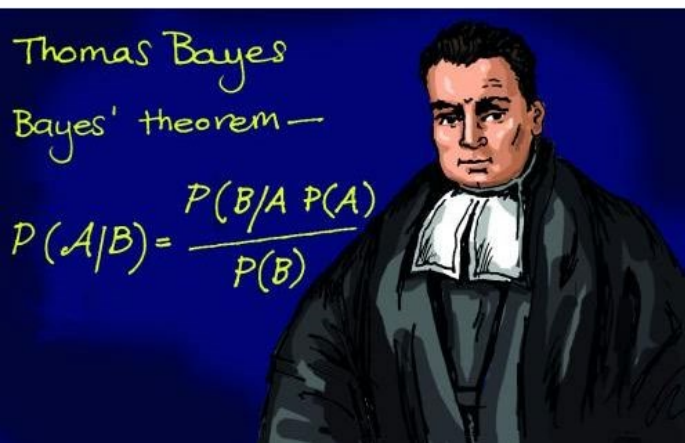


第三讲 贝叶斯分类器

授课老师：郭 迟 教授

guochi@whu.edu.cn



- 1 贝叶斯决策理论概念和规则
- 2 类条件概率密度函数的估计
- 3 朴素贝叶斯分类器
- 4 EM算法

第三讲 贝叶斯分类器



贝叶斯决策理论概念和规则

1. 几个基本概念

- **先验概率 (priori probability)**：没有对样本进行任何观测情况下的概率。在一般分类问题中，常以 $\omega_i (i=1,2,3\dots c)$ 表示类型，以 $P(\omega_i)$ 表示其先验概率，

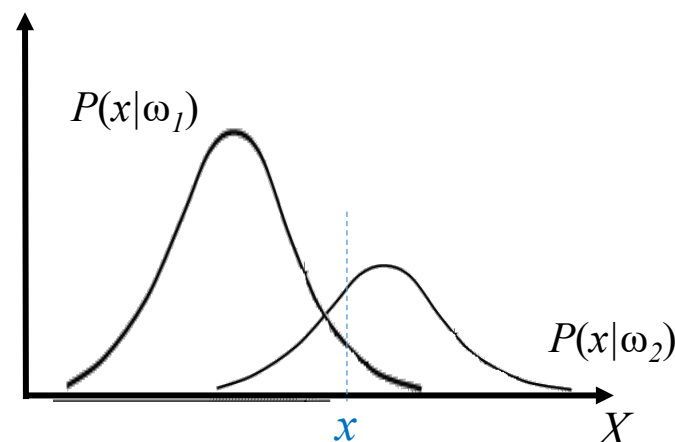
$$\sum_i P(\omega_i)=1$$

这个东西很重要，不同的名称含义也不同

- e.g 鼻咽拭子核酸检测的阴性 $P(\omega_1)$ 与阳性 $P(\omega_2)$ 先验概率分别是0.9和0.1

- **类似然 (class likelihood)**：即属于类 ω_i 的事件具有观测值 X 的条件概率，记作 $P(X|\omega_i)$ 。获得这些条件概率的分布情况，形成**类条件概率密度函数**

- e.g 核酸检测的类条件概率密度函数如下，观测特征 x 经该函数查表可得 $P(x|\omega_1)=0.2$ ， $P(x|\omega_2)=0.4$



提问：观测值 x 到底是阴性还是阳性呢？

第三讲 贝叶斯分类器



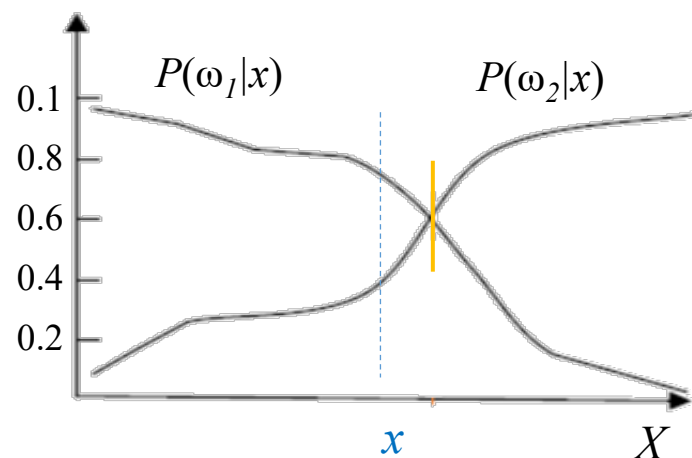
贝叶斯决策理论概念和规则

1. 几个基本概念

(全概率公式)

- **证据 (evidence)** : 观测到 x 的概率 $P(x) = \sum_i P(x, \omega_i) = \sum_i P(x|\omega_i)P(\omega_i)$
- **后验概率 (posterior probability)** : 系统在某个具体样本 x 条件下属于某种类型的概率 (分类的依据) , 需要组合先验知识和证据, 利用贝叶斯公式计算获得, 记作 $P(\omega_i|x)$, 显然 $\sum_i P(\omega_i|x)=1$
- e.g 利用贝叶斯公式分别计算 ω_1 和 ω_2 的后验概率

$$P(\omega_1|x) = \frac{\overset{\text{类条件概率}}{P(x|\omega_1)} \overset{\text{先验概率}}{P(\omega_1)}}{\sum_i \underset{\text{证据}}{P(x|\omega_i)P(\omega_i)}} = \frac{0.2 \times 0.9}{0.2 \times 0.9 + 0.4 \times 0.1} = 0.818$$
$$P(\omega_2|x) = \frac{P(x|\omega_2)P(\omega_2)}{\sum_i P(x|\omega_i)P(\omega_i)} = \frac{0.4 \times 0.1}{0.2 \times 0.9 + 0.4 \times 0.1} = 0.182$$



教材P29

2. 基于最小错误率的分类决策

- 贝叶斯分类器 (Bayes' Classifier) 选择了具有最高后验概率的类作为决策结果。这种分类是一种基于最小错误率的分类决策

● 选择 ω_i 如果 $P(\omega_i|x) = \max_i P(\omega_i|x)$ (教材式3-6)

二分类时的几个等价的判决函数形式:

(1) $g(\mathbf{x}) = P(\omega_1/\mathbf{x}) - P(\omega_2/\mathbf{x})$, (后验概率)

(2) $g(\mathbf{x}) = p(\mathbf{x}/\omega_1)P(\omega_1) - p(\mathbf{x}/\omega_2)P(\omega_2)$, (类条件概率密度)

(3) $g(\mathbf{x}) = \frac{p(\mathbf{x}/\omega_1)}{p(\mathbf{x}/\omega_2)} - \frac{P(\omega_2)}{P(\omega_1)}$, (似然比形式)

(4) $g(\mathbf{x}) = \ln \frac{p(\mathbf{x}/\omega_1)}{p(\mathbf{x}/\omega_2)} - \ln \frac{P(\omega_2)}{P(\omega_1)}$, (取对数方法)

第三讲 贝叶斯分类器



贝叶斯决策理论概念和规则

2. 分类错误率

- 假设特征向量 \mathbf{x} 是一维时， t 为 x 轴上的一点。两个决策区域:

$R_1 \sim (-\infty, t)$: 决策为 ω_1 , $P(\text{error} | x) = P(\omega_2 | x)$;

$R_2 \sim (t, +\infty)$: 决策为 ω_2 , $P(\text{error} | x) = P(\omega_1 | x)$;

