提纲

- **概率模型学习**
  - 统计学习
  - 完全数据学习
  - 隐变量学习

贝叶斯学习

- 假设有五种糖果袋:

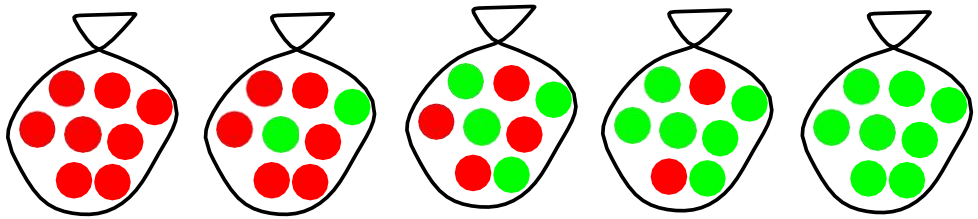
  10% 为 h_1: 100% 樱桃味

  20% 为 h_2: 75% 樱桃味 + 25% 酸橙味

  40% 为 h_3: 50% 樱桃味 + 50% 酸橙味

  20% 为 h_4: 25% 樱桃味 + 75% 酸橙味

  10% 为 h_5: 100% 酸橙味



- 然后我们观测从某个袋子中抽取的糖果:

- 这是什么种类的袋子? 下一块糖果是什么口味?

贝叶斯学习

- 将学习看作假设空间中概率分布的贝叶斯更新：

  H 是假设变量，值为 $h_1, h_2, \ldots$，先验分布 $\mathrm{P}(H)$

  第j个观测 d_j给出了随机变量的输出 D_j
  训练数据 $\mathrm{d} = d_1, \ldots, d_N$

- 给定到目前为止的数据，每一个假设有一个后验分布：

  $P(h_i|\mathrm{d}) = aP(\mathrm{d}|h_i)P(h_i)$

  这里 $P(\mathrm{d}|h_i)$ 被称为**似然**

- 预测为在假设上的概率加权平均：

  $\mathrm{P}(X|\mathrm{d}) = \Sigma_i\ \mathrm{P}(X|\mathrm{d}, h_i)P(h_i|\mathrm{d}) = \Sigma_i\ \mathrm{P}(X|h_i)P(h_i|\mathrm{d})$

贝叶斯学习

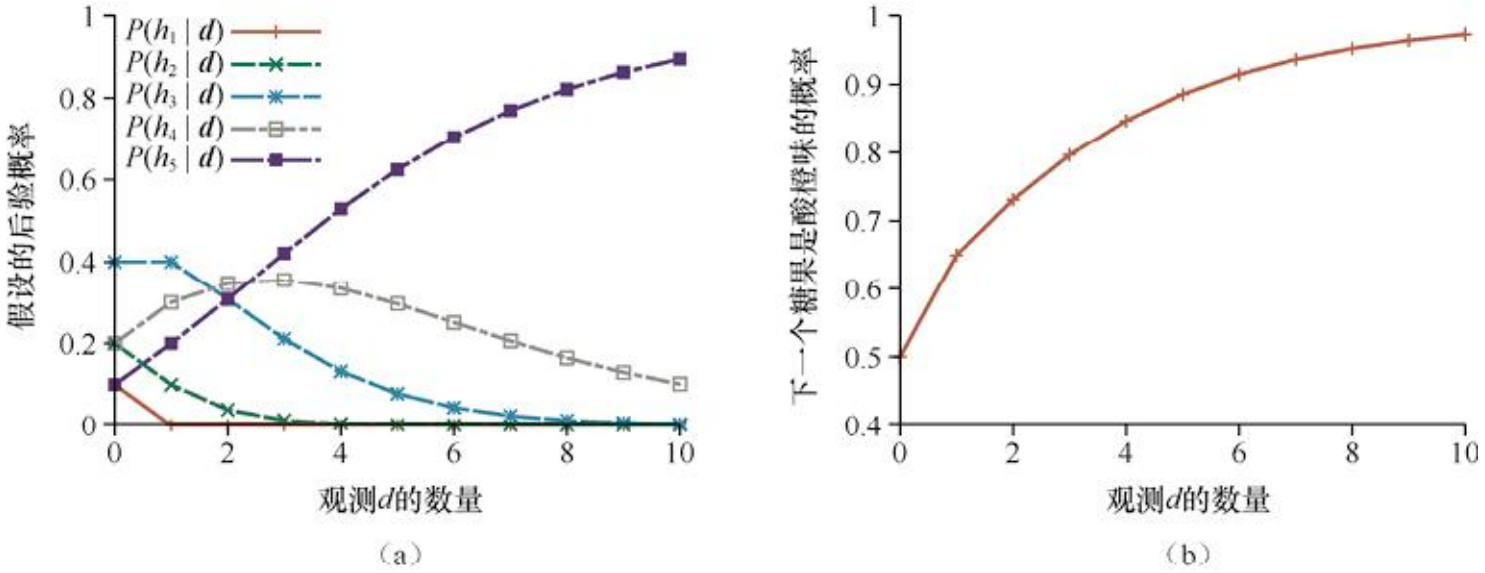


图20-1 (a) 根据式 (20-1) 得到的后验概率$P(h_i \mid d_1, \cdots, d_N)$。观测数量$N$为1～10，且每一个观测都是酸橙味的糖果。 (b) 基于式 (20-2) 的贝叶斯预测$P(D_{N+1} = \text{lime} \mid d_1, \cdots, d_N)$

$$P(\boldsymbol{d} \mid h_i) = \prod_j P(d_j \mid h_i) \quad P(h_i \mid \boldsymbol{d}) = \alpha P(\boldsymbol{d} \mid h_i) P(h_i)$$

最大后验学习

- 在假设空间上求和通常是非常困难的 $\Sigma_i \ \mathrm{P}(X|h_i)P(h_i|\mathrm{d})$
- **最大后验** (MAP)学习：选择 h_{MAP} 来最大化 $P(h_i|\mathrm{d})$

即, 最大化 $P(\mathrm{d}|h_i)P(h_i)$ 或 $\log P(\mathrm{d}|h_i) + \log P(h_i)$

$$\boldsymbol{P}(X \mid \boldsymbol{d}) \approx \boldsymbol{P}(X \mid h_{\mathrm{MAP}})$$

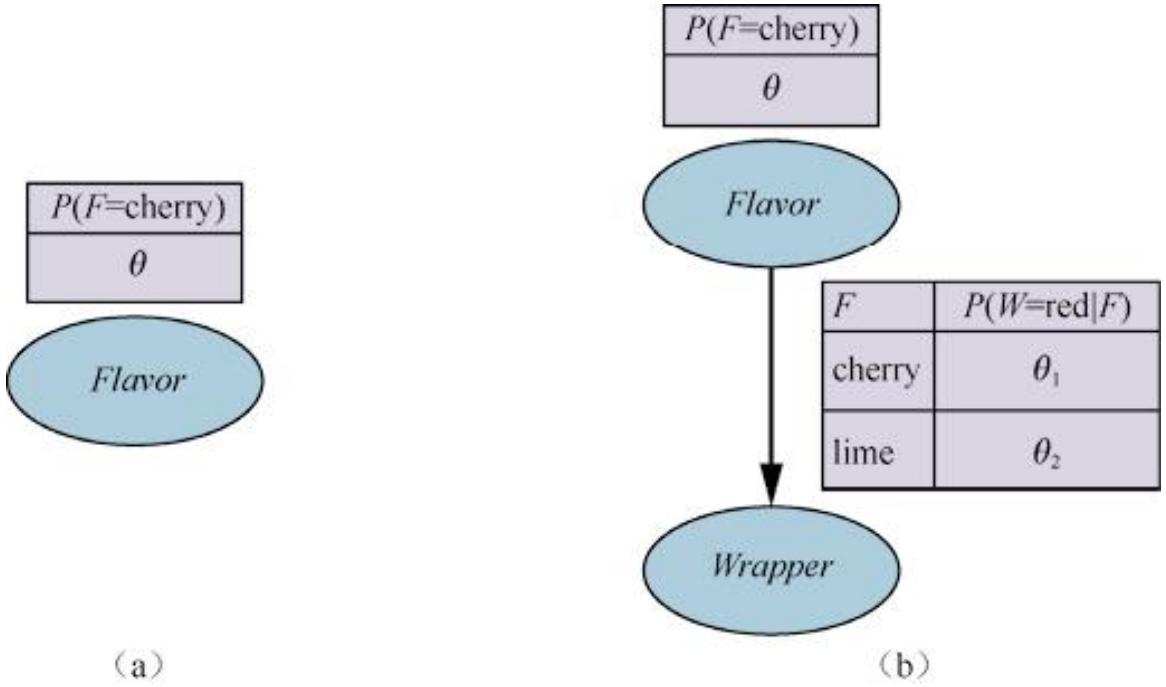
- $-\log_2 P(\boldsymbol{d} \mid h_i) - \log_2 P(h_i)$ 负对数项可以被看作
  - 给定假设编码数据所需的比特数 + 编码假设所需的比特数
  - 这是最小描述长度(MDL)学习的基本思想

