesian Network) , 或信念网络 (Belief Network, BN) 。

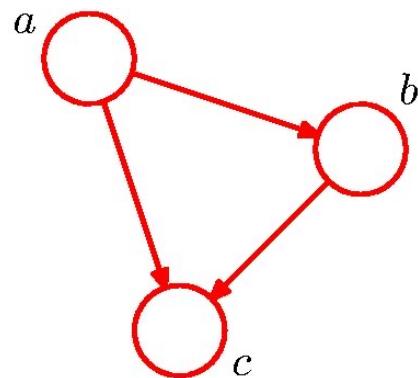
定义 11.1 - 贝叶斯网络: 对于一个 K 维随机向量 \mathbf{X} 和一个有 K 个节点的有向非循环图 G , G 中的每个节点都对应一个随机变量, 每个连接 e_{ij} 表示两个随机变量 X_i 和 X_j 之间具有非独立的因果关系. 令 \mathbf{X}_{π_k} 表示变量 X_k 的所有父节点变量集合, $P(X_k|\mathbf{X}_{\pi_k})$ 表示每个随机变量的局部条件概率分布 (Local Conditional Probability Distribution). 如果 \mathbf{X} 的联合概率分布可以分解为每个随机变量 X_k 的局部条件概率的连乘形式, 即

$$p(\mathbf{x}) = \prod_{k=1}^K p(x_k|\mathbf{x}_{\pi_k}), \quad (11.8)$$

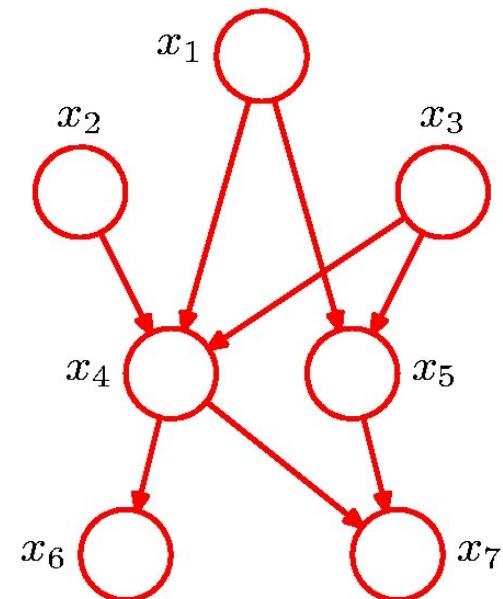
那么 (G, \mathbf{X}) 构成了一个贝叶斯网络.



练习



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



$$\begin{aligned} p(x_1, \dots, x_7) &= p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ &\quad p(x_5|x_1, x_3)p(x_6|x_4)p(x_7|x_4, x_5) \end{aligned}$$

局部马尔可夫性质

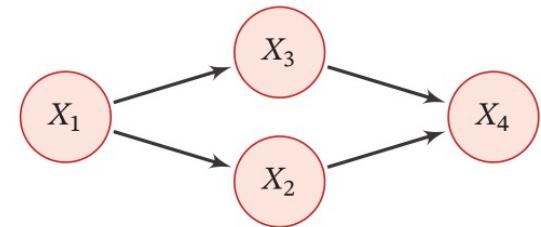
- ▶ 贝叶斯网络的局部马尔可夫性质：每个随机变量在给定父节点的情况下，条件独立于它的非后代节点。

$$X_k \perp\!\!\!\perp Z | X_{\pi_k}$$

- ▶ 利用局部马尔可夫性，可以对多元变量的联合概率进行简化，从而降低建模的复杂度。

- ▶ 例子：

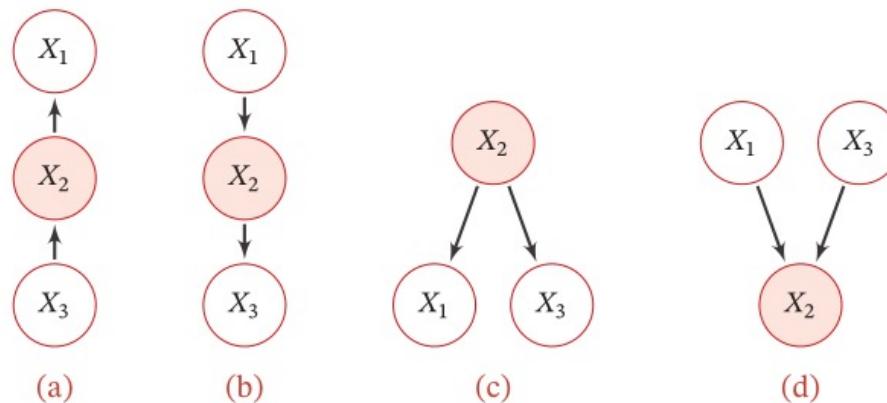
$$\begin{aligned} p(\mathbf{x}) &= p(x_1)p(x_2|x_1)p(x_3|x_1, x_2)p(x_4|x_1, x_2, x_3), \\ &= p(x_1)p(x_2|x_1)p(x_3|x_1)p(x_4|x_2, x_3), \end{aligned}$$



- ▶ 是4个局部条件概率的乘积，这样只需要 $1 + 2 + 2 + 4 = 9$ 个独立参数。

条件独立性

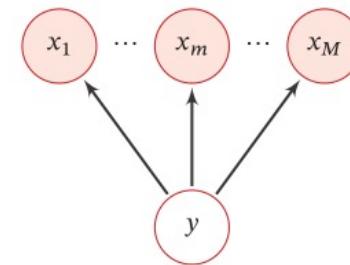
- ▶ 在贝叶斯网络中，如果两个节点是直接连接的，它们肯定是非条件独立的，是直接因果关系。
- ▶ 点是“因”，子节点是“果”。
- ▶ 如果两个节点不是直接连接的，但是它们之间有一条经过其他节点的路径连接互连接，它们之间的条件独立性就比较复杂。



常见的有向图模型

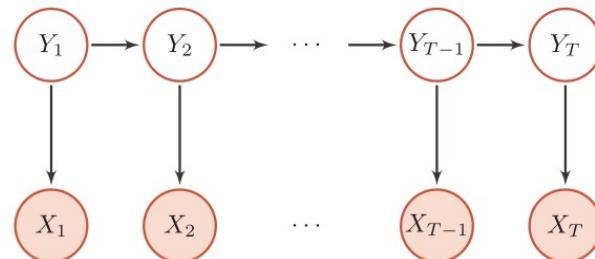
- ▶ 朴素贝叶斯分类器
- ▶ 给定一个有 M 维特征的样本 x 和类别 y ，类别的后验概率为

$$p(y|x; \theta) \propto p(y|\theta_c) \prod_{m=1}^M p(x_m|y; \theta_m)$$



隐马尔可夫模型 (Hidden Markov Model , HMM)

- ▶ 表示一种含有隐变量的马尔可夫过程



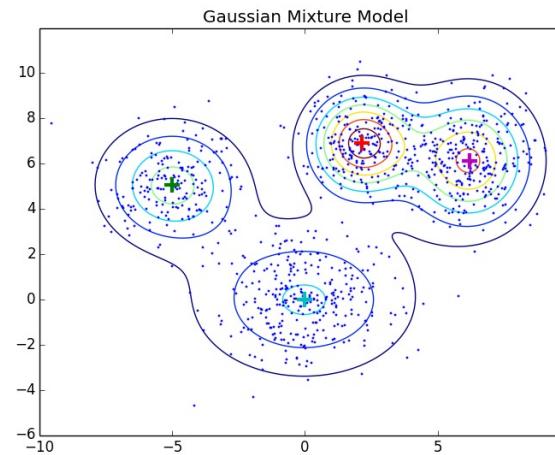
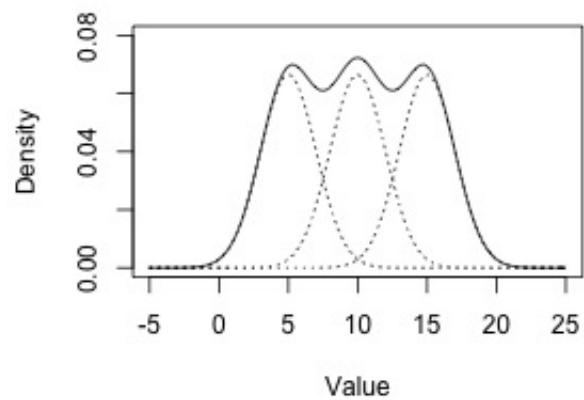
- ▶ 隐马尔可夫模型的联合概率可以分解为

$$p(\mathbf{x}, \mathbf{y}; \theta) = \prod_{t=1}^T p(y_t | y_{t-1}, \theta_s) p(x_t | y_t, \theta_t)$$