- e.g 刚才例子里根据后验概率 0.818 作出属于 ω_1 决策时，实际就蕴含了 0.182 的分类错误概率

平均错误率 (期望)

$$P(e) = \int_{-\infty}^t P(\omega_2 | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} + \int_t^{\infty} P(\omega_1 | \mathbf{x}) p(\mathbf{x}) d\mathbf{x} \quad \leftarrow \text{期望的定义}$$

$$= \int_{-\infty}^t p(\mathbf{x} | \omega_2) P(\omega_2) d\mathbf{x} + \int_t^{\infty} p(\mathbf{x} | \omega_1) P(\omega_1) d\mathbf{x} \quad \leftarrow \text{贝叶斯公式展开}$$

$$P(e) = P(\omega_2) P_2(e) + P(\omega_1) P_1(e)$$

第二类错误

第一类错误

第三讲 贝叶斯分类器

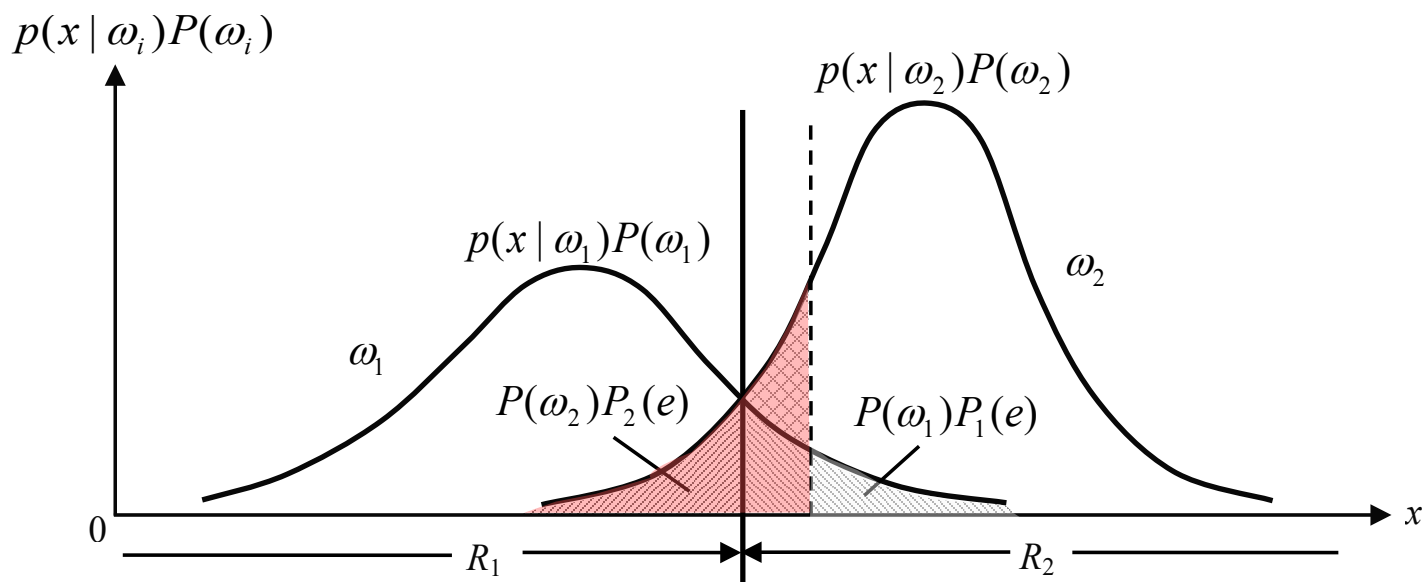


贝叶斯决策理论概念和规则

2. 分类错误率

$$P(e) = P(\omega_2)P_2(e) + P(\omega_1)P_1(e)$$

- 当 t 满足 $P(t|\omega_1)P(\omega_1) = P(t|\omega_2)P(\omega_2)$ 即 $P(\omega_1|x=t) = P(\omega_2|x=t)$ 时, 面积和最小, 说明用后验概率进行决策就是一种最小分类错误的决策



提问：最小错误率的分类决策是不是最理想的呢？

3. 基于最小风险的分​​类决策

- 错误率最小决策在很多情况下不是最佳选择。将作出判决的依据从单纯考虑后验概率最大值修改为对观测值 x 条件下各状态后验概率加权和的方式，引入一个与损失有关联的概念——风险
- 给定 K 个类别，令 λ_{ij} 代表将实属于第 j 类样本，但误分类为第 i 类所产生的**损失**，则基于后验概率将样本 x 分到第 i 类的**期望风险**为

$$R(\omega_i|x) = \sum_{j=1}^K \lambda_{ij} P(\omega_j|x)$$

- 如果希望尽可能避免将 ω_j 判断为 ω_i 则将 λ_{ij} 权重设大，表明损失的严重性

● 选择 ω_i 如果 $R(\omega_i|x) = \min_i R(\omega_i|x)$ (教材式3-8)


3. 基于最小风险的分类决策

- e.g 在前面核酸检测的实例中, $\lambda_{11} = 0$, $\lambda_{12} = 6$, $\lambda_{21} = 1$, $\lambda_{22} = 0$, 重新计算归于各类的期望风险

$$R(\omega_1|x) = 6 \times 0.182 = 1.092$$

$$R(\omega_2|x) = 1 \times 0.818 = 0.818$$

} 将样本归于 ω_2 (阳性) 的风险更小

- 在两类问题中, 如果 $\lambda_{12} - \lambda_{22} = \lambda_{21} - \lambda_{11}$, 则称为对称损失, 那么最小风险决策与最小错误率决策等价
- 一般多类问题中, 出现0-1损失 (zero-one loss) 情况  教材P29
采用0-1损失函数时, 最小风险贝叶斯决策就等价于最小错误率贝叶斯决策

第三讲 贝叶斯分类器



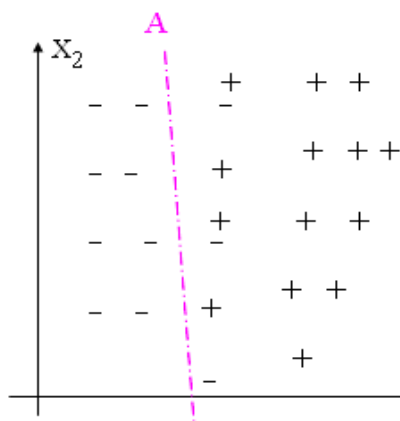
贝叶斯决策理论概念和规则

3. 基于最小风险的分分类决策

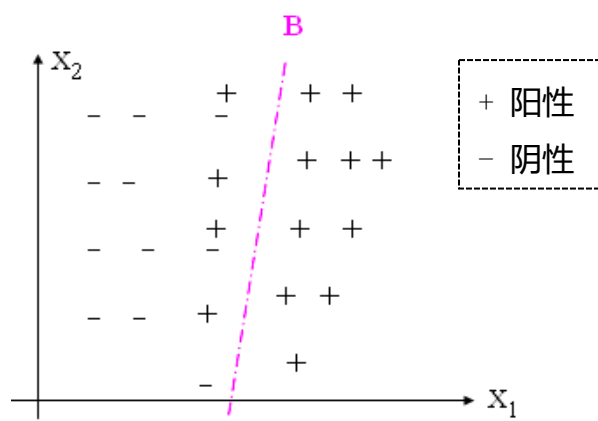
行动 α_i : 表示把模式 x 判决为 ω_i 类的一次动作

损失函数 λ_{ij} 表示模式 X 本来属于 ω_j 类判为 ω_i 所受损失。 $i \neq j$ 时都是错误判决, 故损失大于正确判断

损失 行动 \ 类别	ω_1	ω_2	ω_3		ω_m
α_1	λ_{11}	λ_{12}	λ_{13}		λ_{1m}
α_2	λ_{21}	λ_{22}	λ_{23}		λ_{2m}
α_3					
				λ_{ij}	
α_a	λ_{a1}	λ_{a2}	λ_{a3}		λ_{am}