最大似然学习

- 当数据集很大时，假设的先验分布就不那么重要了，因为来自数据的证据足够强大，足以淹没假设的先验分布。
- 最大似然 (ML) 学习: 选择 h_{ML} 来最大化 $P(\mathrm{d}|h_i)$
- 即, 简单的获取对数据的最佳拟合；对于假设空间具有均匀先验分布，等同于最大后验学习 (例如所有的假设都同样复杂)
- 最大似然学习是“标准”的（非贝叶斯） 统计学习方法

完全数据学习

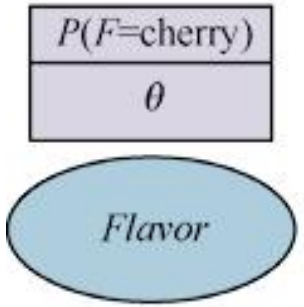
最大似然参数学习：离散模型



（a）樱桃味糖果和酸橙味糖果比例未知情况下的贝叶斯网络。（b）包装颜色（依概率）与糖果口味相关情况下的模型

完全数据学习

最大似然参数学习：离散模型



- 可能含有樱桃味和酸橙味糖果的糖果袋，其中糖果口味的比例完全未知。
- 参数θ表示樱桃味糖果所占的比例，其对应的假设为h_θ
- 如果我们假设所有的比例有相同的先验可能性，那么采用最大似然估计是合理的
- 现在假设我们已经打开了N颗糖果，其中有c颗为樱桃味，则该特定数据集的似然为：

$$P(\boldsymbol{d} \mid h_\theta) = \prod_{j=1}^{N} P(d_j \mid h_\theta) = \theta^c \cdot (1-\theta)^\ell$$

- 最大似然假设所需的参数即为使得上式最大化的参数 。由于log函数是单调函数， 我们可以最大化对数似然来简化计算：

$$L(\boldsymbol{d} \mid h_\theta) = \log P(\boldsymbol{d} \mid h_\theta) = \sum_{j=1}^{N} \log P(d_j \mid h_\theta) = c \log\theta + \ell \log(1-\theta)$$

- 对上式关于θ进行求导，并令导数为0可得：

$$\frac{\mathrm{d}L(\boldsymbol{d} \mid h_\theta)}{\mathrm{d}\theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell} = \frac{c}{N}$$

完全数据学习

最大似然参数学习：离散模型

- 最大似然参数学习的标准流程：

（1）将数据的似然写成关于参数的函数的形式。

（2）计算对数似然关于每个参数的导数。

（3）解出使得导数为0的参数。

- 最大似然学习中普遍存在的一个重要问题：当数据集非常小以至于一些事件还未发生时，最大似然假设将把这些事件的概率置为0。

完全数据学习

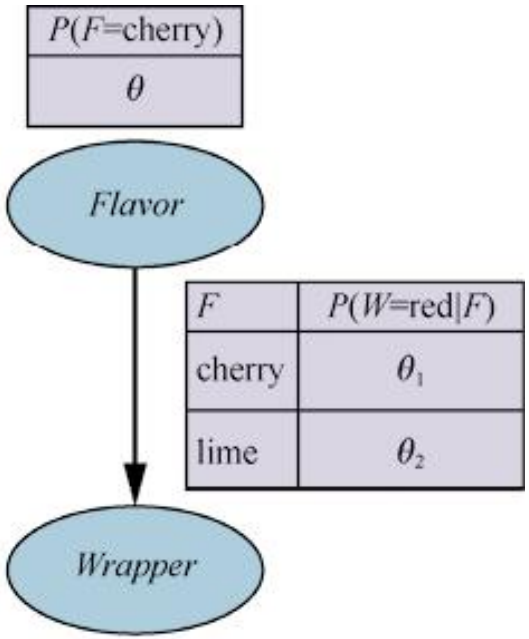
最大似然参数学习：离散模型

- 红、绿两种不同颜色的糖果包装；其包装在概率上服从某个条件分布，该分布取决于糖果的口味。
- 观测到一颗带有绿色包装的樱桃味糖果的似然：

$$P(Flavor = \text{cherry}, Wrapper = green \mid h_{\theta,\theta_1,\theta_2})$$
$$= P(Flavor = \text{cherry} \mid h_{\theta,\theta_1,\theta_2}) P(Wrapper = \text{green} \mid Flavor = \text{cherry}, h_{\theta,\theta_1,\theta_2})$$
$$= \theta \cdot (1-\theta_1)$$

- N颗糖果，r_c颗包装为红色的樱桃味糖果等等，则该数据的似然为：

$$P(\boldsymbol{d} \mid h_{\theta,\theta_1,\theta_2}) = \theta^c (1-\theta)^\ell \cdot \theta_1^{r_c} (1-\theta_1)^{g_c} \cdot \theta_2^{r_\ell} (1-\theta_2)^{g_\ell}$$



完全数据学习

最大似然参数学习：离散模型

- 对似然取对数：

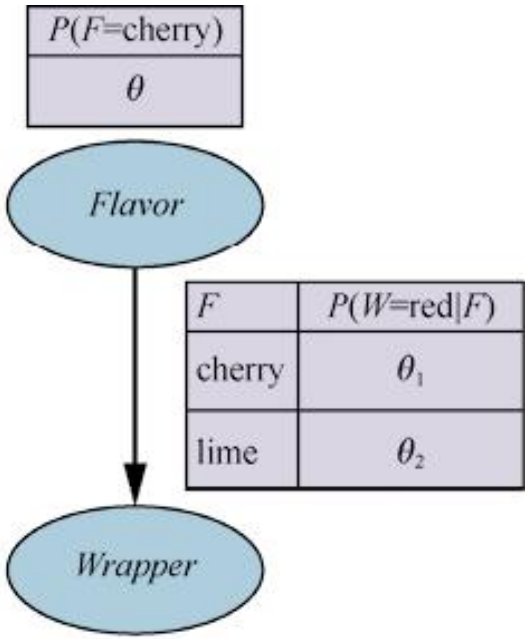
$$L = [c \log \theta + \ell \log(1-\theta)] + [r_c \log \theta_1 + g_c \log(1-\theta_1)] + [r_\ell \log \theta_2 + g_\ell \log(1-\theta_2)]$$

- 对数似然的具体形式是3项求和，其中每一项包含单独的一个参数。
- 令对数似然对每个参数求导并置为0时，可以得到3个独立的方程，其中每一个方程只含有一个参数：

$$\frac{\partial L}{\partial \theta} = \frac{c}{\theta} - \frac{\ell}{1-\theta} = 0 \quad \Rightarrow \quad \theta = \frac{c}{c+\ell}$$

$$\frac{\partial L}{\partial \theta_1} = \frac{r_c}{\theta_1} - \frac{g_c}{1-\theta_1} = 0 \quad \Rightarrow \quad \theta_1 = \frac{r_c}{r_c + g_c}$$

$$\frac{\partial L}{\partial \theta_2} = \frac{r_\ell}{\theta_2} - \frac{g_\ell}{1-\theta_2} = 0 \quad \Rightarrow \quad \theta_2 = \frac{r_\ell}{r_c + g_\ell}$$



- 一旦有了完全数据，贝叶斯网络的最大似然参数学习问题将可以被分解为一些分离的学习问题，每个问题对应一 个参数。

完全数据学习

南京大学 NANJING UNIVERSITY

朴素贝叶斯模型

- 假设变量为布尔变量, 其参数为

$$\theta = P(C = true), \theta_{i1} = P(X_i = true \mid C = true), \theta_{i2} = P(X_i = true \mid C = false).$$