转移概率 输出概率

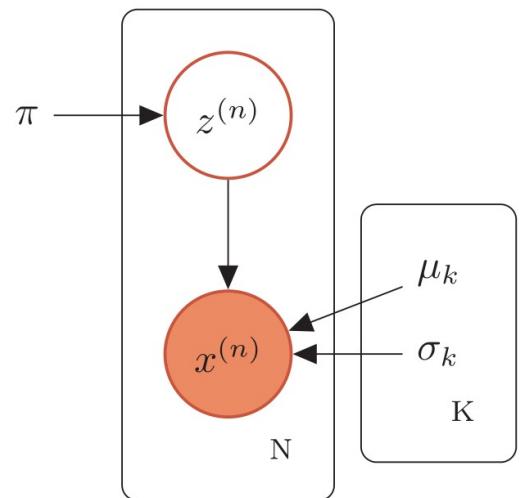
高斯混合模型

- ▶ 高斯混合模型（Gaussian Mixture Model， GMM）是由多个高斯分布组成的模型，其密度函数为多个高斯密度函数的加权组合。



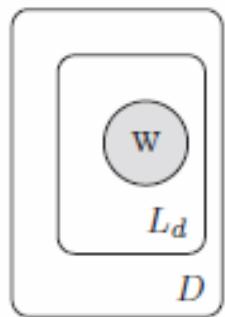
高斯混合模型

► 图模型表示

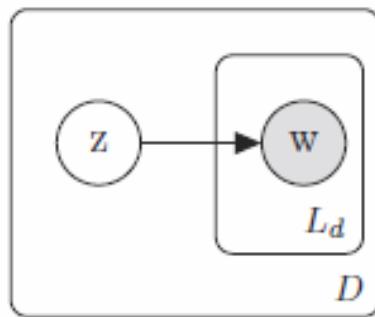


$$\mathcal{N}(x|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right)$$

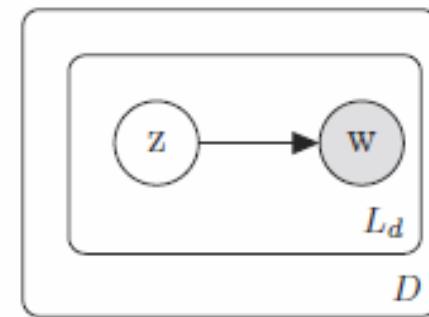
概率主题模型



(a) 一元语言模型



(b) 一元混合语言模型



(c) 概率主题模型

概率主题模型

Topics

gene	0.04
dna	0.02
genetic	0.01
...	

life	0.02
evolve	0.01
organism	0.01
...	

brain	0.04
neuron	0.02
nerve	0.01
...	

data	0.02
number	0.02
computer	0.01
...	

Documents

Seeking Life's Bare (Genetic) Necessities

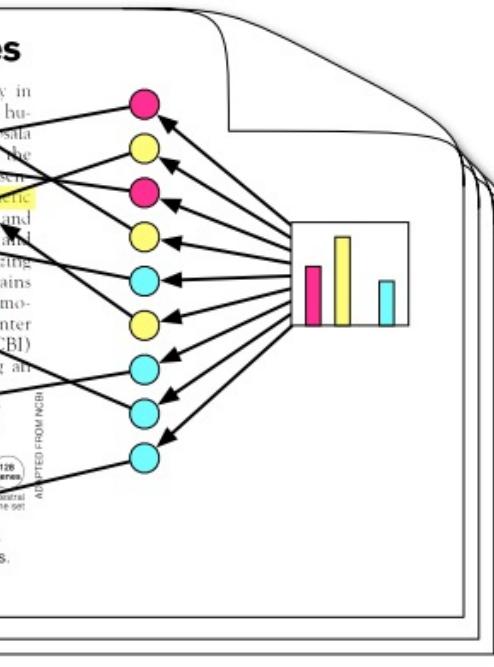
COLD SPRING HARBOR, NEW YORK—How many genes does an organism need to **survive**? Last week at the genome meeting here,⁹ two genome researchers with radically different approaches presented complementary views of the basic genes needed for **life**. One research team, using **computer** analyses to compare known **genomes**, concluded that today's organisms can be sustained with just 250 genes, and that the earliest life forms required a mere 128 **genes**. The other researcher mapped genes in a simple parasite and estimated that for this organism, 800 genes are plenty to do the job—but that anything short of 100 wouldn't be enough.

Although the numbers don't match precisely, those **predictions**

"are not all that far apart," especially in comparison to the 75,000 **genes** in the human genome, notes Siv Andersson of Uppsala University in Sweden, who arrived at the 800 number. But coming up with a consensus answer may be more than just a **genetic numbers game**, particularly as more and more **genomes** are completely mapped and sequenced. "It may be a way of organizing any newly **sequenced genome**," explains Arcady Mushegian, a **computational molecular biologist** at the National Center for Biotechnology Information (NCBI) in Bethesda, Maryland. Comparing all

Diagram illustrating the process of stripping down genomes to find minimal sets. It shows the Rickettsia genome (1763 genes) and the Mycoplasma genome (469 genes). A Venn diagram shows their overlap of 232 genes. Arrows show the removal of redundant and parasite-specific genes from the Rickettsia genome to reach a minimal set of 128 genes, which is then compared with the 122 genes of the Mycoplasma genome.

Topic proportions and assignments



SCIENCE • VOL. 272 • 24 MAY 1996

马尔可夫随机场

- ▶ 马尔可夫随机场，也称无向图模型，是一类用无向图来表示一组具有马尔可夫性质的随机变量 \mathbf{X} 的联合概率分布模型。

定义 11.2 - 马尔可夫随机场：对于一个随机向量 $\mathbf{X} = [X_1, \dots, X_K]^\top$ 和一个有 K 个节点的无向图 $G(\mathcal{V}, \mathcal{E})$ (可以存在循环)，图 G 中的节点 k 表示随机变量 $X_k, 1 \leq k \leq K$. 如果 (G, \mathbf{X}) 满足**局部马尔可夫性质**，即一个变量 X_k 在给定它的邻居的情况下独立于所有其他变量，

$$p(x_k | \mathbf{x}_{\setminus k}) = p(x_k | \mathbf{x}_{\mathcal{N}(k)}), \quad (11.15)$$

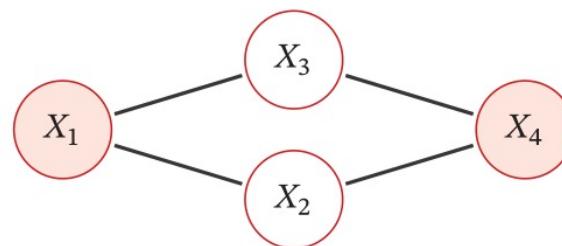
其中 $\mathcal{N}(k)$ 为变量 X_k 的邻居集合， $\setminus k$ 为除 X_k 外其他变量的集合，那么 (G, \mathbf{X}) 就构成了一个**马尔可夫随机场**.