(a) 基于最小风险 (扩大错误率, 减小损失)



(b) 基于最小错误

+ 阳性
- 阴性

第三讲 贝叶斯分类器



贝叶斯决策理论概念和规则

4. 限定错误率的两类决策*



仅作参考

- 在实际工作中，有的时候要求限制某一类错误率，并使另一类错误尽可能小，这类决策称为 “Neyman-Pearson” 决策
- e.g 设 $P_2(e)=\varepsilon_0$ 条件下，求 $P_1(e)$ 的极小值

$$L=P_1(e)+\mu(P_2(e)-\varepsilon_0)$$

← Lagrange乘数法

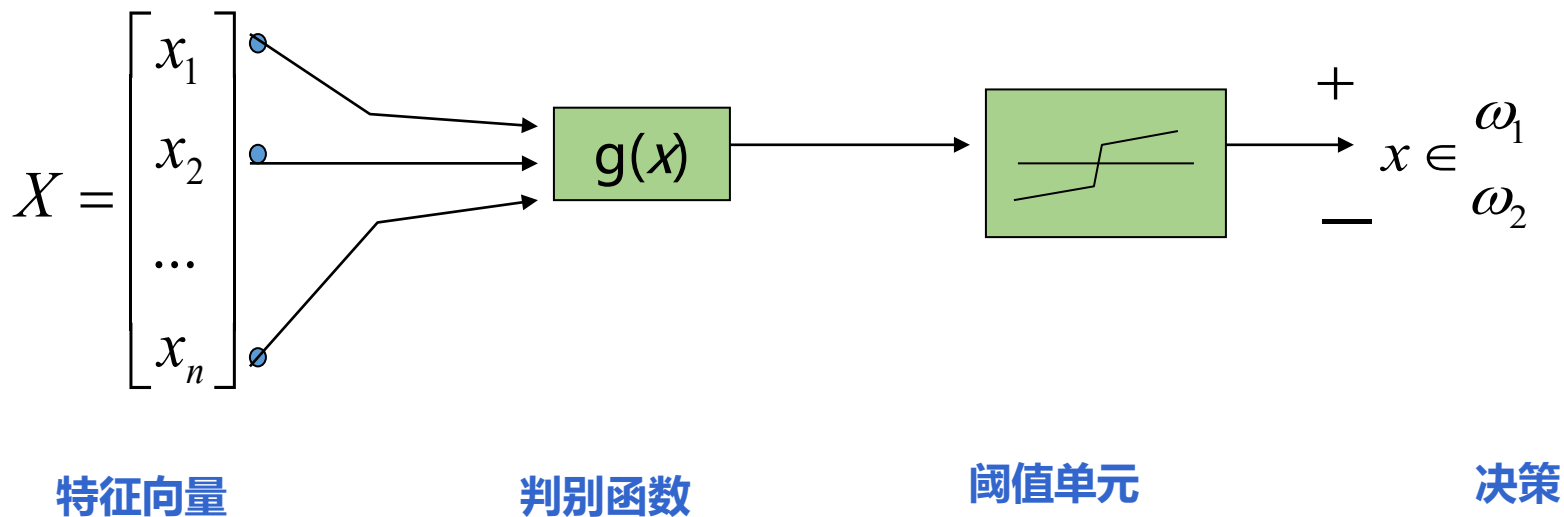
第三讲 贝叶斯分类器



贝叶斯决策理论概念和规则

5. 贝叶斯分类器

- 分类器可以视为一个计算 c 个类别的判别函数并选取最大判别值所对应的类别为决策结果的一种“机器”