- 当观测到的属性值为 $x_1, \ldots, x_n$, 其属于某一类的概率由下式给出：

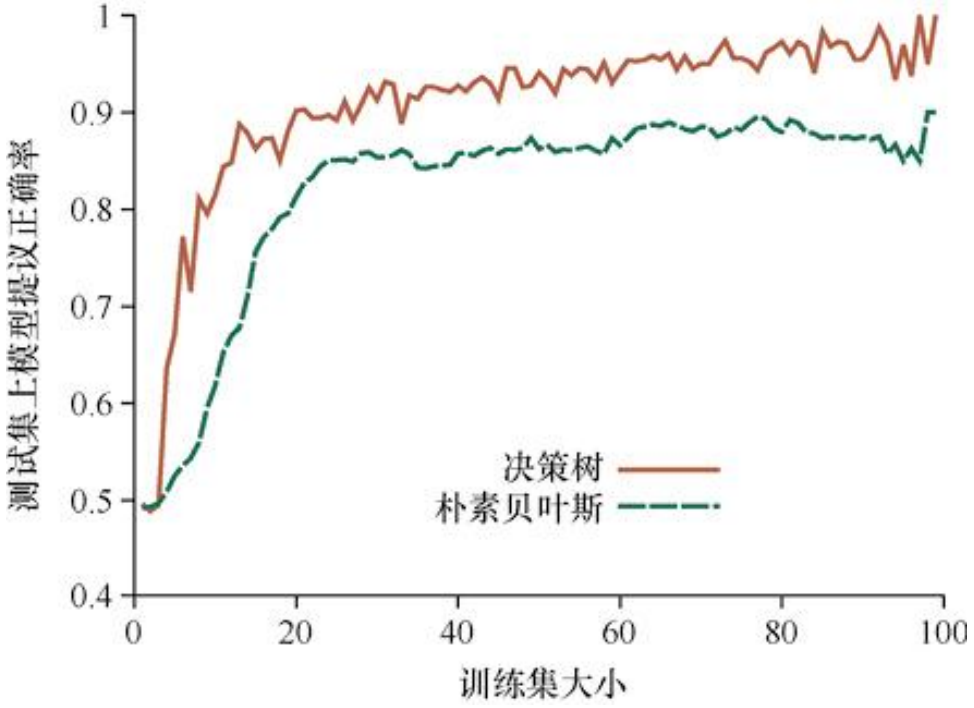
$$\mathbf{P}(C \mid x_1, \ldots, x_n) = \alpha \, \mathbf{P}(C) \prod_i \mathbf{P}(x_i \mid C) \, .$$

通过选择可能性最大的类，我们可以获得一个确定性的预测。

- 朴素贝叶斯可以很好地推广到大规模的问题上：当有n个布尔属性时，我们只需要2n + 1个参数，且不需要任何的搜索就能找到朴素贝叶斯最大似然假设。

完全数据学习

朴素贝叶斯模型



将朴素贝叶斯学习应用于第19章餐厅等待问题得到的学习曲线；决策树的学习曲线也在图中给出，用于比较

完全数据学习

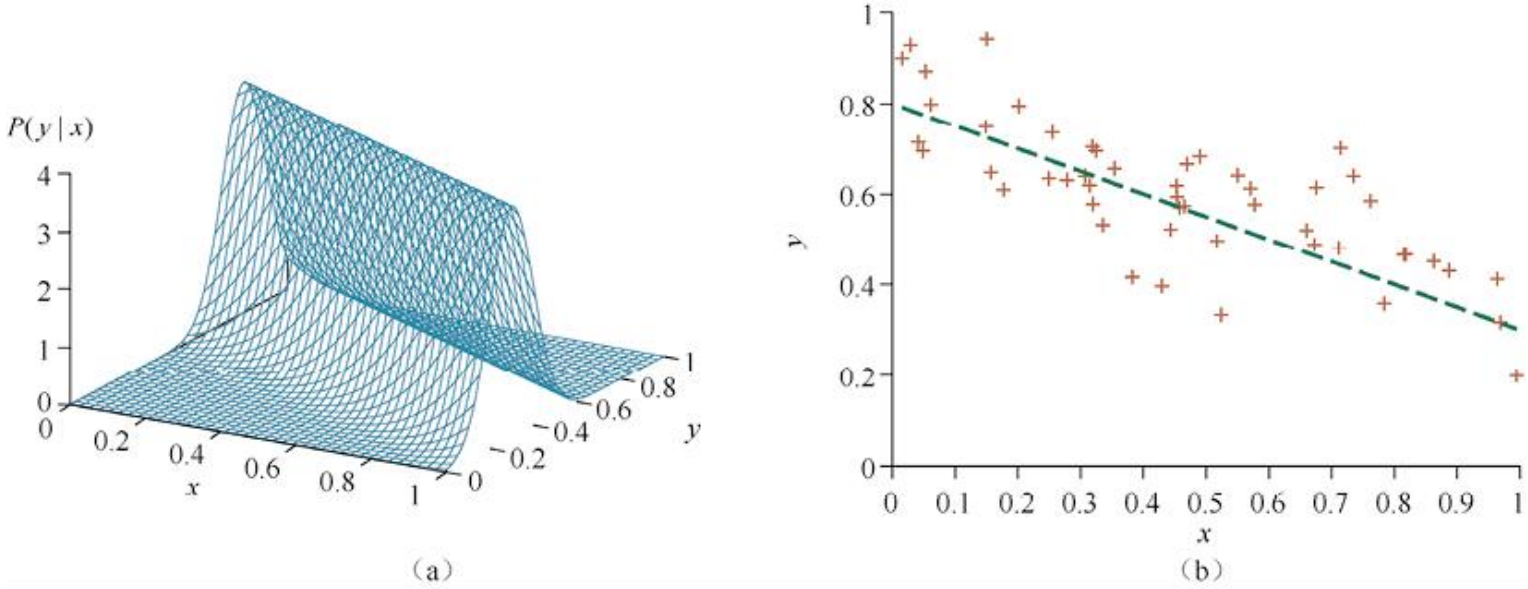
生成模型和判别模型

- 生成模型（generative model）对每一类的概率分布进行建模。每个模型包含该模型对应类的先验以及对应的条件分布 。根据这些，我们可以计算出联合概率 ，并且我们可以随机生成类别相应的特征。

- 判别模型（discriminative model）直接学习类别之间的决策边界 。给定一个输入样例，一个判别模型将会输出一个类别。逻辑斯谛回归、决策树以及支持向量机都是判别模型。

完全数据学习

最大似然参数学习：连续模型



(a) 高斯线性模型，它表述为$y = \theta_1 x + \theta_2$加上固定方差的高斯噪声。 (b) 由该模型生成的50个数据点，以及它的最佳拟合直线

完全数据学习

南京大学 NANJING UNIVERSITY

最大似然参数学习：连续模型

- 学习单变量高斯密度函数的参数。
- 假设数据按如下分布生成：$P(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
- 观测数据的对数似然：

$$L=\sum_{j=1}^{N}\log\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x_j-\mu)^2}{2\sigma^2}}=N(-\log\sqrt{2\pi}-\log\sigma)-\sum_{j=1}^{N}\frac{(x_j-\mu)^2}{2\sigma^2}$$

- 令对数似然导数为0：

$$\frac{\partial L}{\partial\mu}=-\frac{1}{\sigma^2}\sum_{j=1}^{N}(x_j-\mu)=0 \quad\Rightarrow\quad \mu=\frac{\sum_j x_j}{N}$$

$$\frac{\partial L}{\partial\sigma}=-\frac{N}{\sigma}+\frac{1}{\sigma^3}\sum_{j=1}^{N}(x_j-\mu)^2=0 \quad\Rightarrow\quad \sigma=\sqrt{\frac{\sum_j(x_j-\mu)^2}{N}}$$

完全数据学习

最大似然参数学习：连续模型

- 考虑一个线性高斯模型，它有一个连续的父变量X和一个连续的子变量Y，它的条件分布：

$$P(y \mid x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(y-(\theta_1 x+\theta_2))^2}{2\sigma^2}}$$

