人工智能原理与算法

第20章-概率模型学习

中科院自动化研究所　朱翔昱

- 统计学习
- 带完整数据的学习
- 隐变量学习

20.1 统计学习

- **贝叶斯网络的学习方法**
  - 目标：贝叶斯网络的条件概率分布。
  - 离散情况：条件概率分布表格
  - 连续情况：线性高斯分布
- **数据：**
  - 描述领域的某些或全部随机变量的示例，即随机变量的值。
- **假设：**
  - 关于领域是如何工作的概率性理论，即随机变量的分布。

20.1 统计学习 - 贝叶斯学习

- **贝叶斯学习:**
  - 给定数据，贝叶斯学习计算每个假设的概率，并基于这些概率做决策。**它使用所有假设做预测，并用概率加权，而不是使用单个“最好”的假说。**
  - 令$\boldsymbol{d}$为观察到的所有数据，$(h_1, h_2, \dots, h_i)$为一系列关于数据的假设，则贝叶斯规则下每个假说的概率：

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

  - 如果希望做关于未知随机变量X的预测，则有

$$\mathbf{P}(X \mid \boldsymbol{d}) = \sum_i \mathbf{P}(X \mid \boldsymbol{d}, h_i)\mathbf{P}(h_i \mid \boldsymbol{d}) = \sum_i \mathbf{P}(X \mid h_i)P(h_i \mid \boldsymbol{d})$$

20.1 统计学习 - 贝叶斯学习

贝叶斯学习：

$$\mathbf{P}(X \mid \boldsymbol{d}) = \sum_i \mathbf{P}(X \mid \boldsymbol{d}, h_i)\mathbf{P}(h_i \mid \boldsymbol{d}) = \sum_i \mathbf{P}(X \mid h_i)P(h_i \mid \boldsymbol{d})$$

- 每个假设都确定了X上的一个概率分布，对X的预测是每个单独假设预测的加权平均。
- **假设本身就是原始数据和预测之间的过渡。**

20.1 统计学习 - 贝叶斯学习

示例：糖果比例预测

- 一袋糖果共5种类型的组合形式
  - h1：100% 樱桃味
  - h2：75% 樱桃味 + 25% 酸橙味
  - h3：50% 樱桃味 + 50% 酸橙味
  - h4：25% 樱桃味 + 75% 酸橙味
  - h5：100% 酸橙味
- 给定一袋未拆袋的糖果，用随机变量 H（代表假设）表示糖果袋的类型，其可能的值为从 h1 至 h5。
- 外观上没有任何信息表明一袋糖果的类型，但随着袋中的糖果逐颗被打开与辨认，越来越多的证据支持其中一种假设。

20.1 统计学习 - 贝叶斯学习

示例：糖果比例预测

- 使用后验概率$P(h_i|\boldsymbol{d})$评估假设的可能性
- 假设的先验：设5种糖果比例h1, ..., h5的先验概率为

$P(h_i) \quad = < 0.1, 0.2, 0.4, 0.2, 0.1 >$ （糖果商广告给出）

- 假设的似然：每个糖果独立同分布（糖果足够多等价于有放回采样）

$$P(\boldsymbol{d}|h_i) = \prod_j P(d_i|h_i)$$

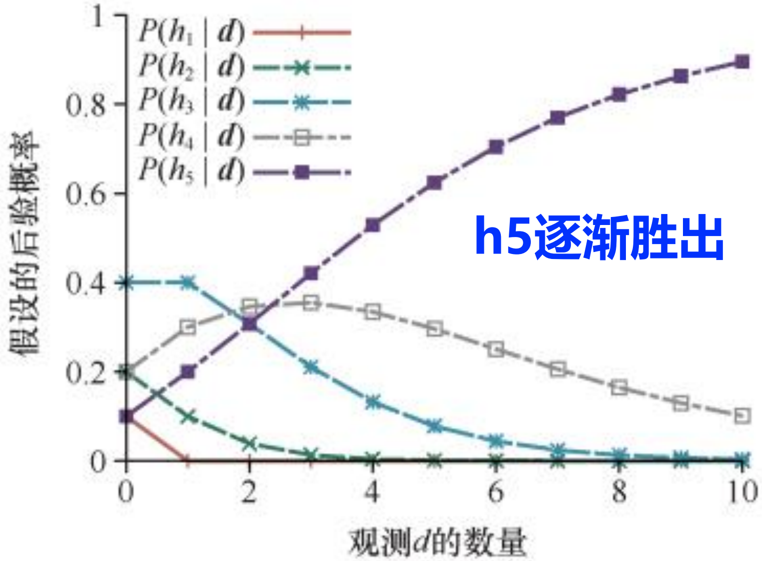
  - 举例：如果从糖果袋中取出的前10颗糖果都是酸橙味，那么假设h3（一半苹果一半酸橙）的似然为0.5^{10}