二类分类器

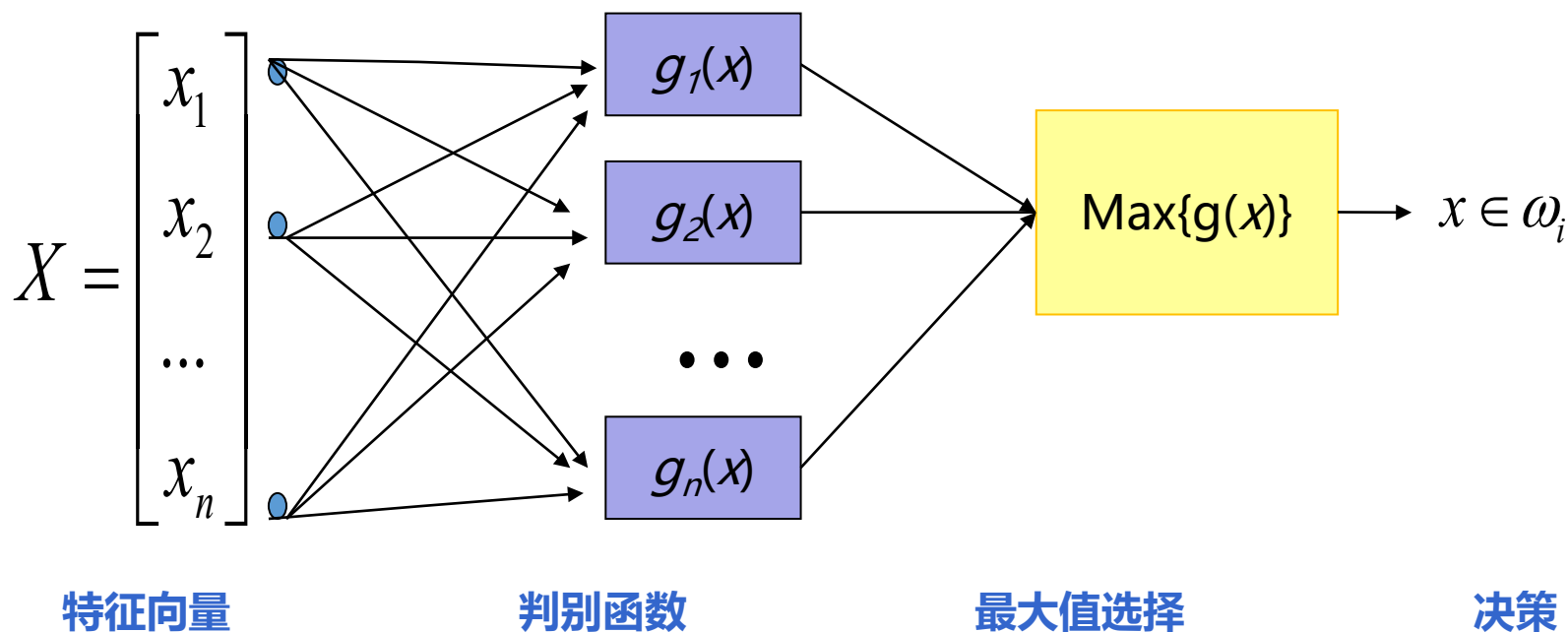
第三讲 贝叶斯分类器



贝叶斯决策理论概念和规则

5. 贝叶斯分类器

- 对于多分类问题，按照决策规则可以把多维特征空间分成 n 个类别区域，划分这些区域的边界面称之为决策面，在数学上用解析形式表示为决策面方程



多类分类器

5. 贝叶斯分类器

$P(c | x)$ 在现实中通常难以直接获得，机器学习所要实现的是基于有限的训练样本尽可能准确地估计出后验概率

两种基本策略：

判别式 (discriminative)模型

思路：直接对 $P(c | x)$ 建模

代表：

- 线性判别式
- 决策树
- BP 神经网络
- SVM

生成式(generative)模型

思路：先对联合概率分布 $P(x, c)$ 建模，再由此获得 $P(c | x)$

$$P(c | x) = \frac{P(x, c)}{P(x)}$$

代表：贝叶斯分类器

注意：贝叶斯分类器 \neq 贝叶斯学习

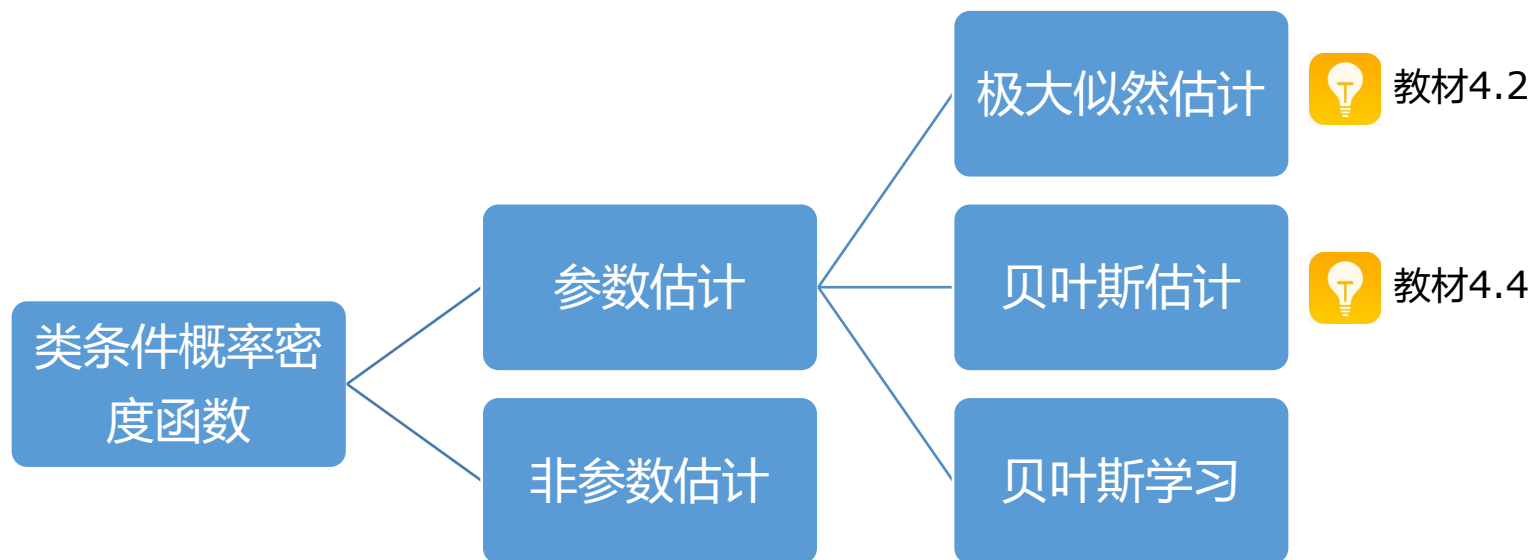
- 1 贝叶斯决策理论概念和规则
- 2 类条件概率密度函数的估计
- 3 朴素贝叶斯分类器
- 4 EM算法

第三讲 贝叶斯分类器



类条件概率密度函数的估计

- 先验概率 $P(\omega_i)$ 表达了样本空间中各类样本的比例，可以通过各类样本占总体的频率统计去获得
- 在上一节案例中，我们假设类条件概率密度函数 $P(X|\omega_i)$ 是已知的，然后去设计贝叶斯分类器。但在实际情况中类条件概率密度函数往往必须利用统计推断方法，从样本集数据中估计出来



第三讲 贝叶斯分类器



类条件概率密度函数的估计

1. 极大似然估计 (Maximum Likelihood Estimation, MLE)

- 先假设某种概率分布形式，再基于训练样例对参数进行估计。概率模型的训练过程即参数估计过程
- 假定 $P(x|\omega_i)$ 具有确定的概率分布形式，且被参数 θ_i 唯一确定，则任务就是利用训练集 D 来估计参数 θ_i 。 θ_i 对于训练集 D 中第 i 类样本组成的集合 D_i 的类似然 (likelihood) 为

$$P(D_i|\theta_i) = \prod_{x \in D_i} P(x|\theta_i)$$

寻找能最大化似然的参数值

因连乘易造成下溢，因此通常使用对数似然(log-likelihood)

$$LL(\theta_i) = \log P(D_i|\theta_i) = \sum_{x \in D_i} \log P(x|\theta_i)$$

于是 θ_i 的极大似然估计为 $\hat{\theta}_i = \underset{\theta_i}{\operatorname{argmax}} LL(\theta_i)$

由 $\frac{dLL(\theta_i)}{d\theta_i} = 0$ 求得

第三讲 贝叶斯分类器



类条件概率密度函数的估计

1. 极大似然估计

➤ 几种常见分布的极大似然估计



教材P38

(1) 伯努利密度

$$\hat{p} = \frac{\sum_t x^t}{N}$$

(2) 多项式密度

$$\hat{p}_i = \frac{\sum_t x_i^t}{N}$$

(3) 高斯密度

$$\hat{\mu}_c = \frac{1}{|D_c|} \sum_{x \in D_c} x$$

$$\hat{\sigma}_c^2 = \frac{1}{|D_c|} \sum_{x \in D_c} (x - \hat{\mu}_c)(x - \hat{\mu}_c)^T$$

第三讲 贝叶斯分类器



类条件概率密度函数的估计

1. 极大似然估计

设总体的概率密度函数是

$$f(x) = \begin{cases} 3\lambda x^2 \exp\{-\lambda x^3\}, & x > 0 \\ 0, & \text{其它} \end{cases}$$

其中 $\lambda > 0$ 是未知参数, $x_1, x_2, x_3, \dots, x_n$ 是一组样本值, 求参数 λ 的最大似然估计。

解: 似然函数 $L = \prod_{i=1}^n (3\lambda x_i^2 \exp\{-\lambda x_i^3\}) = (3^n \lambda^n \prod_{i=1}^n x_i^2 \exp\{-\lambda \sum_{i=1}^n x_i^3\})$

$$\ln L = n \ln(3\lambda) + \sum_{i=1}^n \ln x_i^2 - \lambda \sum_{i=1}^n x_i^3$$

$$\frac{d \ln L}{d \lambda} = \frac{n}{\lambda} - \sum_{i=1}^n x_i^3 = 0$$

$$\hat{\lambda} = \frac{n}{\sum_{i=1}^n x_i^3}$$

2. 贝叶斯估计

- 频率主义学派：参数看作是某种确定而未知的量，最好的估计量是在获得实际观测样本的概率为最大的条件下得到的
- 贝叶斯学派：认为参数是一个具有某种先验分布的随机变量，样本的观察使先验分布转化为后验分布，再根据后验分布来修正对参数值的估计。因此贝叶斯估计可以看作是，在假定 θ 服从 $\pi(\theta)$ 的先验分布前提下，根据样本信息去校正先验分布，得到后验分布 $P(\theta|x)$ 。由于后验分布是一个条件分布，通常我们取后验分布的期望作为参数的估计值

$$P(\theta|D) = \frac{P(D|\theta) \cdot \pi(\theta)}{\int P(D|\theta) \pi(\theta) d\theta} \quad (\text{教材式4-13})$$

$$\hat{\theta} = \int \theta \cdot P(\theta|D) d\theta = E[\theta|D] \quad (\text{教材式4-16})$$