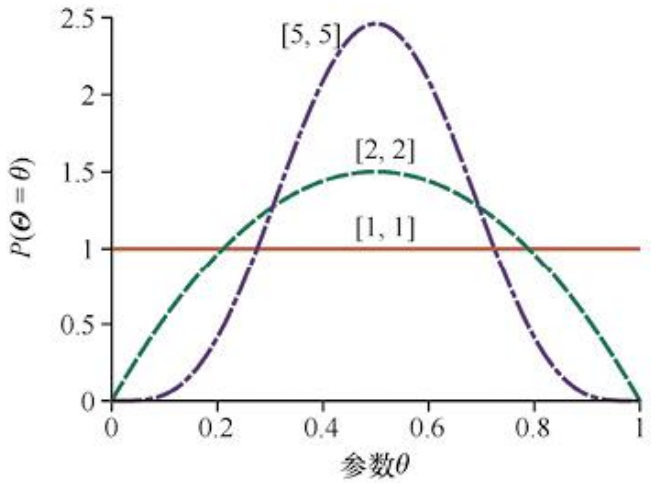
- 运用一般的最大似然学习方法，我们可以找到参数的最大似然值。

- 如果仅考虑定义x和y之间线性关系的参数，那么最大化这些参数的对数似然与最小化条件分布中指数的分子是等价的。

贝叶斯参数学习

- 基于贝叶斯方法的参数学习过程从一个关于假设的先验分布开始，随着新数据出现而不断更新该分布。
- 从贝叶斯角度看，随机变量 Θ 定义了假设空间，θ 是 Θ 的一个未知值。
- 假设先验是先验分布 $\mathbf{P}(\Theta)$。因此, $P(\Theta = \theta)$ 是糖果袋中含有比例θ的樱桃味糖果的先验概率。
- $P(\theta) = Uniform[0,1](\theta)$, 均匀分布是 β分布的一个特例.
- β分布由两个超参数 a 和 b 定义：

$$\text{beta}[a,b](\theta) = \alpha\, \theta^{a-1}(1-\theta)^{b-1}\ ,$$



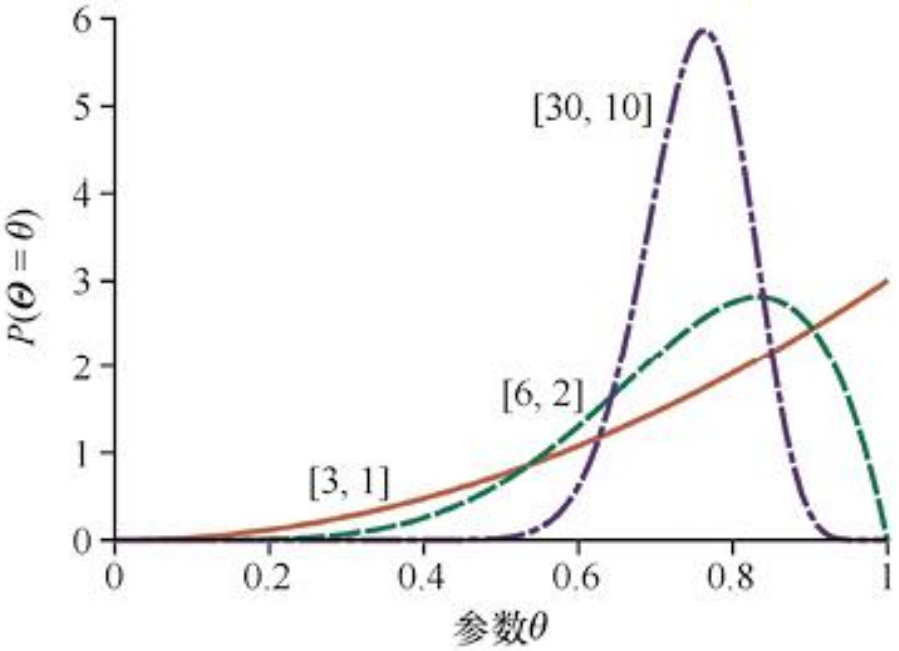
完全数据学习

南京大学 NANJING UNIVERSITY

贝叶斯参数学习

- 假设我们观测到了一颗樱桃味的糖果，那么我们有

$$
\begin{aligned}
P(\theta \mid D_1 = cherry) &= \alpha\, P(D_1 = cherry \mid \theta) P(\theta) \\
&= \alpha'\, \theta \cdot \mathrm{beta}[a, b](\theta) = \alpha'\, \theta \cdot \theta^{a-1}(1-\theta)^{b-1} \\
&= \alpha'\, \theta^{a}(1-\theta)^{b-1} = \mathrm{beta}[a+1, b](\theta)\ .
\end{aligned}
$$



完全数据学习

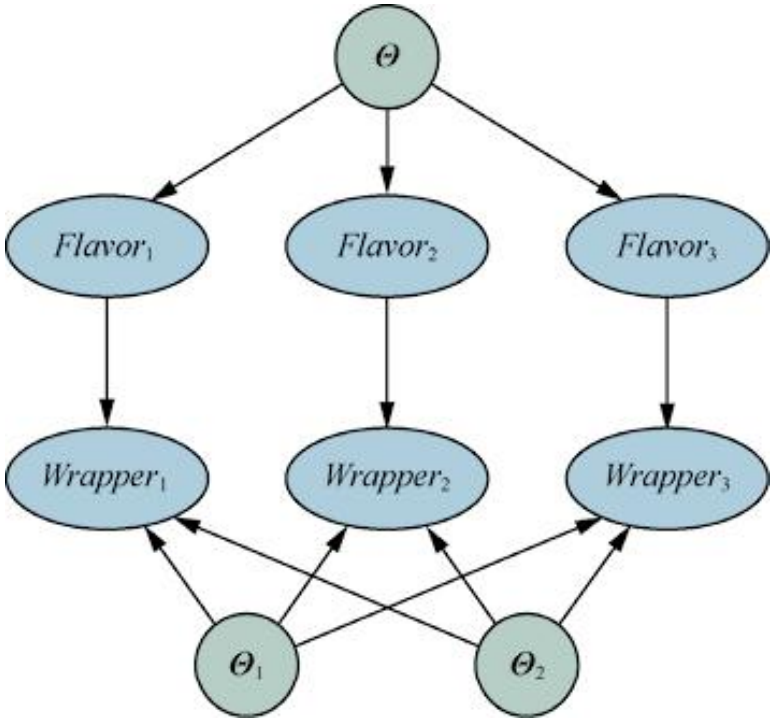
贝叶斯参数学习

- 假设有3个参数，根据参数独立性

$$\mathbf{P}(\Theta, \Theta_1, \Theta_2) = \mathbf{P}(\Theta)\mathbf{P}(\Theta_1)\mathbf{P}(\Theta_2) .$$

$$P(Flavor_i = cherry \mid \Theta = \theta) = \theta .$$

$$P(Wrapper_i = red \mid Flavor_i = cherry, \Theta_1 = \theta_1) = \theta_1$$
$$P(Wrapper_i = red \mid Flavor_i = lime, \Theta_2 = \theta_2) = \theta_2 .$$



与贝叶斯学习过程对应的贝叶斯网络。后验分布的参数Θ、Θ_1和Θ_2将根据它们的先验分布以及数据$Flavor_i$和$Wrapper_i$进行推断

完全数据学习

南京大学 NANJING UNIVERSITY

贝叶斯线性回归

$$P(\theta \mid \boldsymbol{d}) \propto P(\boldsymbol{d} \mid \theta) P(\theta)$$

$$P(y \mid x, \theta) = \mathcal{N}(y; \theta x, \sigma_y^2) = \frac{1}{\sigma\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{(y-\theta x)^2}{\sigma^2}\right)}$$

$$P(\theta) = \mathcal{N}(\theta; \theta_0, \sigma_0^2) = \frac{1}{\sigma_0\sqrt{2\pi}} \mathrm{e}^{-\frac{1}{2}\left(\frac{(\theta-\theta_0)^2}{\sigma_0^2}\right)}$$

$$P(\boldsymbol{d} \mid \theta)$$

$$P(\theta \mid \boldsymbol{d}) = \left(\prod_i P(x_i)\right) \prod_i P(y_i \mid x_i, \theta) = \alpha \prod_i \mathrm{e}^{-\frac{1}{2}\left(\frac{(y_i-\theta x_i)^2}{\sigma^2}\right)}$$

$$= \alpha \mathrm{e}^{-\frac{1}{2}\sum_i \left(\frac{(y_i-\theta x_i)^2}{\sigma^2}\right)}$$

$$P(\theta \mid \boldsymbol{d}) = \alpha''' \mathrm{e}^{-\frac{1}{2}\left(\frac{(\theta-\theta_N)^2}{\sigma_N^2}\right)} \quad \theta_N = \frac{\sigma^2\theta_0 + \sigma_0^2 \sum_i x_i y_i}{\sigma^2 + \sigma_0^2 \sum_i x_i^2} \quad \text{和} \quad \sigma_N^2 = \frac{\sigma^2\sigma_0^2}{\sigma^2 + \sigma_0^2 \sum_i x_i^2}$$