无向图的马尔可夫性

无向图的马尔可夫性 无向图中的马尔可夫性可以表示为

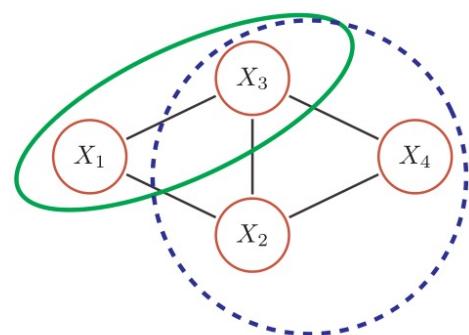
$$X_k \perp \mathbf{X}_{\setminus N(k), \setminus k} \mid \mathbf{X}_{N(k)},$$

其中 $\mathbf{X}_{\setminus N(k), \setminus k}$ 表示除 $\mathbf{X}_{N(k)}$ 和 X_k 外的其它变量。



团 (Clique)

► 团：一个全连通子图，即团内的所有节点之间都连边。



共有7个团

Hammersley-Clifford定理

- ▶ 无向图的联合概率可以分解为一系列定义在最大团上的非负函数的乘积形式。

定理 11.1 – Hammersley-Clifford 定理: 如果一个分布 $p(\mathbf{x}) > 0$ 满足无向图 G 中的局部马尔可夫性质, 当且仅当 $p(\mathbf{x})$ 可以表示为一系列定义在最大团上的非负函数的乘积形式, 即

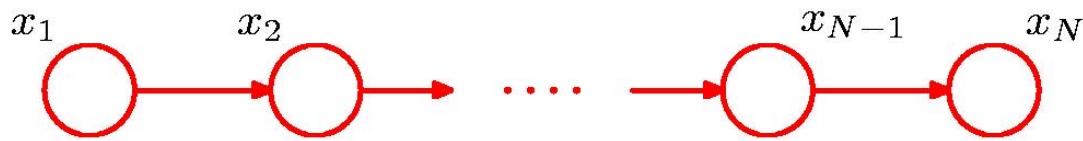
$$p(\mathbf{x}) = \frac{1}{Z} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x}_c), \quad (11.16)$$

其中 \mathcal{C} 为 G 中的最大团集合, $\phi_c(\mathbf{x}_c) \geq 0$ 是定义在团 c 上的势能函数 (Potential Function), Z 是配分函数 (Partition Function), 用来将乘积归一化为概率形式:

$$Z = \sum_{\mathbf{x} \in \mathcal{X}} \prod_{c \in \mathcal{C}} \phi_c(\mathbf{x}_c), \quad (11.17)$$

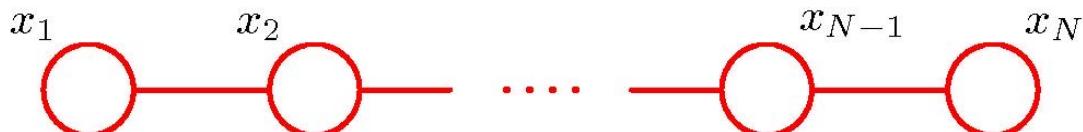
其中 \mathcal{X} 为随机向量 X 的取值空间.

有向图和无向图的转换

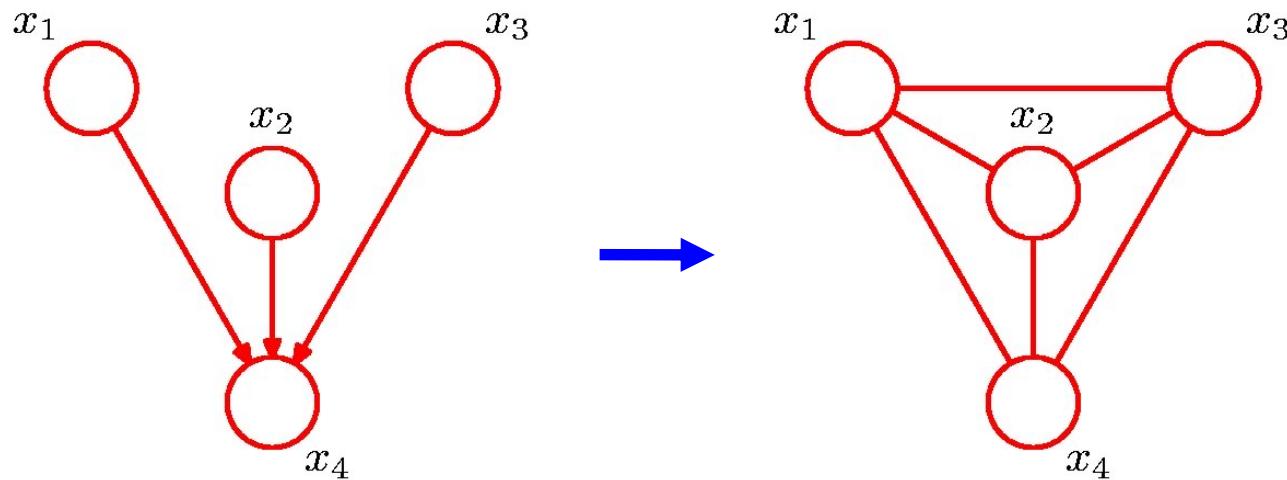


$$p(\mathbf{x}) = \underbrace{p(x_1)p(x_2|x_1) p(x_3|x_2) \cdots p(x_N|x_{N-1})}_{\text{Conditional probabilities}}$$

$$p(\mathbf{x}) = \frac{1}{Z} \psi_{1,2}(x_1, x_2) \psi_{2,3}(x_2, x_3) \cdots \psi_{N-1,N}(x_{N-1}, x_N)$$



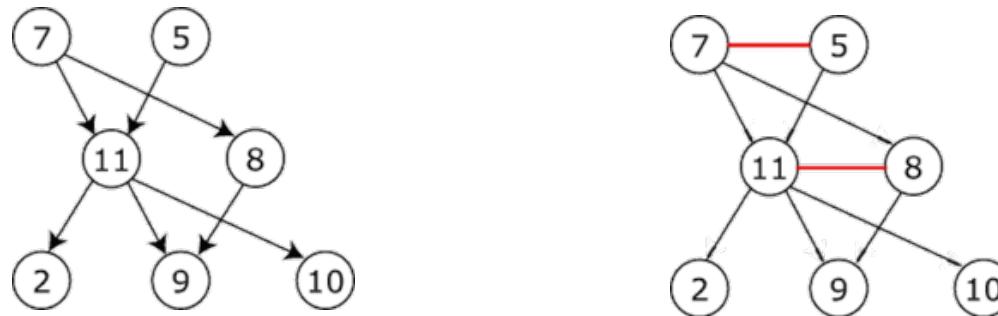
有向图和无向图的转换



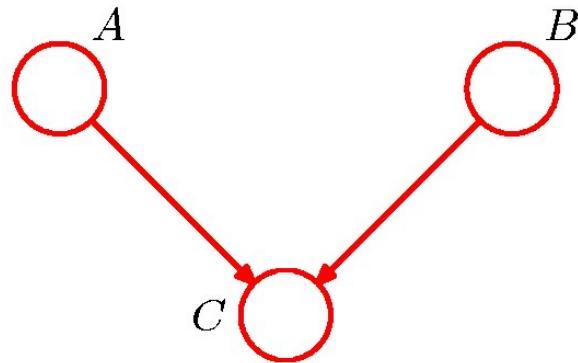
$$\begin{aligned} p(\mathbf{x}) &= p(x_1)p(x_2)p(x_3)p(x_4|x_1, x_2, x_3) \\ &= \frac{1}{Z}\psi_A(x_1, x_2, x_3)\psi_B(x_2, x_3, x_4)\psi_C(x_1, x_2, x_4) \end{aligned}$$

道德图 (Moral Graph)

► 首先完全连接每个结点的父母，然后删除图中箭头的方向。道德化通过将父母连接在一起“联姻”父母。这种“父结点之间结婚”的过程被称为道德化或者伦理化(moralization)，去掉箭头后生成的无向图被称为道德图(moral graph)。很重要的一点是，在这个例子中道德图是完全链接的，因此没有表达出条件独立性质。