- 后验概率：

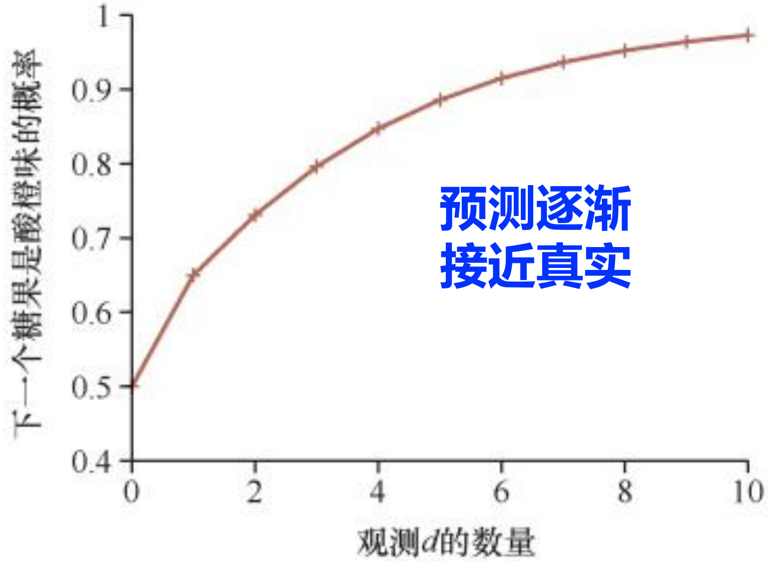
$$P(h_i|\boldsymbol{d}) = \prod_j P(d_i|h_i)P(h_i)$$

20.1 统计学习 - 贝叶斯学习

- 如果取出的糖果全部为酸橙味，则随着糖果的增加各假说后验概率：



假说成立的概率



对下一颗糖果的贝叶斯预测

- **贝叶斯预测最终将与正确的假说一致**。给定任何不排除正确假说的先验，任何错误假说的后验概率将最终消失。
- 贝叶斯预测是最优的，**因为他是所有可能假设的加权和**，任何其他预测（使用单个最优假设）的正确率总是要小些。

20.1 统计学习 - 贝叶斯学习

使用单个最优假设

最大后验假设(maximum a posteriori，MAP)

- 基于单个可能性最大的假设进行预测：

MAP: $$\mathbf{P}(X \mid \boldsymbol{d}) \approx \mathbf{P}(X \mid h_{max})$$

$$h_{max} = argmax_{(h_i)} P(\boldsymbol{d} \mid h_i) P(h_i)$$

最大似然假设(maximum-likelihood，ML)

- 没有关于假设先验的知识，以似然最大的假设为最大可能假设:

ML: $$h_{max} = argmax_{(h_i)} P(\boldsymbol{d} \mid h_i)$$

- 统计学习
- 带完整数据的学习
- 隐变量学习

20.2 带完整数据的学习

- **完全数据下的概率密度估计：**
  - 密度估计：给定从一个概率模型中产生的数据，学习概率模型的任务被称为密度估计。
  - 完全数据：每个数据点包含所有随机变量的值。
- **例子：**
  - 购买一袋樱桃和酸橙糖果，其中樱桃和酸橙的比例是未知的，可以是0和1之间的任何数。
  - 参数θ是樱桃糖所占比例（酸橙的比例是1-θ）假说是h_θ。
  - 所有比例具有相同先验可能性，使用极大似然方法。

20.2 带完整数据的学习

- **例子：**
  - 现打开了N颗糖果，其中c颗为樱桃味，$l = N - c$颗为橙味，则假设h_θ成立的似然为:

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

  - 极大似然，似然取log，不影响最大值：

$$L(\mathbf{d} \mid h_\theta) = \log P(\mathbf{d} \mid h_\theta) = \sum_{j=1}^{N} \log P\left(d_j \mid h_\theta\right) = c\log\theta + l\log(1-\theta)$$

$$\Downarrow$$

$$\frac{dL(\mathbf{d} \mid h_\theta)}{d\theta} = \frac{c}{\theta} - \frac{l}{1-\theta} = 0 \Rightarrow \theta = \frac{c}{c+l} = \frac{c}{N}$$

极大似然假说下，袋中樱桃口味比例的**学习结果**
等于目前撕开的糖果中的樱桃口味的**观察比例**

20.2 带完整数据的学习

- **极大似然参数学习方法：**

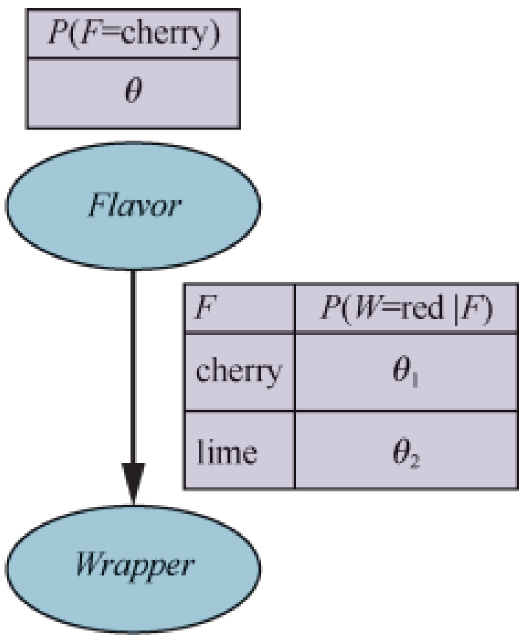
1. 将数据的似然写成关于参数的函数形式。
2. 将似然log化。
3. 计算log似然对每个参数的导数。
4. 解出导数为0的参数。