完全数据学习

南京大学 NANJING UNIVERSITY

贝叶斯线性回归

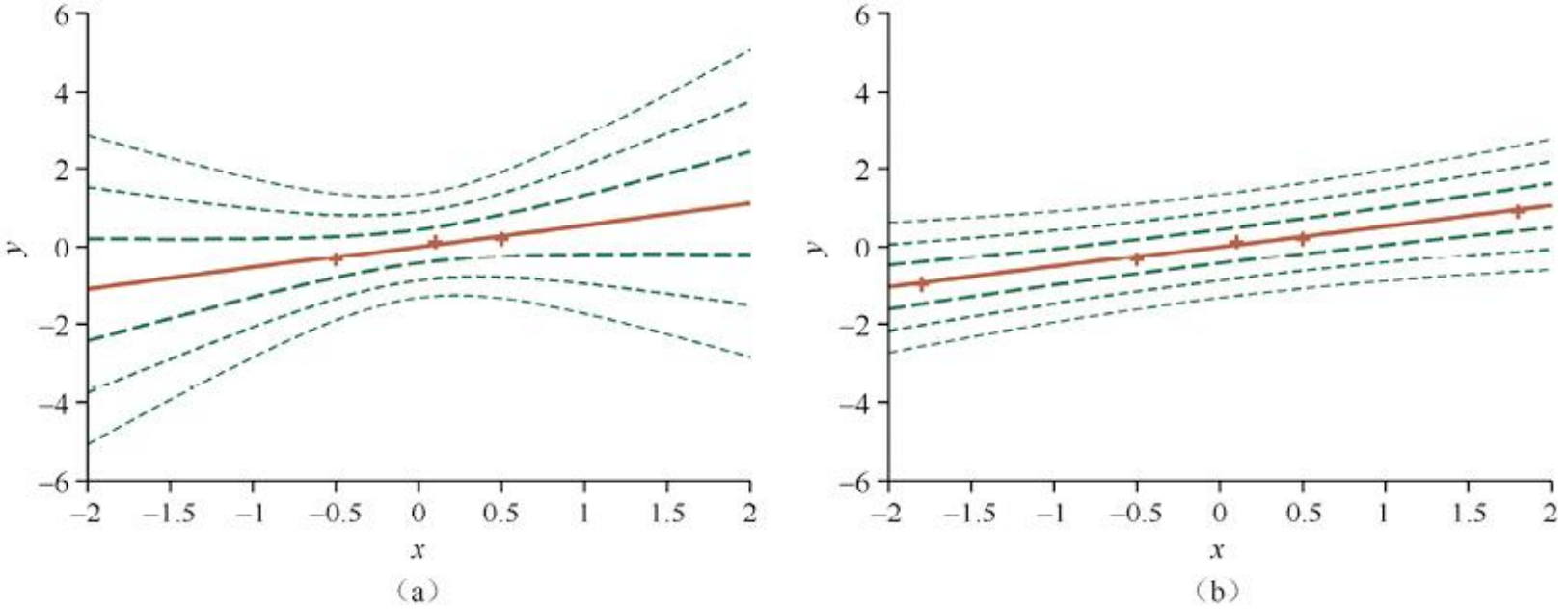
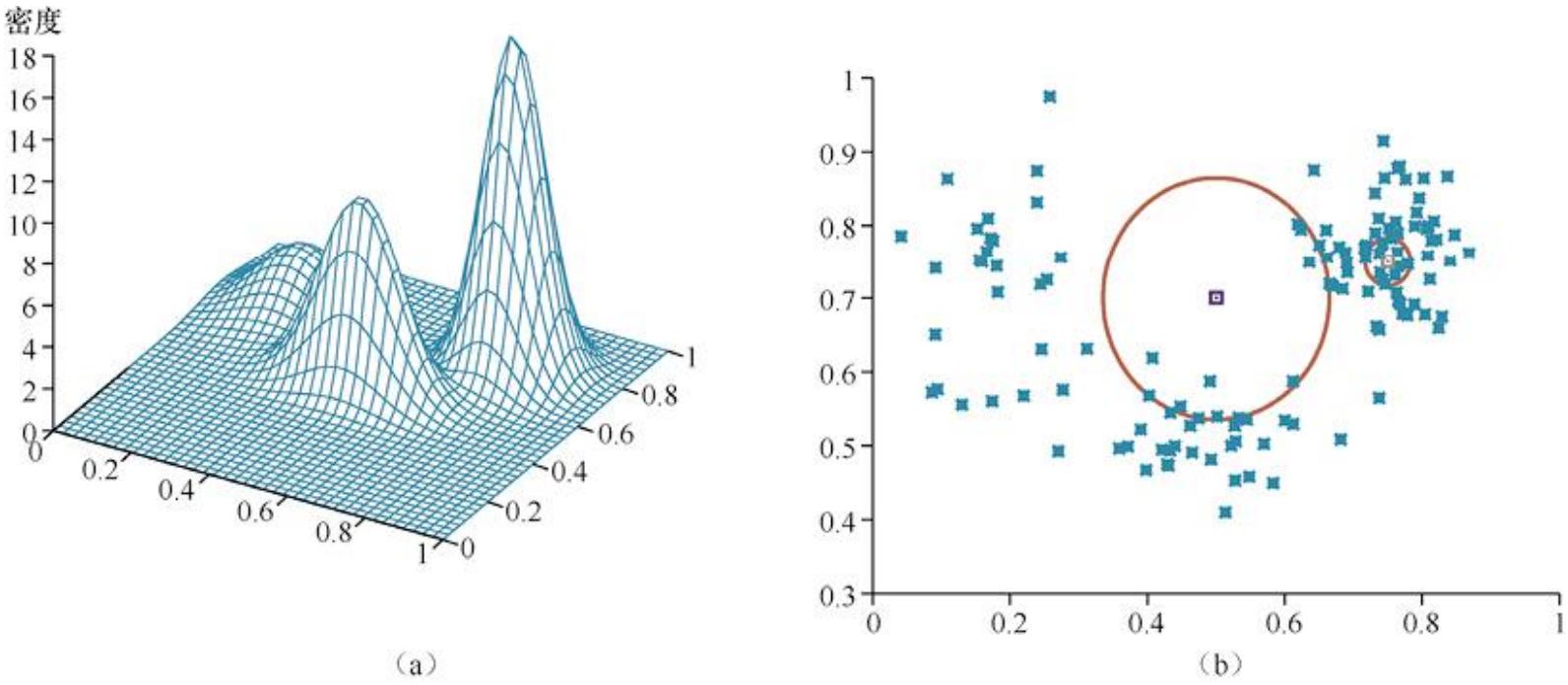


图20-7　贝叶斯线性回归模型，它被约束为经过原点且噪声方差固定为$\sigma^2 = 0.2$。误差为±1、±2和±3个标准差的密度预测等高线也在图中给出。（a）其中3个数据点距离原点较近，因此斜率相当不确定，其方差$\sigma_N^2 \approx 0.3861$。注意，当离观测到的数据点距离增大时，预测的不确定性也逐渐增大。（b）相比前一幅图多出两个距离较远的数据点，此时斜率θ被较严格地约束，其方差为$\sigma_N^2 \approx 0.0286$。密度预测中剩余的方差几乎完全来源于噪声的固定方差σ^2