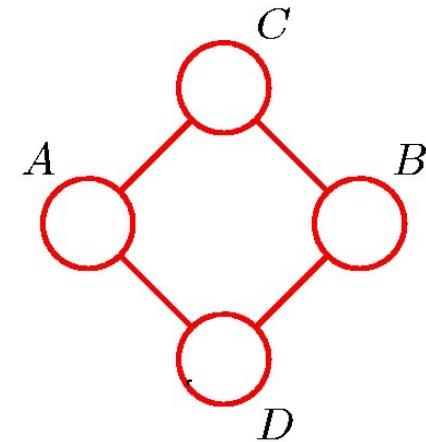


有向图和无向图



$$A \perp\!\!\!\perp B \mid \emptyset$$

$$A \not\perp\!\!\!\perp B \mid C$$



$$A \not\perp\!\!\!\perp B \mid \emptyset$$

$$A \perp\!\!\!\perp B \mid C \cup D$$

$$C \perp\!\!\!\perp D \mid A \cup B$$

无向图模型

- ▶ 无向图模型的联合分布可以表示为

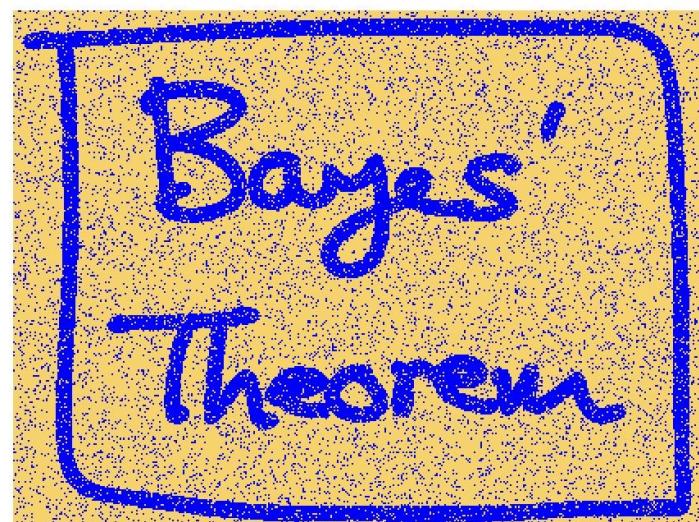
$$\begin{aligned} P(\mathbf{X} = \mathbf{x}) &= \frac{1}{Z} \prod_{c \in \mathcal{C}} \exp(-E(X_c)) \\ &= \frac{1}{Z} \exp\left(\sum_{c \in \mathcal{C}} -E(X_c)\right) \end{aligned}$$

- ▶ 其中 $E(X_c)$ 为能量函数， Z 是配分函数。

Illustration: Image De-Noising (1)

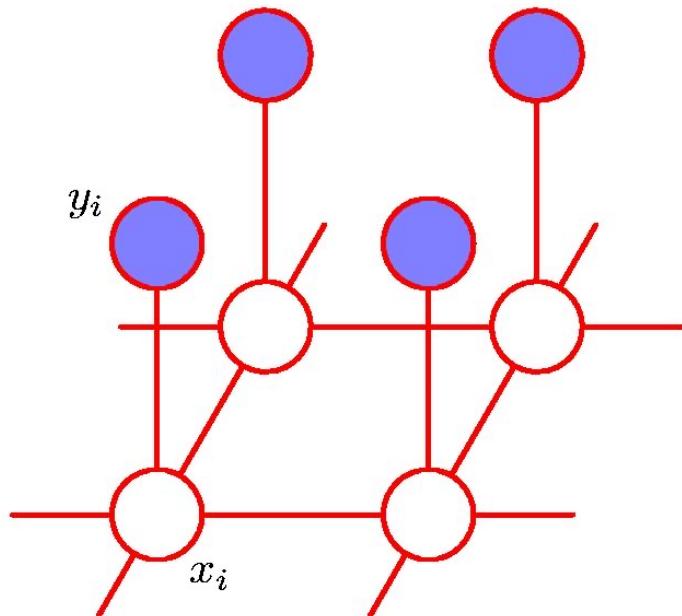


Original Image



Noisy Image

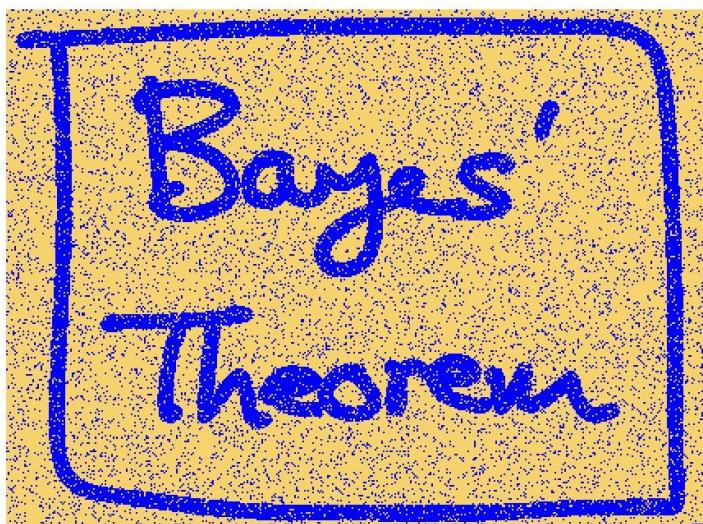
Illustration: Image De-Noising (2)



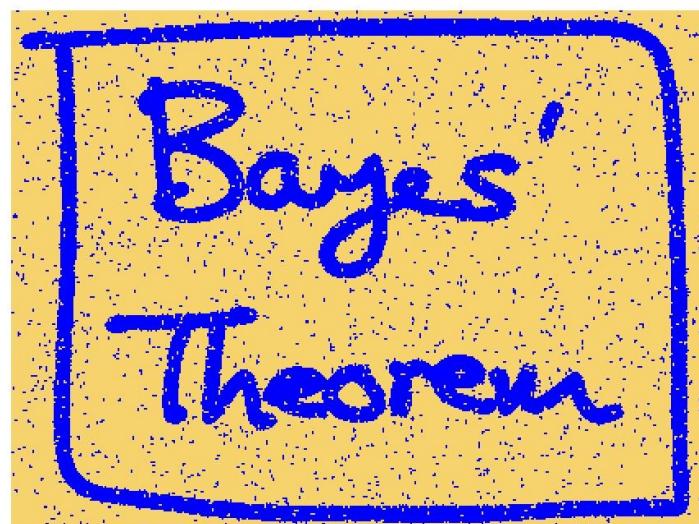
$$E(\mathbf{x}, \mathbf{y}) = h \sum_i x_i - \beta \sum_{\{i,j\}} x_i x_j - \eta \sum_i x_i y_i$$

$$p(\mathbf{x}, \mathbf{y}) = \frac{1}{Z} \exp\{-E(\mathbf{x}, \mathbf{y})\}$$

Illustration: Image De-Noising (2)



Noisy Image



Restored Image (ICM)

常见的无向图模型

► 对数线性模型

► 势能函数的一般定义为

$$\phi_c(\mathbf{x}_c|\theta_c) = \exp\left(\theta_c^T f_c(\mathbf{x}_c)\right)$$

► 联合概率 $p(\mathbf{x})$ 的对数形式为

$$\log p(\mathbf{x}|\theta) = \sum_{c \in C} \theta_c^T f_c(\mathbf{x}_c) - \log Z(\theta)$$

► 也称为最大熵模型

► 条件随机场

► \mathbf{y} 一般为随机向量

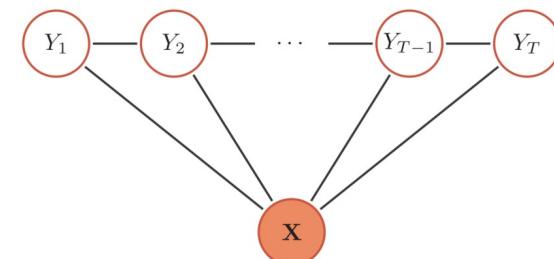
► 条件概率 $p(\mathbf{y}|\mathbf{x})$

$$p(\mathbf{y}|\mathbf{x}, \theta) = \frac{1}{Z(\mathbf{x}, \theta)} \exp\left(\sum_{c \in C} \theta_c^T f_c(\mathbf{x}, \mathbf{y}_c)\right)$$

$$p(\mathbf{y}|\mathbf{x}; \theta) = \frac{1}{Z(\mathbf{x}; \theta)} \exp\left(\sum_{t=1}^T \theta_1^T f_1(\mathbf{x}, y_t) + \sum_{t=1}^{T-1} \theta_2^T f_2(\mathbf{x}, y_t, y_{t+1})\right),$$