20.2 带完整数据的学习-离散模型

■ 贝叶斯网络学习——离散情形：

- 举例2：制造商使用红色或绿色糖纸包装糖果，包装是依赖于味道的未知条件分布。我们希望知道**糖果比例**和**包装规则。**
- 概率模型表达为贝叶斯网络，那么任意数据似然为：

$$P\left(Flavor, Wrapper \mid h_{\theta,\theta_1,\theta_2}\right) = P(Flavor \mid h_{\theta,\theta,\theta_2})P(Wrapper \mid Flavor, h_{\theta,\theta_1,\theta_2})$$



有三个参数

- θ：樱桃味糖果的概率（比例）
- θ_1：给定一颗糖果为樱桃味，包装为红的概率。
- θ_2：给定一颗糖果为酸橙味，包装为红的概率。

20.2 带完整数据的学习-离散模型

■ 贝叶斯网络学习——离散情形：

- 现在撕开N个糖块，数据如下：
  - N个糖中：c个樱桃和l个酸橙
  - c个樱桃中：r_c个红色包装， g_c个绿色包装
  - l个酸橙中： r_l个红色包装，g_l个绿色包装
- 数据似然$P(Flavor)P(Wrapper \mid Flavor)$为：

$$P\left(d \mid h_{\theta,a_1,a_2}\right) = \theta^c(1-\theta)^l \cdot \theta_1^{r_c}(1-\theta_1)^{g_c} \cdot \theta_2^{r_l}(1-\theta_2)^{g_l}$$

- 取对数：

$$\begin{aligned} L &= [c\log\theta + l\log(1-\theta)] \\ &+ [r_c\log\theta_1 + g_c\log(1-\theta_1)] \\ &+ [r_l\log\theta_2 + g_l\log(1-\theta_2)] \end{aligned}$$

20.2 带完整数据的学习-离散模型

■ 贝叶斯网络学习——离散情形：

- 求偏导数为0，获得结果

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

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

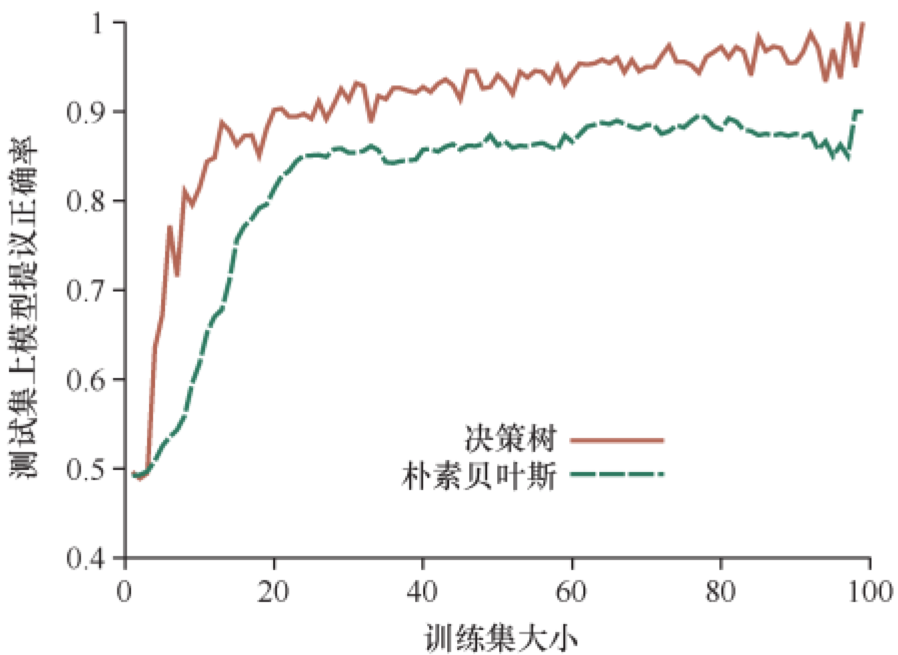
$$\frac{\partial L}{\partial \theta_2} = \frac{r_i}{\theta_2} - \frac{g_i}{1-\theta_2} = 0 \Rightarrow \theta_2 = \frac{r_l}{r_l+g_l}$$