完全数据学习

南京大学 NANJING UNIVERSITY

非参数模型密度估计



（a）混合高斯模型的三维样貌。（b）从混合高斯模型中采样的128个 数据点、两个查询点（小方块）以及它们的10近邻（大圆圈以及右边的小圆圈）

完全数据学习

南京大学 NANJING UNIVERSITY

非参数模型密度估计

直观上理解，概率密度就是单位区域数据出现的概率，公式表示如下

$$p(x) \cong \frac{k}{NV}$$

这里k是面积是V的区域内数据的个数，N是数据的总个数。

固定 k的值，利用数据确定 V的值，这引出了*k*-近邻算法。
固定 V的值，利用数据确定 k的值，这引出了核密度估计方法。

- *k*-近邻模型

为估计某个查询点***x***的未知概率密度，我们可以简单地估计数据点落在查询点***x***附近的密度。

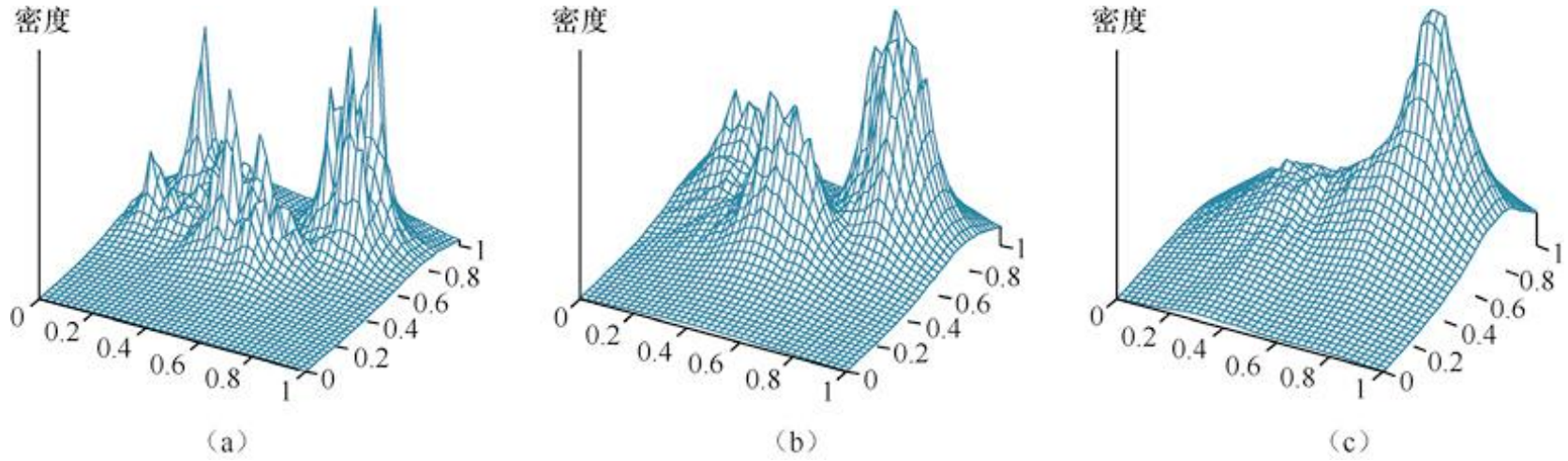
- 使用核函数

$$P(\boldsymbol{x}) = \frac{1}{N}\sum_{j=1}^{N}\mathcal{K}(\boldsymbol{x}, \boldsymbol{x}_j) \quad \mathcal{K}(\boldsymbol{x}, \boldsymbol{x}_j) = \frac{1}{(w^2\sqrt{2\pi})^d} e^{-\frac{D(\boldsymbol{x}, \boldsymbol{x}_j)^2}{2w^2}}$$

完全数据学习

南京大学 NANJING UNIVERSITY

非参数模型密度估计

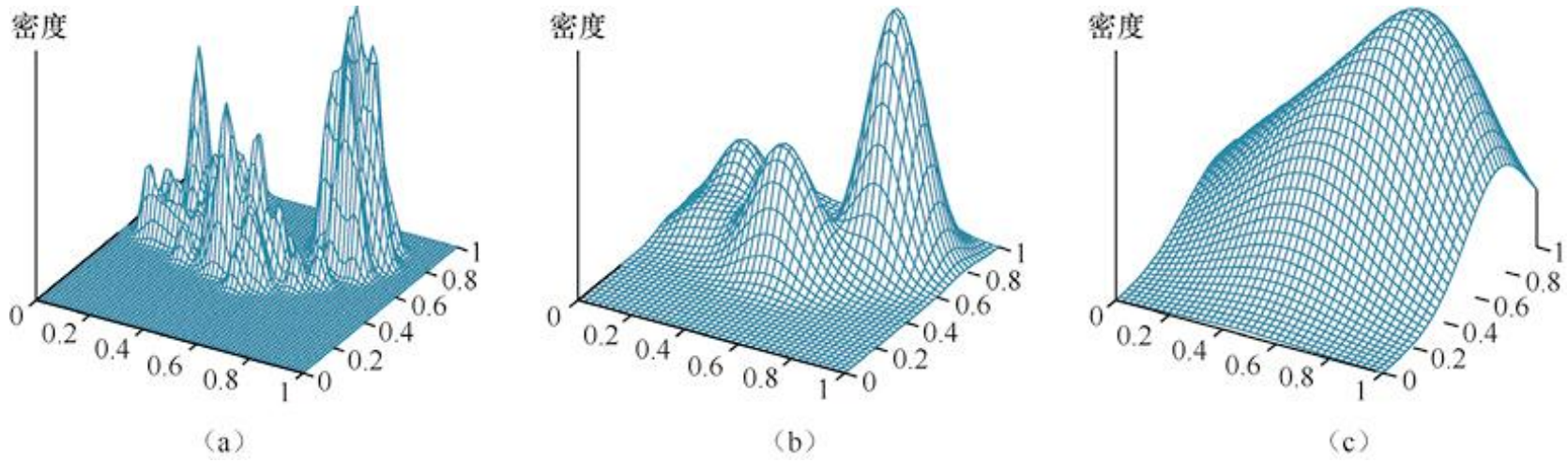


应用k近邻进行密度估计，所用的数据为上图b中的数据，分别对应$k = 3$、10和 40。$k = 3$的结果过于尖锐，40的结果过于光滑，而10的结果接近真实情况。最好的k值可以通过交叉验证进行选择。