2. 贝叶斯估计

- 贝叶斯估计 $\hat{\theta}$ 要进行积分计算，计算困难。考虑后验分布极大化而求解，这一类方法称为最大后验概率估计（Maximum a posteriori estimation, MAP）方法

$$\text{MLE: } \hat{\theta}_{mle} = \arg \max_{\theta} L(\theta|\mathbf{x}) \quad L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = f(x_1, x_2, \dots, x_n|\theta) = \prod_{i=1}^n f(x_i|\theta)$$

$$\text{MAP: } \hat{\theta}_{map} = \arg \max_{\theta} \pi(\theta|\mathbf{x}) = \arg \max_{\theta} \frac{f(\mathbf{x}|\theta)\pi(\theta)}{m(\mathbf{x})} = \arg \max_{\theta} f(\mathbf{x}|\theta)\pi(\theta)$$

- 如果将机器学习结构风险中的正则化项对应为上式的，那么带有正则化项的最大似然学习就可以被解释为MAP。MAP与MLE最大的不同在于 $\pi(\theta)$ 项，正好可以解决MLE缺乏先验知识的缺点。 $\pi(\theta)$ 项正好起到了正则化的作用。如：如果假设 $\pi(\theta)$ 服从高斯分布，则相当于加了一个L2 norm；如果假设 $\pi(\theta)$ 服从拉普拉斯分布，则相当于加了一个L1 norm

- 1 贝叶斯决策理论概念和规则
- 2 类条件概率密度函数的估计
- 3 朴素贝叶斯分类器
- 4 EM算法

1. 朴素贝叶斯分类的含义

- 贝叶斯公式估计后验概率困难在于，类条件概率是所有属性上的联合概率，难以从有限的样本之间估计而得。朴素 (Naive) 贝叶斯推断是在贝叶斯推断的基础上，对条件概率分布做了条件独立性的假设。即，假设每个属性独立地对分类结果发生影响

$$P(c | \mathbf{x}) = \frac{P(c)P(\mathbf{x} | c)}{P(\mathbf{x})} = \frac{P(c)}{P(\mathbf{x})} \prod_{i=1}^d P(x_i | c)$$

贝叶斯判定准则：

$$h_{nb}(x) = \arg \max_{c \in Y} P(c) \prod_{i=1}^d P(x_i | c)$$

- 朴素贝叶斯分类器的训练过程就是基于训练集D来估计类先验概率 $P(c)$ ，并为每个属性估计类条件概率

1. 朴素贝叶斯分类的含义

- 令 D_c 表示训练集 D 中第 c 类样本组成的集合，若有充足的独立同分布样本，则可以估计出类先验概率

$$P(c) = \frac{|D_c|}{|D|}$$

- 对离散属性，则条件概率 $P(x_i|c)$ 可估计为

$$P(x_i | c) = \frac{|D_{c_i x_i}|}{|D_c|}$$

- 对连续属性，假设样本分布满足正态分布，则条件概率 $P(x_i|c)$ 估计可利用均值和方差

$$P(x_i | c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right)$$

第三讲 贝叶斯分类器



朴素贝叶斯分类器

2. 朴素贝叶斯分类的例子



西瓜书P151

帅？	性格好？	身高？	上进？	嫁与否
帅	不好	矮	不上进	不嫁
不帅	好	矮	上进	不嫁
帅	好	矮	上进	嫁
不帅	好	高	上进	嫁
帅	不好	矮	上进	不嫁
帅	不好	矮	上进	不嫁
帅	好	高	不上进	嫁
不帅	好	中	上进	嫁
帅	好	中	上进	嫁
不帅	不好	高	上进	嫁
帅	好	矮	不上进	不嫁
帅	好	矮	不上进	不嫁

问题：一个{不帅、性格不好、身高矮、不上进}的男生嫁还是不嫁

第三讲 贝叶斯分类器



朴素贝叶斯分类器

2. 朴素贝叶斯分类的例子



西瓜书P151

➤ 判别器：

$P1\{\text{嫁} | (\text{不帅、性格不好、身高矮、不上进})\}$

$P2\{\text{不嫁} | (\text{不帅、性格不好、身高矮、不上进})\}$ 的概率

➤ 计算：

$$\begin{aligned} p(\text{嫁} | \text{不帅、性格不好、身高矮、不上进}) &= \frac{p(\text{不帅、性格不好、身高矮、不上进} | \text{嫁}) * p(\text{嫁})}{p(\text{不帅、性格不好、身高矮、不上进})} \\ &= \frac{p(\text{不帅} | \text{嫁}) * p(\text{性格不好} | \text{嫁}) * p(\text{身高矮} | \text{嫁}) * p(\text{不上进} | \text{嫁}) * p(\text{嫁})}{p(\text{不帅}) * p(\text{性格不好}) * p(\text{身高矮}) * p(\text{不上进})} \end{aligned}$$

关键在于：

$P(\text{不帅、性格不好、身高矮、不上进} | \text{嫁})$

$= P(\text{不帅} | \text{嫁}) \times P(\text{性格不好} | \text{嫁}) \times P(\text{身高矮} | \text{嫁}) \times P(\text{不上进} | \text{嫁})$