- 该例子可以推广到任意离散贝叶斯网络。
- 一旦有了完全数据，贝叶斯网络的最大似然参数学习问题可以分解为一些**分离的学习问题，每个问题对应一个参数。**

20.2 带完整数据的学习-朴素贝叶斯模型

- 朴素贝叶斯：假设给定类时，属性相互条件独立

$$P(C \mid x_1, \cdots, x_n) = \alpha P(C) \prod_i P(x_i \mid C)$$



即使真实函数是决策树（完全不满足条件独立性），但朴素贝叶斯仍然取得了不错的效果

20.2 带完整数据的学习-连续模型

贝叶斯网络学习——连续情形：

- 简单情形：假设h为高斯分布，要学习的参数为均值μ和标准差σ

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

- 数据d为$x_1 \dots x_j$，则log似然为：

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

- 求偏导数取0：极大似然下，均值是样本的平均值，标准偏是样本偏差的平方根，与常识一致。

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

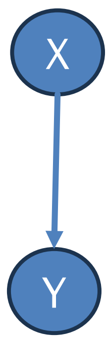
$$\frac{\partial L}{\partial \sigma} = -\frac{N}{\sigma} + \frac{1}{\sigma^3}\sum_{j=1}^{N}(x_j-\mu)^2 = 0 \Rightarrow \sigma = \sqrt{\frac{\sum_j (x_j-\mu)^2}{N}}$$

20.2 带完整数据的学习-连续模型

贝叶斯网络学习——连续情形：

- 线性高斯模型，y的均值线性依赖于x，且有固定的标准差。
- 参数为：σ，θ_1，θ_2：

连续变量



连续变量

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

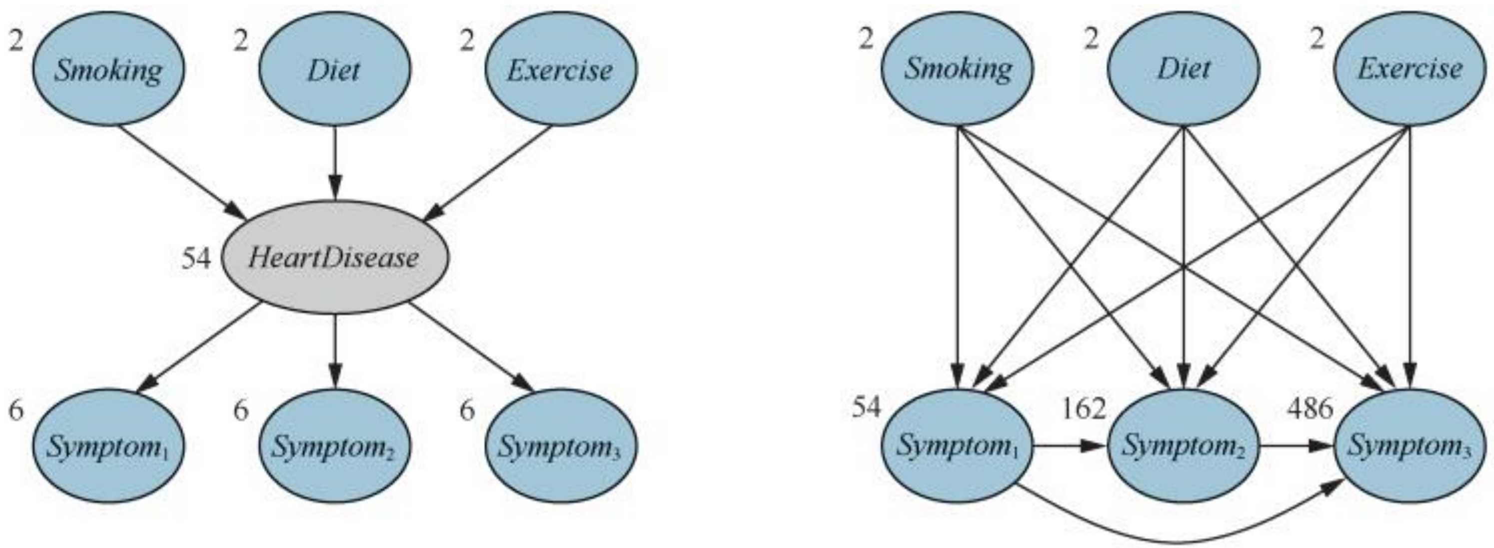
θ_1 ,θ_2极大似然值等于$(y-(\theta_1 x+\theta_2))^2$的最小值