完全数据学习

南京大学 NANJING UNIVERSITY

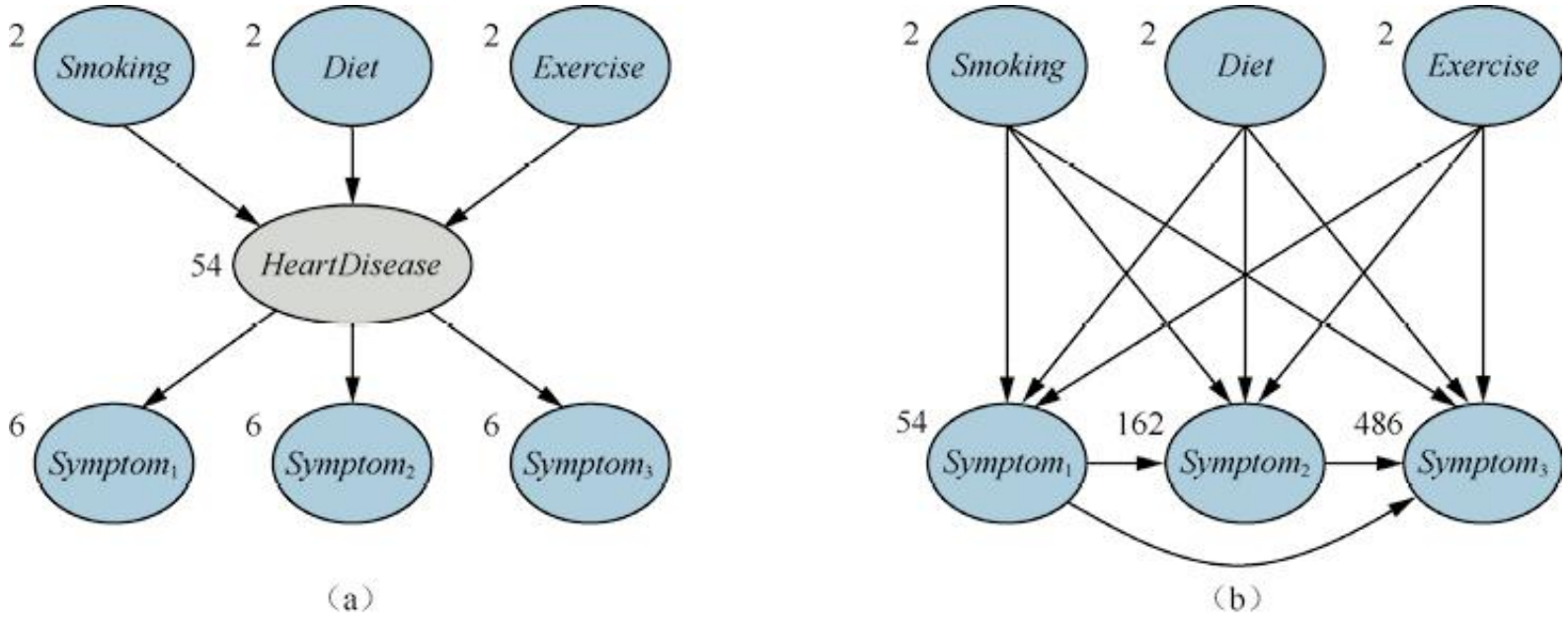
非参数模型密度估计



使用核函数进行密度估计，所用数据为上图b中的数据，分别采用了$w = 0.02$、0.07和0.20的高斯核。其中$w = 0.07$的结果最接近真实情况

隐变量学习：EM算法

- 在现实生活中，许多问题存在隐变量（hidden variable），这些变量在数据中是未被观测的。
- 隐变量可以大大减少确定一个贝叶斯网络所需参数的个数。同样也可以大大减少所需学习的参数的个数。



（a）一个简单的心脏病诊断网络，其中*HeartDisease*是一个隐变量。每个变量有3个可能的值，并标明了每个变量对应的条件独立参数的个数，其总数为78。（b）去除隐变量*HeartDisease*之后的等效网络。注意，给定了父变量值后，症状对应的变量不再是条件独立的。这个网络有708个参数

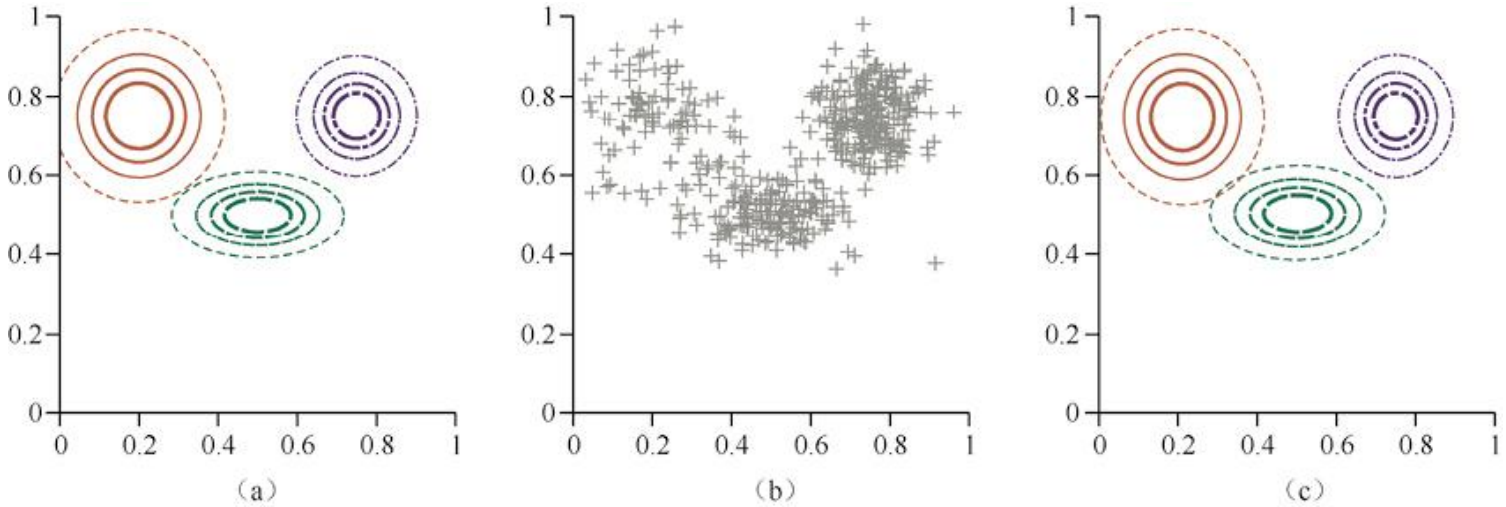
隐变量学习：EM算法

南京大学 NANJING UNIVERSITY

无监督聚类：学习混合高斯

- 无监督聚类是在一个对象集合中识别多个类别的问题，在学习过程中数据没有被赋予类别标签。
- 聚类假设了数据是从某个混合分布*P*中生成的。该分布由*k*个分量组成，每个分量本身是一个分布：

$$P(\boldsymbol{x}) = \sum_{i=1}^{k} P(C=i)\, P(\boldsymbol{x} \mid C=i)$$



（a）由3个分量组成的混合高斯模型，其权重（从左到右）分别为0.2、0.3和0.5。（b）采样于（a）中模型的500个数据点。（c）根据（b）中的数据点，使用EM算法重建出的模型。

隐变量学习：EM算法

南京大学 NANJING UNIVERSITY

无监督聚类：学习混合高斯

- 对于**混合高斯模型**，我们可以任意地初始化混合模型参数，然后进行以下两个步骤的迭代：

（1）**E步**：计算概率$p_{ij} = P(C = i \mid \boldsymbol{x}_j)$，即数据点$\boldsymbol{x}_j$是由分量$i$生成的概率。根据贝叶斯法则，我们有$p_{ij} = \alpha P(\boldsymbol{x}_j \mid C = i)P(C = i)$。其中$P(\boldsymbol{x}_j \mid C = i)$项是$\boldsymbol{x}_j$在第$i$个高斯分量中的概率，$P(C = i)$项是第$i$个高斯分量的权重。定义$n_i = \sum_j p_{ij}$，即分配至第$i$个分量的数据点的有效个数。

（2）**M步**：按照以下式子计算新的均值、方差和各分量的权重。

$$\boldsymbol{\mu}_i \leftarrow \sum_j p_{ij}\mathbf{x}_j / n_i$$

$$\boldsymbol{\Sigma}_i \leftarrow \sum_j p_{ij}(\mathbf{x}_j - \boldsymbol{\mu}_i)(\mathbf{x}_j - \boldsymbol{\mu}_i)^\top / n_i$$

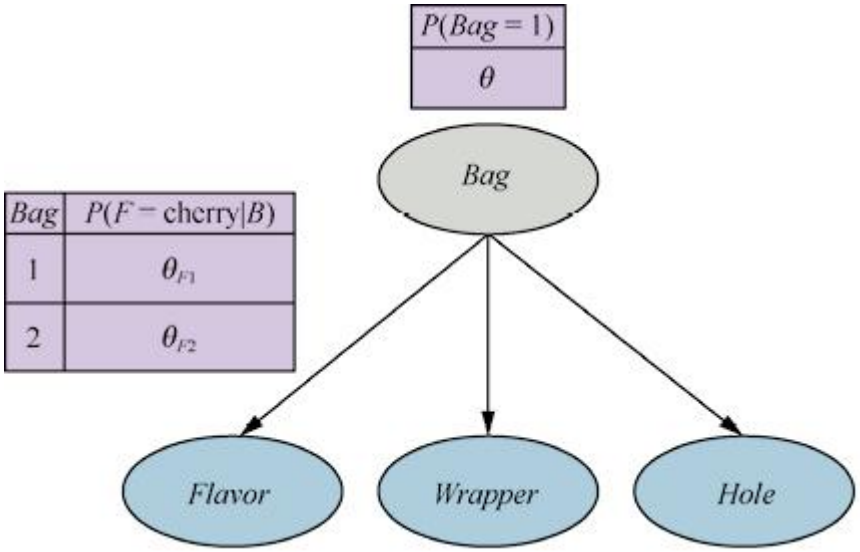
$$w_i \leftarrow n_i / N$$

隐变量学习：EM算法

学习带隐变量的贝叶斯网络参数值

示例: 有两袋混合在一起的糖果

- 糖果有3个特征：除口味（Flavor）和包装（Wrapper）外，一些糖果中间还有夹心（Holes），而有些糖果没有。
- 糖果在每个糖果袋中的分布状况可以用**朴素贝叶斯**模型进行描述：在给定糖果袋的情况下，特征之间是独立的，但每个特征的条件概率取决于这个糖果袋的状况。