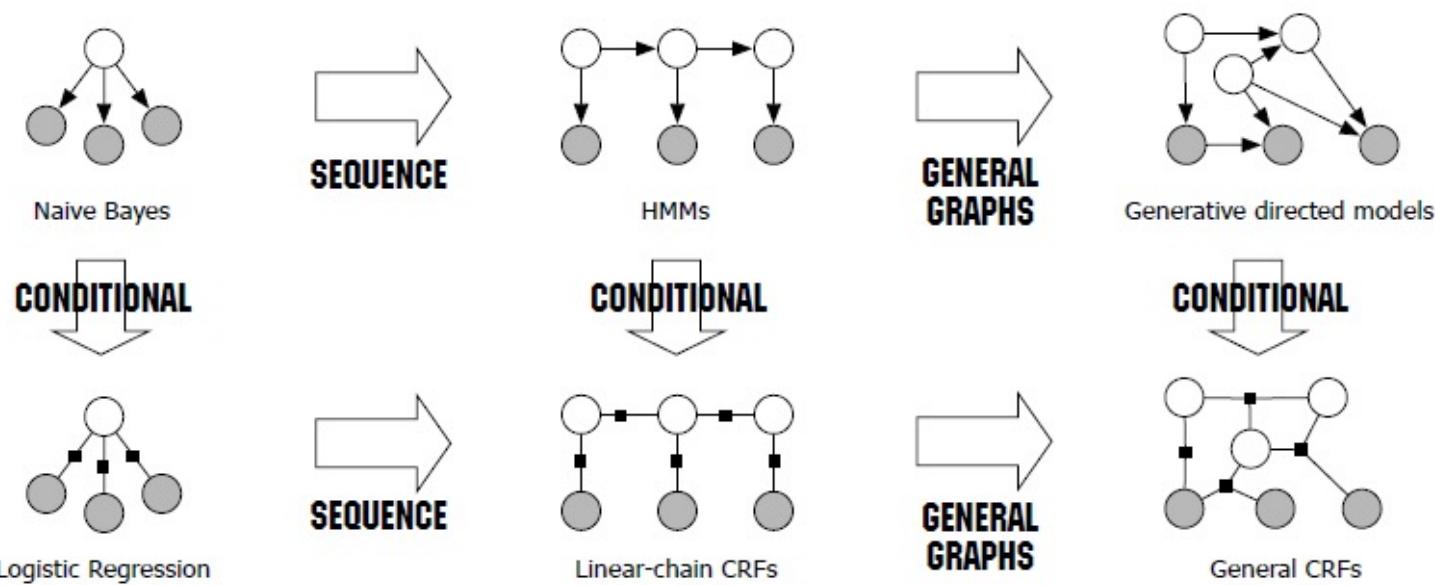


(a) 最大熵模型

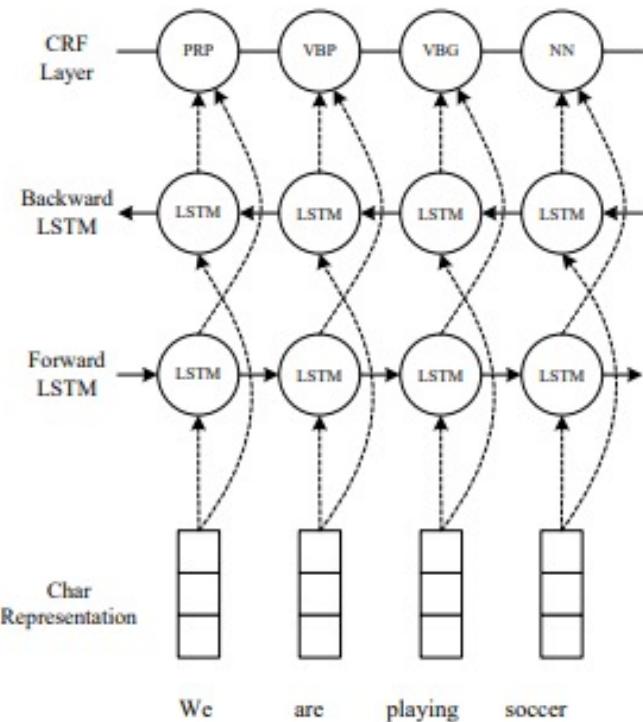


(b) 线性链的条件随机场

模型对比



LSTM-CRF for Sequence Labeling



$$p(\mathbf{y}|\mathbf{z}; \mathbf{W}, \mathbf{b}) = \frac{\prod_{i=1}^n \psi_i(y_{i-1}, y_i, \mathbf{z})}{\sum_{y' \in \mathcal{Y}(\mathbf{z})} \prod_{i=1}^n \psi_i(y'_{i-1}, y'_i, \mathbf{z})}$$

$$\psi_i(y', y, \mathbf{z}) = \exp(\mathbf{W}_{y', y}^T \mathbf{z}_i + \mathbf{b}_{y', y})$$

Xuezhe Ma, Eduard Hovy. End-to-end Sequence labeling via Bi-directional LSTM-CNNs-CRF. ACL 2016



有向图模型

- ▶ 在贝叶斯网络中，所有变量 x 的联合概率分布可以分解为每个随机变量 x_k 的局部条件概率的连乘形式。
- ▶ 假设每个局部条件概率 $p(x_k | x_{\pi_k})$ 的参数为 θ_k ，则对数似然函数为

$$\begin{aligned}\mathcal{L}(\mathcal{D}; \theta) &= \frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}^{(n)}; \theta) \\ &= \frac{1}{N} \sum_{n=1}^N \sum_{k=1}^K \log p(x_k^{(n)} | x_{\pi_k}^{(n)}; \theta_k),\end{aligned}$$

- ▶ 分别最大化每个变量的条件似然来估计其参数

$$\theta_k = \arg \max \sum_{n=1}^N \log p(x_k^{(n)} | x_{\pi_k}^{(n)}; \theta_k)$$

无向图模型

► 以对数线性模型为例,

$$p(\mathbf{x}; \theta) = \frac{1}{Z(\theta)} \exp\left(\sum_{c \in \mathcal{C}} \theta_c^\top f_c(\mathbf{x}_c)\right),$$

给定 N 个训练样本 $\mathcal{D} = \{\mathbf{x}^{(n)}\}_{n=1}^N$, 其对数似然函数为

$$\begin{aligned}\mathcal{L}(\mathcal{D}; \theta) &= \frac{1}{N} \sum_{n=1}^N \log p(\mathbf{x}^{(n)}; \theta) \\ &= \frac{1}{N} \sum_{n=1}^N \left(\sum_{c \in \mathcal{C}} \theta_c^\top f_c(\mathbf{x}_c^{(n)}) \right) - \log Z(\theta),\end{aligned}$$

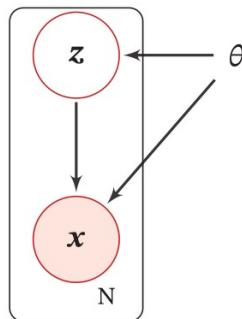
► 偏导数

$$\begin{aligned}\frac{\partial \mathcal{L}(\mathcal{D}; \theta)}{\partial \theta_c} &= \frac{1}{N} \sum_{n=1}^N f_c(\mathbf{x}_c^{(n)}) - \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}; \theta)} [f_c(\mathbf{x}_c)] \\ &= \mathbb{E}_{\mathbf{x} \sim \tilde{p}(\mathbf{x})} [f_c(\mathbf{x}_c)] - \mathbb{E}_{\mathbf{x} \sim p(\mathbf{x}; \theta)} [f_c(\mathbf{x}_c)],\end{aligned}$$