线性高斯模型的极大似然优化等价于L2损失下的线性回归

- 统计学习
- 带完整数据的学习
- 隐变量学习

20.3 隐变量学习：EM算法

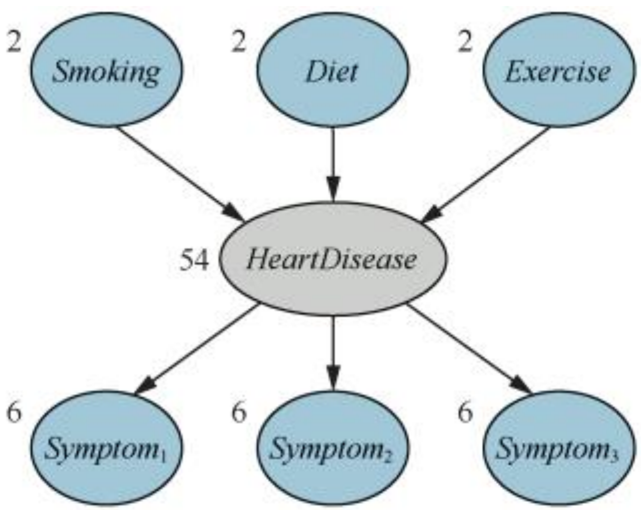
- **隐变量：在数据中不可观察的随机变量。**
  - 医学记录经常包含所观察的症状、医生的诊断、应用的治疗方案，也许还有治疗结果，但很少包含疾病本身的直接观察。



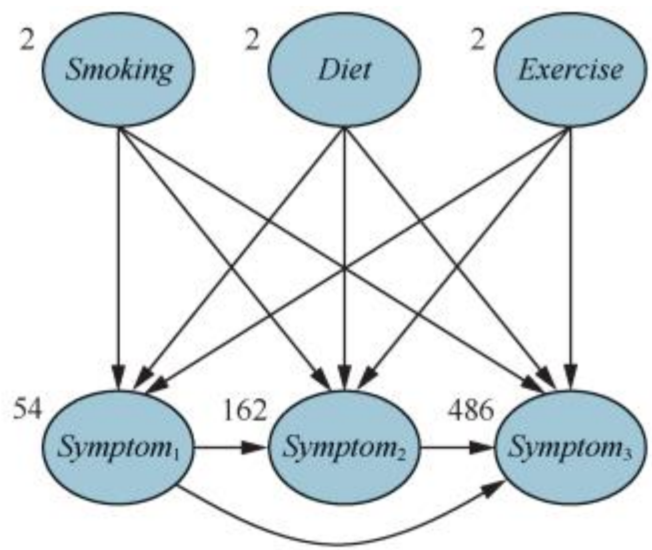
  - 心脏病诊断：可观察变量
    - 心脏病因素：吸烟(smoking)、饮食不规律(diet)、运动少（exercise）
    - 心脏病症状：Symptom1, Symptom2, Symptom3.

20.3 隐变量学习：EM算法

- **隐变量：在数据中不可观察的随机变量。**
  - 为什么要有隐变量：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。
  - 但隐变量在数据中没有数值。需要包含隐变量的新的学习方法，即期望最大化算法（EM算法）。