	$W = red$		$W = green$	
	$H = 1$	$H = 0$	$H = 1$	$H = 0$
$F = cherry$	273	93	104	90
$F = lime$	79	100	94	167

隐变量学习：EM算法

南京大学 NANJING UNIVERSITY

学习带隐变量的贝叶斯网络参数值

糖果个数的期望 $\hat{N}(Bag=1)$是每个糖果来自于糖果袋1的概率之和:

$$\theta^{(1)} = \hat{N}(Bag=1)/N = \sum_{j=1}^{N} P(Bag=1 \mid flavor_j, wrapper_j, holes_j)/N .$$

以利用贝叶斯法则以及条件独立性计算得到

$$\theta^{(1)} = \frac{1}{N}\sum_{j=1}^{N} \frac{P(flavor_j \mid Bag=1)P(wrapper_j \mid Bag=1)P(holes_j \mid Bag=1)P(Bag=1)}{\sum_i P(flavor_j \mid Bag=i)P(wrapper_j \mid Bag=i)P(holes_j \mid Bag=i)P(Bag=i)} .$$

糖果袋1中的樱桃味糖果数量的期望可以由下式给出

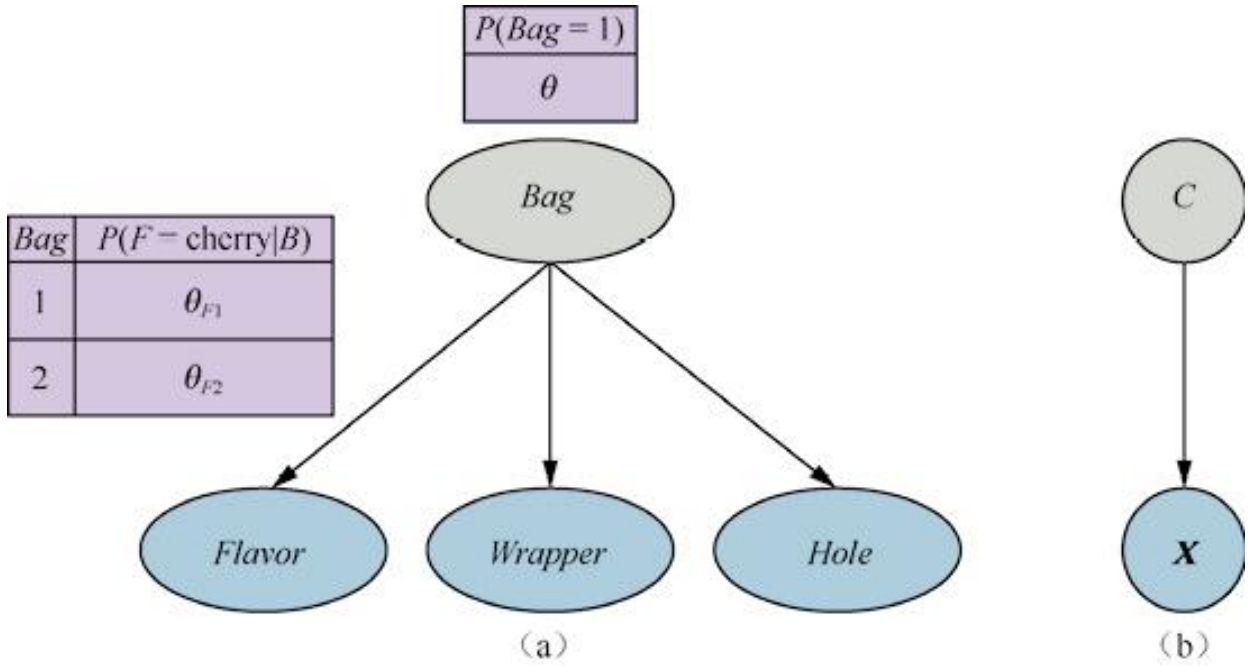
$$\sum_{j:Flavor_j=cherry} P(Bag=1 \mid Flavor_j=cherry, wrapper_j, holes_j) .$$

更新由计数期望归一化后给出

$$\theta_{ijk} \leftarrow \hat{N}(X_i=x_{ij}, \mathbf{U}_i=\mathbf{u}_{ik})/\hat{N}(\mathbf{U}_i=\mathbf{u}_{ik}) .$$

隐变量学习：EM算法

学习带隐变量的贝叶斯网络参数值



（a）关于糖果的混合模型。不同口味、包装的比例以及是否有夹心取决于糖果袋，该变量是不可观测的。（b）混合高斯模型的贝叶斯网络。可观测变量$\boldsymbol{X}$的均值和协方差取决于分量C

隐变量学习：EM算法

南京大学 NANJING UNIVERSITY

学习隐马尔可夫模型

EM算法可以应用于学习隐马尔可夫模型（HMM）中的转移概率

隐马尔可夫模型可以用一个带有单个离散状态变量的动态贝叶斯网络来表示

每个数据点为一个长度有限的观测序列

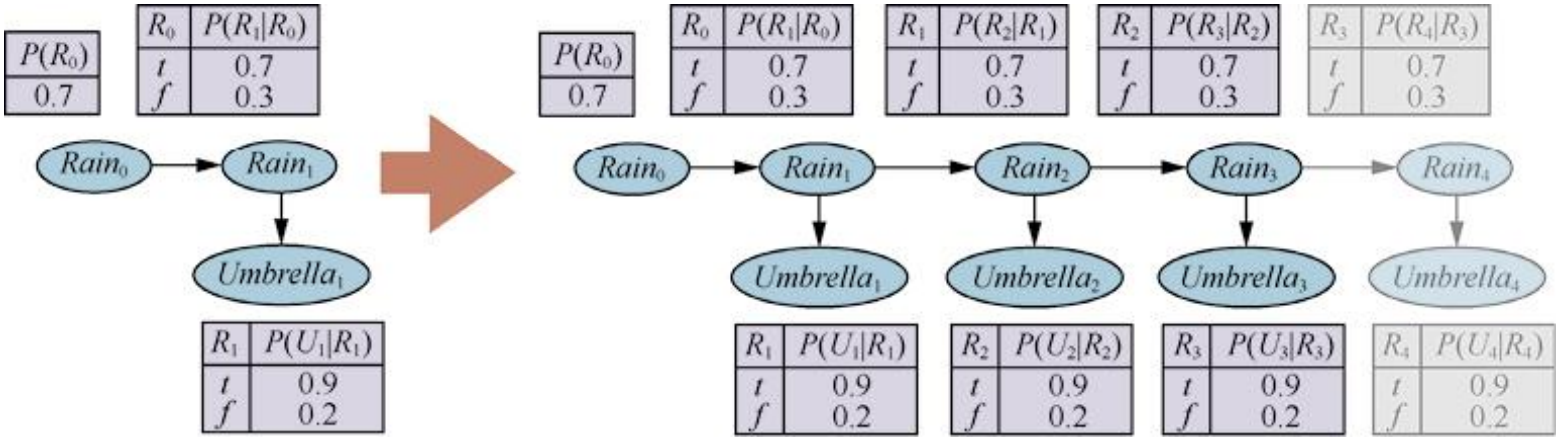
从状态 i 到状态 j 的转移概率

- 计算系统在状态 i 经过一次转移后到达状态 j 的次数比例的期望:

$$\theta_{ij} \leftarrow \sum_t \hat{N}(X_{t+1}=j, X_t=i) / \sum_t \hat{N}(X_t=i) \,.$$

隐变量学习：EM算法

学习隐马尔可夫模型



表示隐马尔可夫模型的动态贝叶斯网络展开图

隐变量学习：EM算法

南京大学 NANJING UNIVERSITY

EM算法的一般形式

X：所有样例中的所有观测值,
Z：所有样例中的所有隐变量,
θ：概率模型中的所有参数

$$\boldsymbol{\theta}^{(i+1)} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \sum_{\mathbf{z}} P(\mathbf{Z}=\mathbf{z} \mid \mathbf{x}, \boldsymbol{\theta}^{(i)}) L(\mathbf{x}, \mathbf{Z}=\mathbf{z} \mid \boldsymbol{\theta}) \ .$$

E步是求和计算

M步则是选取参数使得该对数似然的期望达到最大.