特征
独立

第三讲 贝叶斯分类器



朴素贝叶斯分类器

2. 朴素贝叶斯分类的例子



西瓜书P151

$$p(\text{嫁}) = 6/12 \text{ (总样本数)} = 1/2$$

$$p(\text{不帅}|\text{嫁}) = 3/6 = 1/2$$

$$p(\text{性格不好}|\text{嫁}) = 1/6$$

$$p(\text{矮}|\text{嫁}) = 1/6$$

$$p(\text{不上进}|\text{嫁}) = 1/6$$

$$p(\text{不帅}) = 4/12 = 1/3$$

$$p(\text{性格不好}) = 4/12 = 1/3$$

$$p(\text{身高矮}) = 7/12$$

$$p(\text{不上进}) = 4/12 = 1/3$$

$$p(\text{不嫁}) = 6/12 = 1/2$$

$$p(\text{不帅}|\text{不嫁}) = 1/6$$

$$p(\text{性格不好}|\text{不嫁}) = 3/6 = 1/2$$

$$p(\text{矮}|\text{不嫁}) = 6/6 = 1$$

$$p(\text{不上进}|\text{不嫁}) = 3/6 = 1/2$$

$$p(\text{不帅}) = 4/12 = 1/3$$

$$p(\text{性格不好}) = 4/12 = 1/3$$

$$p(\text{身高矮}) = 7/12$$

$$p(\text{不上进}) = 4/12 = 1/3$$

$P2\{\text{不嫁}|\text{不帅、性格不好、身高矮、不上进}\} > P1\{\text{嫁}|\text{不帅、性格不好、身高矮、不上进}\}$

第三讲 贝叶斯分类器



朴素贝叶斯分类器

2. 朴素贝叶斯分类的例子



西瓜书P151

表 4.4 西瓜数据集 2.0 α

编号	色泽	根蒂	敲声	纹理	脐部	触感	好瓜
1	-	蜷缩	浊响	清晰	凹陷	硬滑	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	-	是
3	乌黑	蜷缩	-	清晰	凹陷	硬滑	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	是
5	-	蜷缩	浊响	清晰	凹陷	硬滑	是
6	青绿	稍蜷	浊响	清晰	-	软粘	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	是
8	乌黑	稍蜷	浊响	-	稍凹	硬滑	是
9	乌黑	-	沉闷	稍糊	稍凹	硬滑	否
10	青绿	硬挺	清脆	-	平坦	软粘	否
11	浅白	硬挺	清脆	模糊	平坦	-	否
12	浅白	蜷缩	-	模糊	平坦	软粘	否
13	-	稍蜷	浊响	稍糊	凹陷	硬滑	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	否
15	乌黑	稍蜷	浊响	清晰	-	软粘	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	否
17	青绿	-	沉闷	稍糊	稍凹	硬滑	否

表 4.5 西瓜数据集 3.0 α

编号	密度	含糖率	好瓜
1	0.697	0.460	是
2	0.774	0.376	是
3	0.634	0.264	是
4	0.608	0.318	是
5	0.556	0.215	是
6	0.403	0.237	是
7	0.481	0.149	是
8	0.437	0.211	是
9	0.666	0.091	否
10	0.243	0.267	否
11	0.245	0.057	否
12	0.343	0.099	否
13	0.639	0.161	否
14	0.657	0.198	否
15	0.360	0.370	否
16	0.593	0.042	否
17	0.719	0.103	否

判决测1样本是不是好瓜

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
测 1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	?

第三讲 贝叶斯分类器



朴素贝叶斯分类器

2. 朴素贝叶斯分类的例子



西瓜书P151

$$P(\text{好瓜} = \text{是}) = \frac{8}{17} \approx 0.471,$$

$$P(\text{好瓜} = \text{否}) = \frac{9}{17} \approx 0.529.$$

先验概率

$$P_{\text{青绿}|\text{是}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{是}) = \frac{3}{8} = 0.375,$$

$$P_{\text{青绿}|\text{否}} = P(\text{色泽} = \text{青绿} | \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333,$$

$$P_{\text{蜷缩}|\text{是}} = P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{是}) = \frac{5}{8} = 0.375,$$

$$P_{\text{蜷缩}|\text{否}} = P(\text{根蒂} = \text{蜷缩} | \text{好瓜} = \text{否}) = \frac{3}{9} \approx 0.333,$$

$$P_{\text{浊响}|\text{是}} = P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750,$$

$$P_{\text{浊响}|\text{否}} = P(\text{敲声} = \text{浊响} | \text{好瓜} = \text{否}) = \frac{4}{9} \approx 0.444,$$

$$P_{\text{清晰}|\text{是}} = P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{是}) = \frac{7}{8} = 0.875,$$

$$P_{\text{清晰}|\text{否}} = P(\text{纹理} = \text{清晰} | \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222,$$

$$P_{\text{凹陷}|\text{是}} = P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750,$$

$$P_{\text{凹陷}|\text{否}} = P(\text{脐部} = \text{凹陷} | \text{好瓜} = \text{否}) = \frac{2}{9} \approx 0.222,$$

$$P_{\text{硬滑}|\text{是}} = P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{是}) = \frac{6}{8} = 0.750,$$

$$P_{\text{硬滑}|\text{否}} = P(\text{触感} = \text{硬滑} | \text{好瓜} = \text{否}) = \frac{6}{9} \approx 0.667,$$

请大家自己算一下
这几个数字

$$P_{\text{密度: 0.697}|\text{是}} = p(\text{密度} = 0.697 | \text{好瓜} = \text{是})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.129} \exp\left(-\frac{(0.697 - 0.574)^2}{2 \cdot 0.129^2}\right) \approx 1.959,$$

$$P_{\text{密度: 0.697}|\text{否}} = p(\text{密度} = 0.697 | \text{好瓜} = \text{否})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.195} \exp\left(-\frac{(0.697 - 0.496)^2}{2 \cdot 0.195^2}\right) \approx 1.203,$$

$$P_{\text{含糖: 0.460}|\text{是}} = p(\text{含糖率} = 0.460 | \text{好瓜} = \text{是})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.101} \exp\left(-\frac{(0.460 - 0.279)^2}{2 \cdot 0.101^2}\right) \approx 0.788,$$

$$P_{\text{含糖: 0.460}|\text{否}} = p(\text{含糖率} = 0.460 | \text{好瓜} = \text{否})$$

$$= \frac{1}{\sqrt{2\pi} \cdot 0.108} \exp\left(-\frac{(0.460 - 0.154)^2}{2 \cdot 0.108^2}\right) \approx 0.066.$$

离散特征的类似然

连续特征的类似然

第三讲 贝叶斯分类器



朴素贝叶斯分类器

2. 朴素贝叶斯分类的例子



西瓜书P151

$$\begin{aligned} P(\text{好瓜} = \text{是}) \times P_{\text{青绿}|\text{是}} \times P_{\text{蜷缩}|\text{是}} \times P_{\text{浊响}|\text{是}} \times P_{\text{清晰}|\text{是}} \times P_{\text{凹陷}|\text{是}} \\ \times P_{\text{硬滑}|\text{是}} \times P_{\text{密度: 0.697}|\text{是}} \times P_{\text{含糖: 0.460}|\text{是}} \approx 0.038, \\ P(\text{好瓜} = \text{否}) \times P_{\text{青绿}|\text{否}} \times P_{\text{蜷缩}|\text{否}} \times P_{\text{浊响}|\text{否}} \times P_{\text{清晰}|\text{否}} \times P_{\text{凹陷}|\text{否}} \\ \times P_{\text{硬滑}|\text{否}} \times P_{\text{密度: 0.697}|\text{否}} \times P_{\text{含糖: 0.460}|\text{否}} \approx 6.80 \times 10^{-5}. \end{aligned}$$



- 需要注意的是：若某个属性值在训练集中和某个类别没有一起出现过，这样会抹掉其它的属性信息，因为该样本的类条件概率被计算为0
- 因此在估计概率值时，常常用进行平滑（smoothing）处理，**拉普拉斯修正（Laplacian correction）**就是其中的经典方法，具体计算方法如下：

$$\begin{aligned} \hat{P}(c) &= \frac{|D_c| + 1}{|D| + N}, \quad \text{出现的类别数} & P_{\text{清脆}|\text{是}} &= P(\text{敲声} = \text{清脆} | \text{好瓜} = \text{是}) = \frac{0}{8} = 0 \\ \hat{P}(x_i | c) &= \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}, \quad \text{属性}x_i\text{可能的取值数} & \hat{P}_{\text{清脆}|\text{是}} &= \hat{P}(\text{敲声} = \text{清脆} | \text{好瓜} = \text{是}) = \frac{0+1}{8+3} \approx 0.091 \end{aligned}$$

3. 小结

优点:

1. 朴素贝叶斯模型有稳定的分类效率
2. 对小规模的数据表现很好，能处理多分类任务，适合增量式训练，尤其是数据量超出内存时，可以一批批的去增量训练
3. 对缺失数据不太敏感，算法也比较简单，常用于文本分类

缺点:

1. 需要知道先验概率，且先验概率很多时候取决于假设，假设的模型可以有很多种，因此在某些时候会由于假设的先验模型的原因导致预测效果不佳
2. 因为是通过先验和数据来决定后验的概率从而决定分类，所以分类决策存在一定的错误率
3. 对输入数据的表达形式很敏感

第三讲 贝叶斯分类器



朴素贝叶斯分类器

思考题：

例 4.1 试由表 4.1 的训练数据学习一个朴素贝叶斯分类器并确定 $x = (2, S)^T$ 的类标记 y . 表中 $X^{(1)}$, $X^{(2)}$ 为特征, 取值的集合分别为 $A_1 = \{1, 2, 3\}$, $A_2 = \{S, M, L\}$, Y 为类标记, $Y \in C = \{1, -1\}$.