包含隐变量：78参数

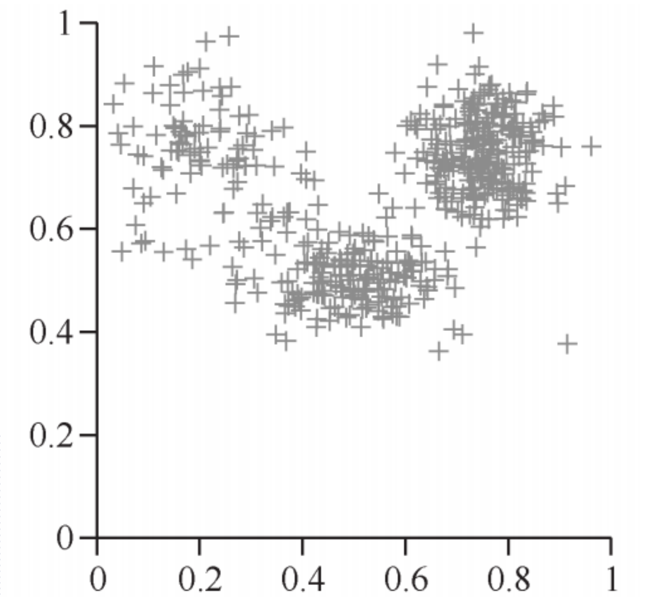


无隐变量：708参数

20.3 隐变量学习：EM算法

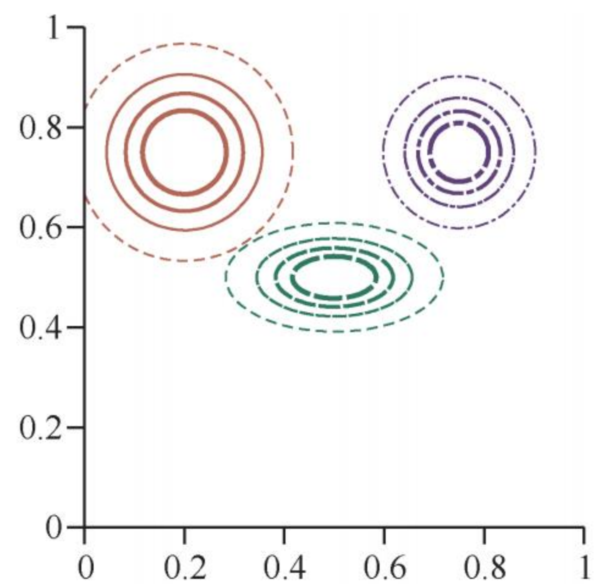
隐变量学习：混合高斯分布

- **定义**：数据的分布为多个子分布的加权和。
- 没有提供数据来自于哪个子分布。



每个数据来自哪个子分布？

每个子分布是？



20.3 隐变量学习：EM算法-

■ 隐变量学习：混合高斯分布

- 假设：数据从某混合分布P中生成，该分布有k个分量，每个分量本身就是一个分布。
- 数据生成方式：首先以一定概率选择某一个分量，然后从该分量采样一个样本，从而生成一个数据点。
- 数据的概率分布为：

$$P(\mathbf{x}) = \sum_{i=1}^{k} P(C=i)P(\mathbf{x}|C=i)$$

选择一个分量　　　　从该分量中采样数据

- 学习任务：给定数据（$x_1 \ldots x_j$），学习**选择分量的概率$P(C=i)$，各分量的均值μ_i和方差σ_i**

20.3 隐变量学习：EM算法

- **问题难点：**
  - 既不知道数据来源于哪个分量，也不知道各分量的模型参数。
- **EM的基本思想：**
  - 分为两个步骤：**补全隐变量值（E步），更新模型参数（M步），** 参数随机初始化，E步和M步交替进行直至收敛。
    a) 先假设我们知道分量模型的均值方差（随机初始化），然后推导出每个数据点属于每个分量的概率（补全变量值）。
    b) 每个数据点用它属于该分量的概率加权，重新计算每个分量的均值方差（计算模型参数）。
    c) 这个过程重复进行，直到收敛。

20.3 隐变量学习：EM算法

- **EM算法：**
  - **E-步**：计算数据x_j由分量i产生的概率$P_{\{ij\}} = P\left(C = i \mid x_j \right)$。
    - 由贝叶斯规则，有$p_{ij} = \alpha P(x_j \mid C = i)P(C = i)$，其中$P(x_j \mid C = i)$是第$i$个高斯能够生成点$x_j$的概率，$P(C = i)$ 是选择第i个高斯的权重。
    - 此时，每个样例**仅部分属于某分量**，**这个“部分”的程度取决于** p_{ij} 。
    - **产生一个加权数据集**：**每个数据被分为了几份，每一份的显变量值相同，隐变量值不同，每个隐变量值对应的那份权重为p_{ij}**

原始数据

权重=1	隐变量：C未知	显变量：x_j

EM下的加权数据

权重=$P\left(C = 1 \mid x_j \right)$	隐变量：C=1	显变量：x_j
权重=$P\left(C = 2 \mid x_j \right)$	隐变量：C=2	显变量：x_j
权重=$P\left(C = 3 \mid x_j \right)$	隐变量：C=2	显变量：x_j

20.3 隐变量学习：EM算法

- **EM算法：**
  - **M-步： 在加权数据集下，使用极大似然**计算新平均值、协方差和分量权重：

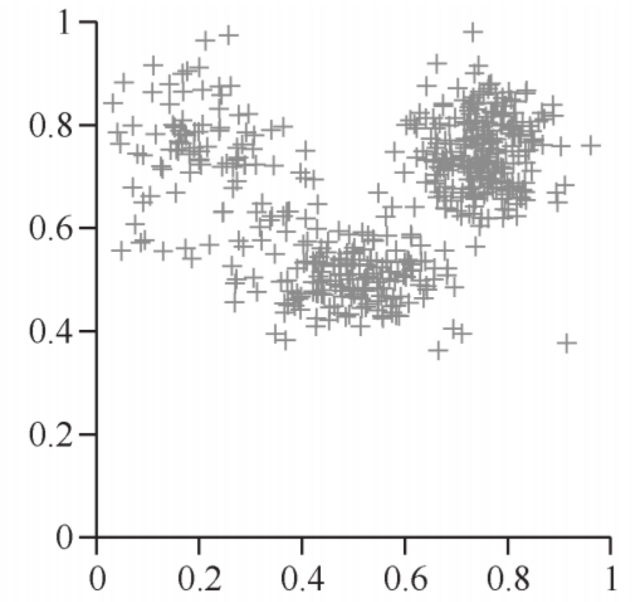
$$w_\mathrm{i} \leftarrow \frac{n_\mathrm{i}}{N}$$