含隐变量的参数学习

- ▶ 隐变量即变量是不可观测的



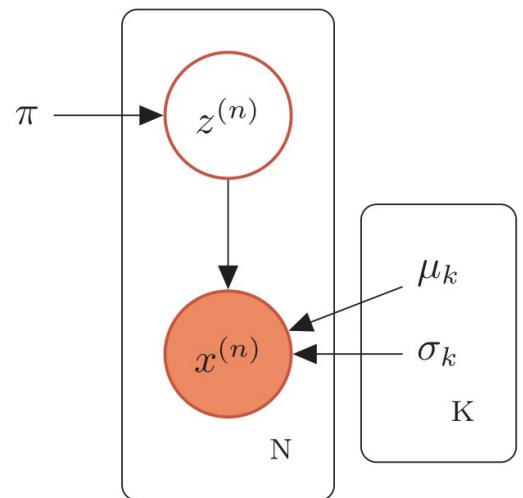
- ▶ 边际似然函数 (Marginal Likelihood)

$$p(\mathbf{x}; \theta) = \sum_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}; \theta)$$

- ▶ 需要用EM算法进行参数估计

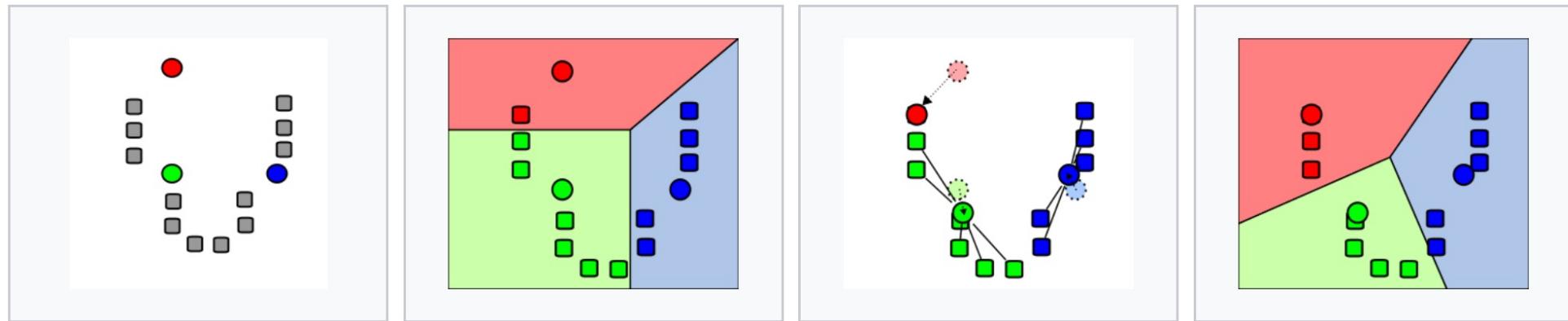
高斯混合模型

► 图模型表示



$$\mathcal{N}(x|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right)$$

一个简单的解法 : K-means



1. k initial "means" (in this case $k=3$) are randomly generated within the data domain (shown in color).
2. k clusters are created by associating every observation with the nearest mean. The partitions here represent the **Voronoi diagram** generated by the means.
3. The **centroid** of each of the k clusters becomes the new mean.
4. Steps 2 and 3 are repeated until convergence has been reached.

K-means算法

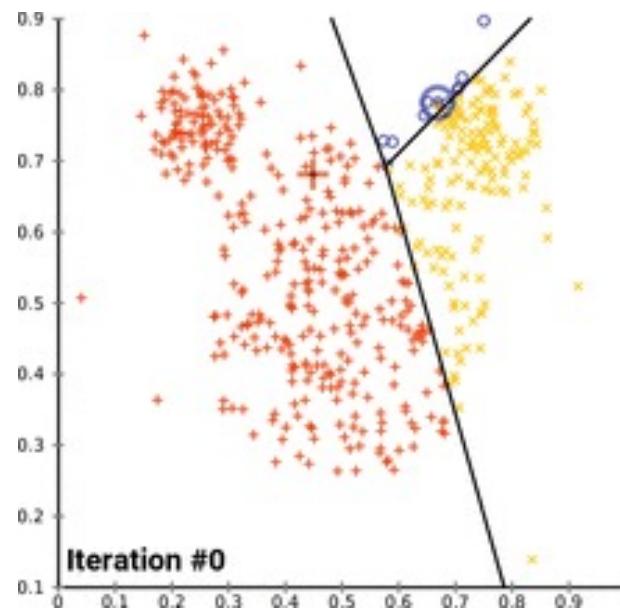
- ▶ 初始化中心点 $m_1^{(1)}, \dots, m_k^{(1)}$
- ▶ 迭代执行下面两步
 - ▶ 分配步 (Assignment step) :

$$S_i^{(t)} = \{x_p : \|x_p - m_i^{(t)}\|^2 \leq \|x_p - m_j^{(t)}\|^2 \forall j, 1 \leq j \leq k\},$$

- ▶ 更新步 (Update step)

$$m_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$$

K-means算法



期望最大化 (Expectation-Maximum, EM) 算法

► 假设有一组变量，有部分变量是不可观测的，如何进行参数估计呢？

D.2.7.1 Jensen 不等式

如果 X 是随机变量, g 是凸函数, 则

$$g(\mathbb{E}[X]) \leq \mathbb{E}[g(X)]. \quad (\text{D.41})$$

等式当且仅当 X 是一个常数或 g 是线性时成立, 这个性质称为 Jensen 不等式.

特别地, 对于凸函数 g 定义域上的任意两点 x_1, x_2 和一个标量 $\lambda \in [0, 1]$, 有

$$g(\lambda x_1 + (1 - \lambda)x_2) \leq \lambda g(x_1) + (1 - \lambda)g(x_2), \quad (\text{D.42})$$

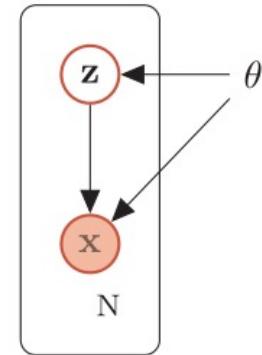
即凸函数 g 上的任意两点的连线位于这两点之间函数曲线的上方.

变分

对数边际似然函数

$$\begin{aligned} \log p(\mathbf{x}; \theta) &= \log \sum_{\mathbf{z}} q(\mathbf{z}) \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} \\ &\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} \\ &\triangleq ELBO(q, \mathbf{x}; \theta), \end{aligned}$$

证据下界



期望最大化 (Expectation-Maximum , EM) 算法

► 假设有一组变量，有部分变量是不可观测的，如何进行参数估计呢？

对数边际似然函数

$$\begin{aligned}\log p(x; \theta) &= \log E_{z \sim q(z)}\left(\frac{p(z, x; \theta)}{q(z)}\right) \\ &\geq E_{z \sim q(z)} \log\left(\frac{p(z, x; \theta)}{q(z)}\right) \\ &= \text{ELBO}\end{aligned}$$

$$\begin{aligned}\frac{p(z, x; \theta)}{q(z)} &= C \\ p(x, z; \theta) &= Cq(z) \\ \sum_z p(x, z; \theta) &= p(x; \theta) = C \\ \frac{p(z, x; \theta)}{q(z)} &= p(x; \theta) \\ q(z) &= \frac{p(z, x; \theta)}{p(x; \theta)} = p(z|x; \theta)\end{aligned}$$

另外一种推导

$$\begin{aligned}\log p(\mathbf{x}; \theta) &= \sum_{\mathbf{z}} q(\mathbf{z}) \log p(\mathbf{x}; \theta) \\&= \sum_{\mathbf{z}} q(\mathbf{z}) \left(\log p(\mathbf{x}, \mathbf{z}; \theta) - \log p(\mathbf{z}|\mathbf{x}; \theta) \right) \\&= \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{x}, \mathbf{z}; \theta)}{q(\mathbf{z})} - \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{z}|\mathbf{x}; \theta)}{q(\mathbf{z})} \\&= ELBO(q, \mathbf{x}; \theta) + KL(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x}; \theta)),\end{aligned}$$