表 4.1 训练数据

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	-1	-1	1	1	-1	-1	-1	1	1	1	1	1	1	1	-1

解 根据算法 4.1, 由表 4.1, 容易计算下列概率:

https://blog.csdn.net/sinat_30353259

第三讲 贝叶斯分类器



朴素贝叶斯分类器

思考题：

$$P(Y=1)=\frac{9}{15}, \quad P(Y=-1)=\frac{6}{15}$$

$$P(X^{(1)}=1|Y=1)=\frac{2}{9}, \quad P(X^{(1)}=2|Y=1)=\frac{3}{9}, \quad P(X^{(1)}=3|Y=1)=\frac{4}{9}$$

$$P(X^{(2)}=S|Y=1)=\frac{1}{9}, \quad P(X^{(2)}=M|Y=1)=\frac{4}{9}, \quad P(X^{(2)}=L|Y=1)=\frac{4}{9}$$

$$P(X^{(1)}=1|Y=-1)=\frac{3}{6}, \quad P(X^{(1)}=2|Y=-1)=\frac{2}{6}, \quad P(X^{(1)}=3|Y=-1)=\frac{1}{6}$$

$$P(X^{(2)}=S|Y=-1)=\frac{3}{6}, \quad P(X^{(2)}=M|Y=-1)=\frac{2}{6}, \quad P(X^{(2)}=L|Y=-1)=\frac{1}{6}$$

对于给定的 $x=(2,S)^T$ 计算：

$$P(Y=1)P(X^{(1)}=2|Y=1)P(X^{(2)}=S|Y=1)=\frac{9}{15} \cdot \frac{3}{9} \cdot \frac{1}{9}=\frac{1}{45}$$

$$P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)=\frac{6}{15} \cdot \frac{2}{6} \cdot \frac{3}{6}=\frac{1}{15}$$

因为 $P(Y=-1)P(X^{(1)}=2|Y=-1)P(X^{(2)}=S|Y=-1)$ 最大，所以 $y=-1$ 。

- 1 贝叶斯决策理论概念和规则
- 2 类条件概率密度函数的估计
- 3 朴素贝叶斯分类器
- 4 EM算法

第三讲 贝叶斯分类器



EM方法

- 在前面讨论中，我们一直假设训练样本所有属性的变量值是可以被观测到的。在实际工作中，有些特征属性因为客观原因无法观测，样本数据“不完整”。此时的参数估计就需要使用期望最大化算法EM(Expectation-Maximization)

- 未观测变量：隐变量(latent variable)

令 X 表示已观测变量集， Z 表示隐变量集，欲对模型参数极大似然估计，则应最大化对数似然函数。 Z 是隐变量，无法直接求解。怎么办？

- e.g1 我们有200个核酸检测数据，但我们并不知道这些样本来自阴性群体 ω_1 还是阳性群体 ω_2
- e.g2 我们有200个身高数据，但我们并不知道这些样本来自成人还是小孩

第三讲 贝叶斯分类器



EM方法

- 在前面讨论中，我们一直假设训练样本所有属性的变量值是可以被观测到的。在实际工作中，有些特征属性因为客观原因无法观测，样本数据“不完整”。此时的参数估计就需要使用期望最大化算法EM(Expectation-Maximization)

- 未观测变量：隐变量(latent variable)

令 X 表示已观测变量集， Z 表示隐变量集，欲对模型参数极大似然估计，则应最大化对数似然函数。 Z 是隐变量，无法直接求解。怎么办？

- e.g1 我们有200个核酸检测数据，但我们并不知道这些样本来自阴性群体 ω_1 还是阳性群体 ω_2
- e.g2 我们有200个身高数据，但我们并不知道这些样本来自成人还是小孩

第三讲 贝叶斯分类器



EM方法

- 对隐变量 Z 计算期望，最大化已观测数据的对数“边际似然” (marginal likelihood)

$$LL(\Theta | X) = \ln P(X | \Theta) = \ln \sum_Z P(X, Z | \Theta)$$

基本思想：

E步（期望）：若参数 Θ 已知，则可根据训练数据推断出最优隐变量 Z 的值；

M步（最大化）：若 Z 值已知，则可对参数 Θ 做极大似然估计

以初始值 Θ^0 为起点，**迭代执行以下步骤直至收敛到局部最优**：

- 基于 Θ^t 推断隐变量 Z 的期望，记为 Z^t
- 基于已观测变量 X 和 Z^t 对参数 Θ 做极大似然估计，记为 Θ^{t+1}

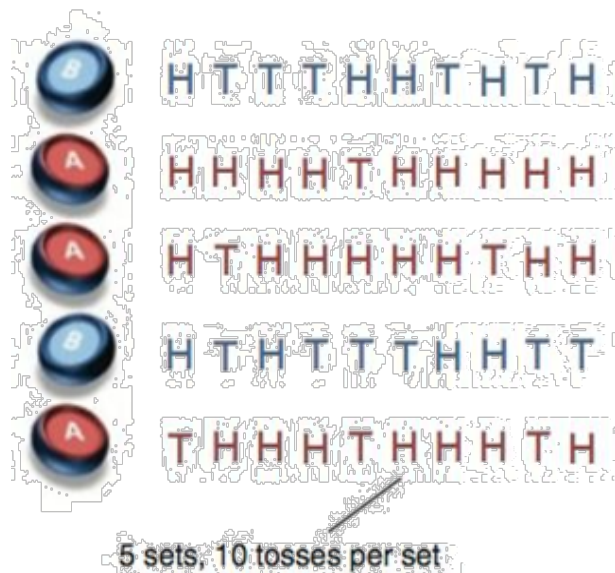
第三讲 贝叶斯分类器



EM方法

- Stanford机器学习经典例子：有A、B两个硬币，根据5组样本，计算其抛至正面（Head, 记为H）与反面（Text, 记为T）的概率

a Maximum likelihood



Coin A	Coin B
	5 H, 5 T
9 H, 1 T	
8 H, 2 T	
	4 H, 6 T
7 H, 3 T	
24 H, 6 T	9 H, 11 T