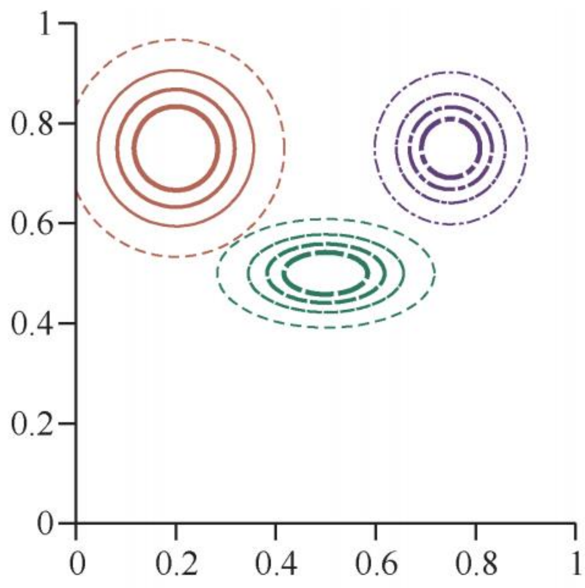
$$\Sigma_i \leftarrow \sum_j \boxed{p_{ij}}(\mathbf{x}_j - \mu_i)(\mathbf{x}_j - \mu_i)^T/n$$

$$\mu_i \leftarrow \sum_j \boxed{p_{ij}}\mathbf{x}_j/n_i$$

20.3 隐变量学习：EM算法

- 将EM算法的结果

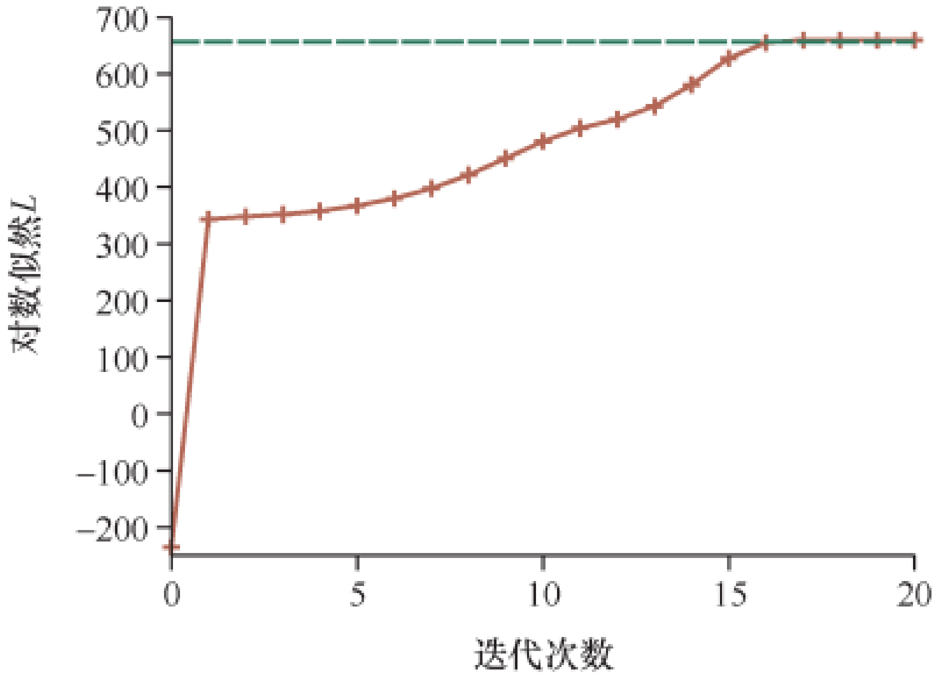




中国科学院自动化研究所
INSTITUTE OF AUTOMATION
CHINESE ACADEMY OF SCIENCES

20.3 隐变量学习：EM算法

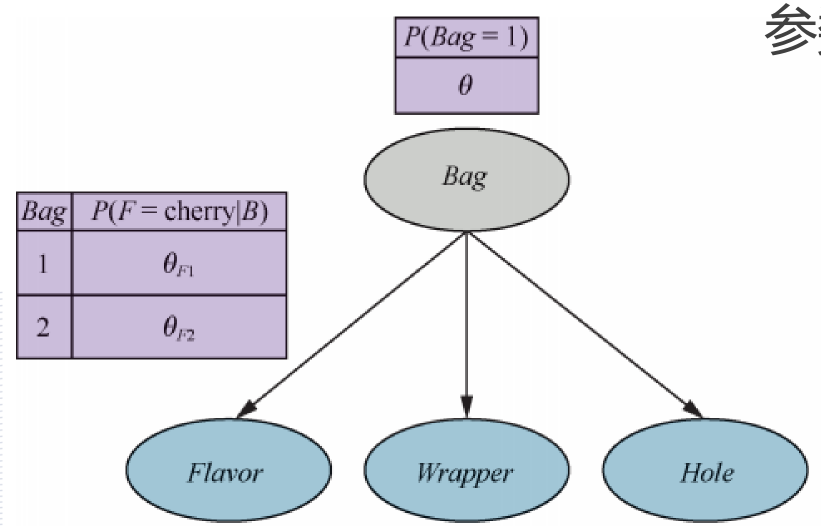
- 将Log似然性曲线



20.3 隐变量学习 – 带隐变量的贝叶斯网络

■ 示例：两袋糖果混合

- 糖果有 3 个特征：口味（Flavor）、包装（Wrapper）、夹心（Holes）
- 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。（朴素贝叶斯）