EM算法

$$ELBO(q, \mathbf{x}; \theta) + KL(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x}; \theta))$$

► E步

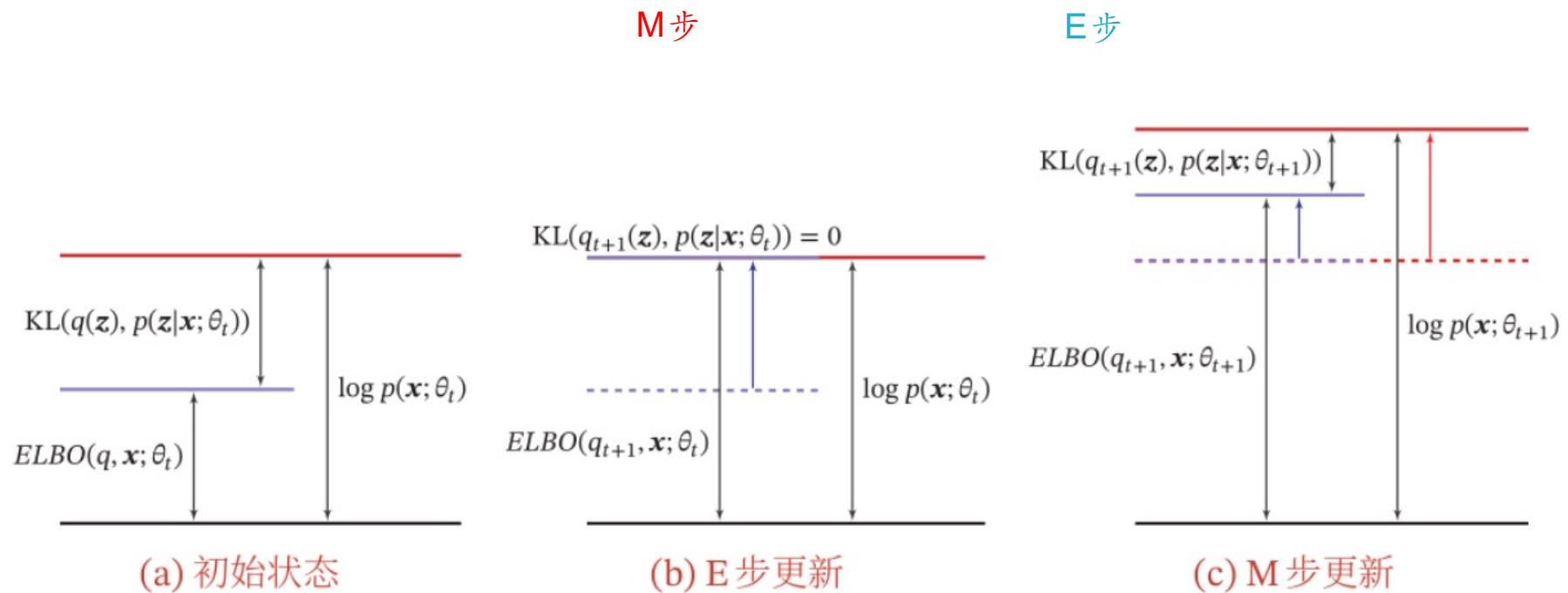
$$q_{t+1}(\mathbf{z}) = \arg \max_q ELBO(q, \mathbf{x}; \theta) + KL(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x}; \theta))$$

► M步

$$\theta_{t+1} = \arg \max_{\theta} ELBO(q, \mathbf{x}; \theta) + KL(q(\mathbf{z}) || p(\mathbf{z}|\mathbf{x}; \theta))$$

收敛性

$$\log p(\mathbf{x}; \theta_{t+1}) \geq ELBO(q_{t+1}, \mathbf{x}; \theta_{t+1}) \geq ELBO(q_{t+1}, \mathbf{x}; \theta_t) = \log p(\mathbf{x}; \theta_t)$$

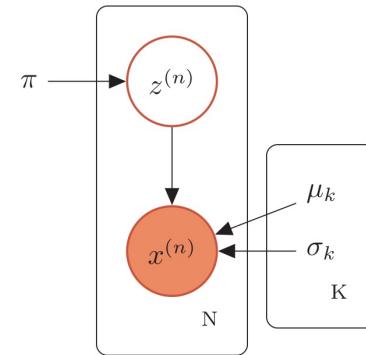


GMM Revisit

E步 先固定参数 μ, σ , 计算后验分布 $p(z^{(n)}|x^{(n)})$

$$\begin{aligned}\gamma_{nk} &\triangleq p(z^{(n)} = k|x^{(n)}) \\ &= \frac{p(z^{(n)})p(x^{(n)}|z^{(n)})}{p(x^{(n)})} \\ &= \frac{\pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k)}{\sum_{k=1}^K \pi_k \mathcal{N}(x^{(n)}|\mu_k, \sigma_k)},\end{aligned}$$

其中 γ_{nk} 定义了样本 $x^{(n)}$ 属于第 k 个高斯分布的后验概率。



$$\mathcal{N}(x|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x - \mu_k)^2}{2\sigma_k^2}\right)$$

GMM Revisit

M步 令 $q(z = k) = \gamma_{nk}$, 训练集 \mathcal{D} 的证据下界为

$$\begin{aligned} ELBO(\gamma, \mathcal{D} | \pi, \mu, \sigma) &= \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \log \frac{p(x^{(n)}, z^{(n)} = k)}{\gamma_{nk}} \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \left(\log \mathcal{N}(x^{(n)} | \mu_k, \sigma_k) + \log \frac{\pi_k}{\gamma_{nk}} \right) \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \left(\frac{-(x - \mu_k)^2}{2\sigma_k^2} - \log \sigma_k + \log \pi_k \right) + C, \end{aligned}$$

其中 C 为和参数无关的常数。

$$\text{s.t. } \sum_{k=1}^K \pi_k = 1.$$

$$\mathcal{N}(x | \mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp \left(-\frac{(x - \mu_k)^2}{2\sigma_k^2} \right)$$

$$p(x^n, z^n = k) = \pi_k \mathcal{N}(x^n | \mu_k, \sigma_k)$$

$$\log \left(\frac{1}{\sqrt{2\pi}\sigma_k} \exp \left(-\frac{(x - \mu_k)^2}{2\sigma_k^2} \right) \right) = \frac{-(x - \mu_k)^2}{2\sigma_k^2} - \log \sigma_k - \log \sqrt{2\pi}$$

$$C = \sum_{n=1}^N \sum_{k=1}^K (-\log \gamma_{nk} - \log \sqrt{2\pi}) * \gamma_{nk}$$

GMM Revisit

M步 令 $q(z = k) = \gamma_{nk}$, 训练集 \mathcal{D} 的证据下界为

$$\begin{aligned} ELBO(\gamma, \mathcal{D} | \pi, \mu, \sigma) &= \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \log \frac{p(x^{(n)}, z^{(n)} = k)}{\gamma_{nk}} \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \left(\log \mathcal{N}(x^{(n)} | \mu_k, \sigma_k) + \log \frac{\pi_k}{\gamma_{nk}} \right) \\ &= \sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \left(\frac{-(x - \mu_k)^2}{2\sigma_k^2} - \log \sigma_k + \log \pi_k \right) + C, \end{aligned}$$

其中 C 为和参数无关的常数。

$$\text{s.t. } \sum_{k=1}^K \pi_k = 1.$$

$$ELBO(\gamma, \mathcal{D} | \pi, \mu, \sigma) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right)$$

$$\sum_{n=1}^N \sum_{k=1}^K \gamma_{nk} \left(\frac{-(x - \mu_k)^2}{2\sigma_k^2} - \log \sigma_k + \log \pi_k \right) + \lambda \left(\sum_{k=1}^K \pi_k - 1 \right) + C$$

$$\frac{\partial L}{\partial \pi_k} = \sum_{n=1}^N \gamma_{nk} \frac{1}{\pi_k} + \lambda = 0$$

$$\frac{\partial L}{\partial \mu_k} = \sum_{n=1}^N \gamma_{nk} \frac{(x^n - \mu_k)}{\sigma_k^2} = 0$$

$$\frac{\partial L}{\partial \sigma_k} = \sum_{n=1}^N \gamma_{nk} \left(\frac{(x - \mu_k)^2}{\sigma_k^3} - \frac{1}{\sigma_k} \right) = 0$$