$$\hat{\theta}_A = \frac{24}{24 + 6} = 0.80$$

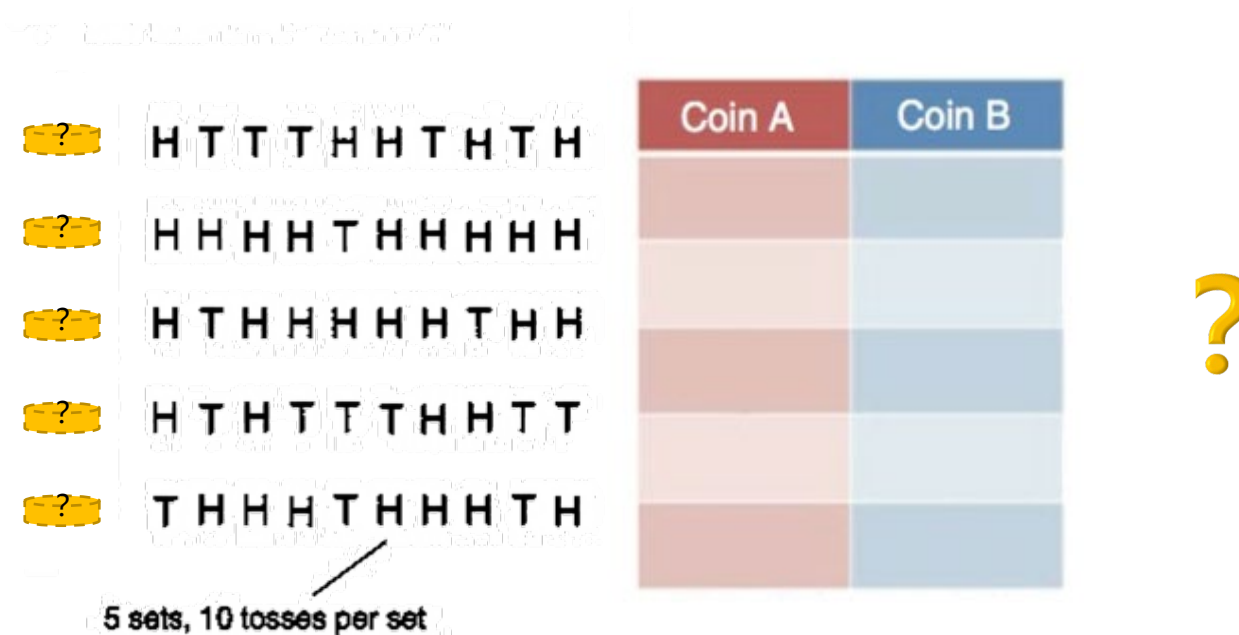
$$\hat{\theta}_B = \frac{9}{9 + 11} = 0.45$$

第三讲 贝叶斯分类器



EM方法

- Stanford机器学习经典例子：有A、B两个硬币，根据5组样本，计算其抛至正面（Head, 记为H）与反面（Text, 记为T）的概率

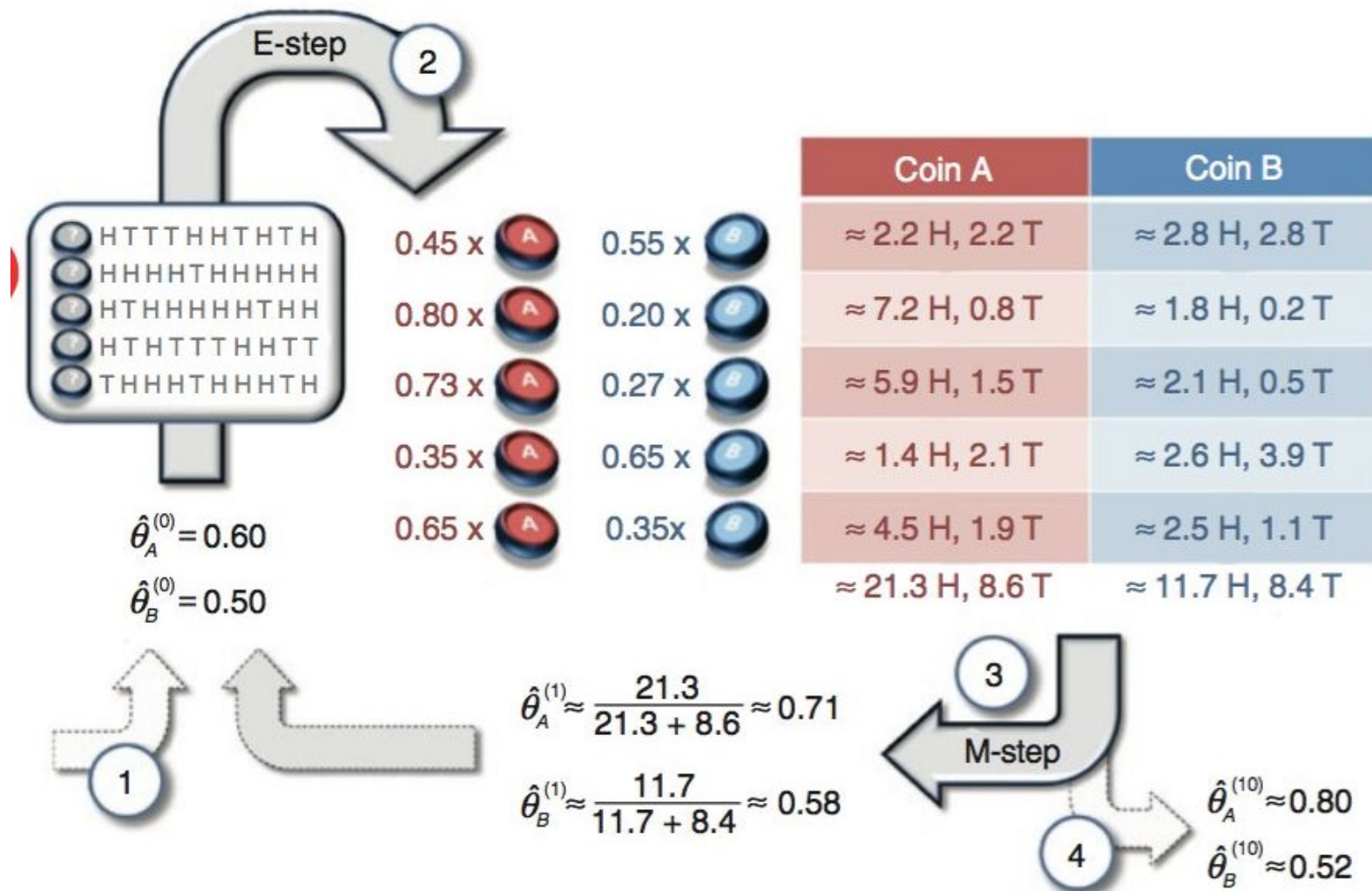


第三讲 贝叶斯分类器



EM方法

b Expectation maximization



第三讲 贝叶斯分类器



EM方法

- 期望步 (E-step) 利用当前估计的参数值来计算似然函数的期望值, 获得 Z 的分布估计; 最大化步 (M-step) 寻找使E步产生的似然期望最大化的参数, 然后将新的参数用于E步...直到收敛到局部最优解
- EM算法一定可以收敛。但是否收敛到最优解以及收敛速度, 与初始值有关
 - e.g2 我们有200个身高数据, 但我们并不知道这些样本来自成人还是小孩
 - (1) 初始化参数: 设成人身高的正态分布的参数: 均值=1.7, 方差=0.1
 - (2) 计算每一个人更可能属于成人分布或者小孩分布;
 - (3) 通过分为成人的 n 个人来重新估计成人身高分布的参数 (最大似然估计), 小孩分布也按照相同的方式估计出来, 更新分布。
 - (4) 这时候两个分布的概率也变了, 然后重复步骤 (1) 至 (3), 直到参数不发生变化为止。

参考阅读《高斯混合模型 (GMM) 及其EM算法的理解》