参数：

- θ：糖果取自袋1的先验概率
- θ_{F1}, θ_{F2} :袋1,袋2中樱桃口味的概率
- θ_{W1}, θ_{W2} :袋1,袋2中红色包装的概率
- θ_{H1} , θ_{H2} :袋1,袋2中有夹心的概率

20.3 隐变量学习 – 带隐变量的贝叶斯网络

示例：两袋糖果混合

- 糖果袋是一个隐变量，因为一旦糖果混合在一起，就不再能知道每个糖果来自哪个糖果袋。
- 通过观测混合物中的糖果来复原这两个袋子的真实情况。
- 观测数据为：

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

20.3 隐变量学习 – 带隐变量的贝叶斯网络

示例：两袋糖果混合

1. 参数随机初始化；

$$\theta^{(0)}=0.6,\ \theta_{F1}^{(0)}=\theta_{W1}^{(0)}=\theta_{H1}^{(0)}=0.6,\ \theta_{F2}^{(0)}=\theta_{W2}^{(0)}=\theta_{H2}^{(0)}=0.4$$

2. 补全隐变量：计算每个糖果来源于糖果袋1的概率p，那么该糖果就有p权重属于袋1，1-p权重属于袋2。

3. 更新参数θ，袋1糖果数量的期望为：

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

4. 更新参数θ_{F1}，袋1樱桃味数量期望：

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

对所有糖果，如果为樱桃味，则将其属于袋1的权重求和

20.3 隐变量学习 – 带隐变量的贝叶斯网络

■ 示例：两袋糖果混合

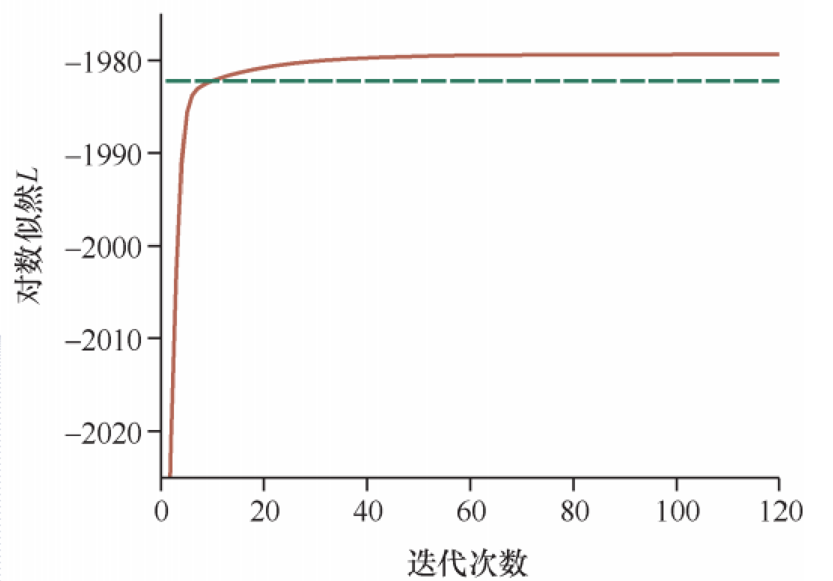
5. 更新其他参数；

$$\theta^{(0)} = 0.6,\ \theta_{F1}^{(0)} = \theta_{W1}^{(0)} = \theta_{H1}^{(0)} = 0.6,\ \theta_{F2}^{(0)} = \theta_{W2}^{(0)} = \theta_{H2}^{(0)} = 0.4$$

$$\Downarrow$$

$$\theta^{(1)} = 0.6124,\ \theta_{F1}^{(1)} = 0.6684,\ \theta_{W1}^{(1)} = 0.6483,\ \theta_{H1}^{(1)} = 0.6558$$

$$\theta_{F2}^{(1)} = 0.3887,\ \theta_{W2}^{(1)} = 0.3817,\ \theta_{H1}^{(1)} = 0.3827$$



模型并不唯一

本章小结

- 贝叶斯学习计算每个假说的概率，并基于这些概率做决策。它使用所有假说做预测，并用概率加权。
- 最大后验（MAP）学习选择给定数据下可能性最大的假设。同样利用了假设的先验分布，并且该方法通常比贝叶斯学习更易处理。
- 最大似然学习（ML）选择使得数据的似然最大的假设；它等价于使用均匀分布作为先验的最大后验学习。
- 当一些变量被隐藏时，期望最大化（EM）算法可以找到局部最大似然解。EM算法包含两步：补全隐变量值（E步），更新模型参数（M步）。参数随机初始化，E步和M步交替进行直至收敛。

THANKS

Q & A