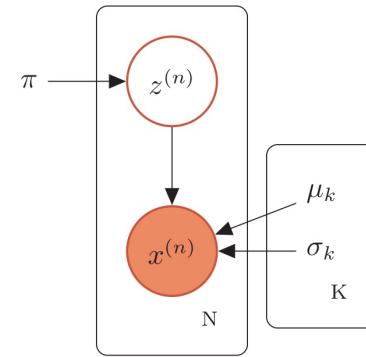
GMM Revisit

$$N_k = \sum_{n=1}^N \gamma_{nk}.$$

$$\pi_k = \frac{N_k}{N},$$

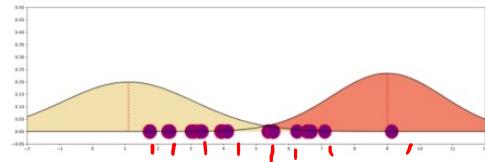
$$\mu_k = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} x^{(n)},$$

$$\sigma_k^2 = \frac{1}{N_k} \sum_{n=1}^N \gamma_{nk} (x^{(n)} - \mu_k)^2,$$

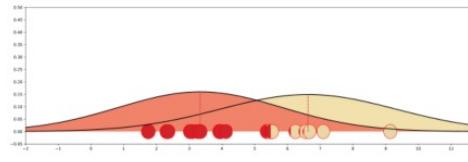


$$\mathcal{N}(x|\mu_k, \sigma_k) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{(x-\mu_k)^2}{2\sigma_k^2}\right)$$

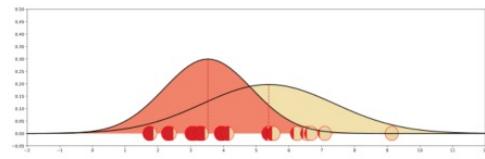
GMM的参数学习



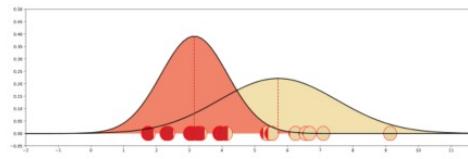
(a) 初始化



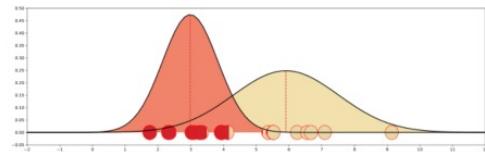
(b) 第1次迭代



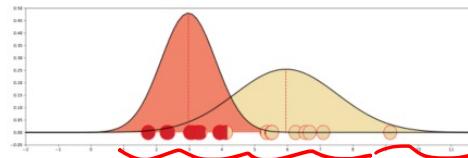
(c) 第4次迭代



(d) 第8次迭代

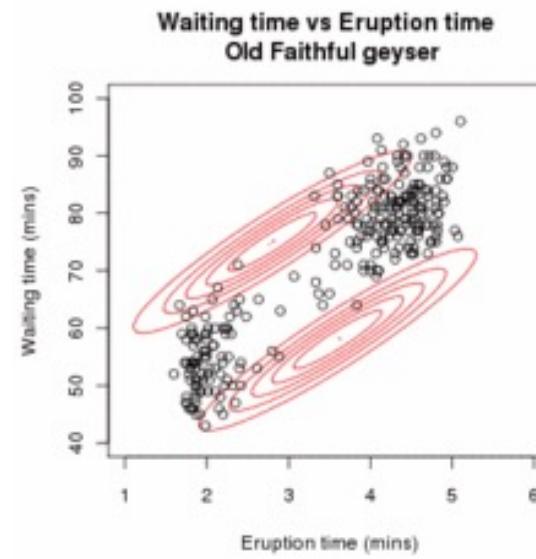
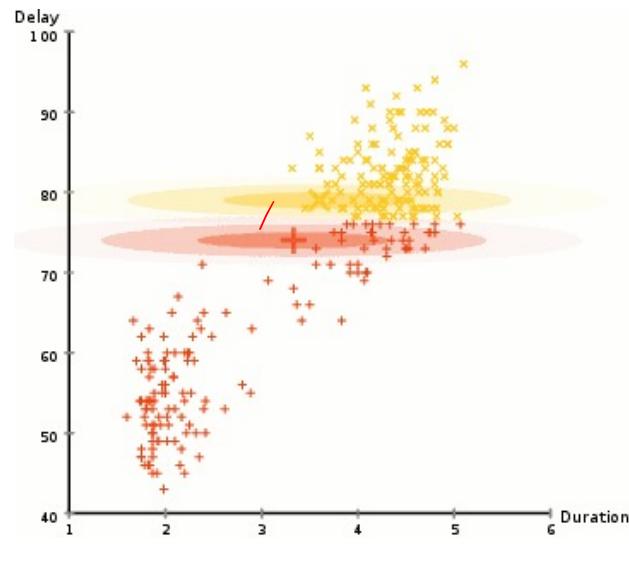


(e) 第12次迭代



(f) 第16次迭代

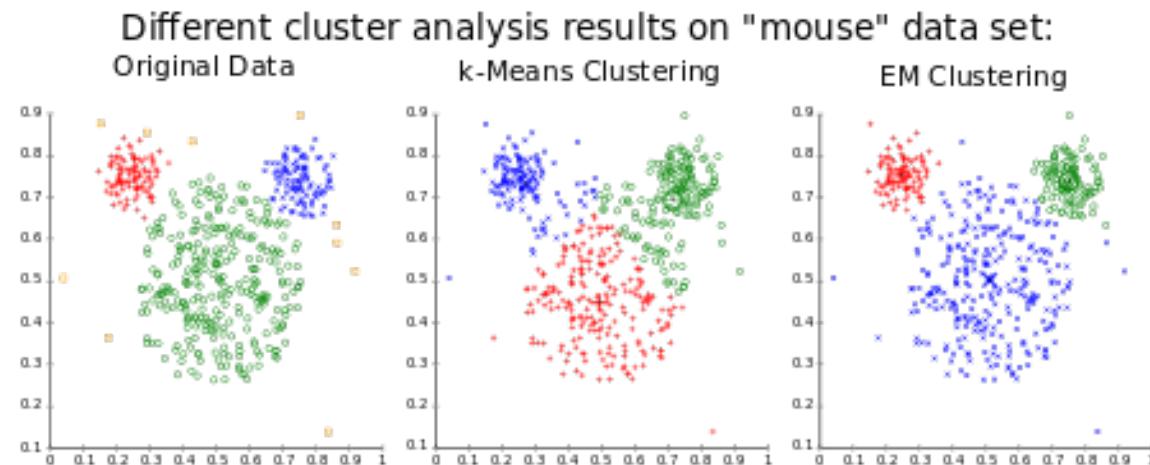
GMM的参数学习



k-means clustering vs. EM clustering

- 随机确定k中心点
 - 计算每个样本点到中心点的距离
 - 把每个样本点分配到离它最近的中心点
 - 对每一个中心点，根据分配到的样本点重新计算中心点位置
 - 迭代直到收敛
-
- 随机初始化k个高斯分布的均值和方差
 - 计算每个样本点属于每个高斯分布的后验概率
 - 用后验概率重新估计每个高斯分布的均值和方差
 - 迭代直到收敛

k-means clustering vs. EM clustering



图模型与神经网络的关系

1. 图模型的节点是随机变量，图结构主要描述变量之间的依赖关系；而神经网络中的节点是神经元，是一个计算节点。
2. 图模型中每个变量一般有着明确的解释，变量之间依赖关系一般是人工来定义；而神经网络中单个神经元没有直观的解释。
3. 神经网络是判别式模型，直接用来分类；而图模型既可以是判别式模型，也可以是生成式模型。
4. 神经网络和概率图的结合越来越紧密。



谢 谢

<https://nndl.github.io/>