第六章 贝叶斯分类

§6.1 背景知识

§6.2 贝叶斯决策

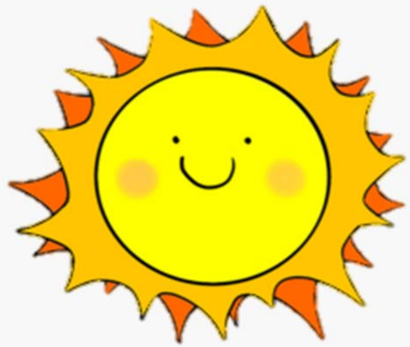
§6.3 朴素贝叶斯分类器

§6.1 背景知识

一、问题引入

问题引入

- 简例1
  - 分类可以看作一种决策
  - 预测明天北京的天气，这就是一个决策问题
  - 假如限定明天只有晴天和雨天两种天气情况，预测明天天气就是一个典型的两类分类问题
  - 不借助任何相关天气预报方面的信息，因为最近北京都是晴天，因此认为晴天的可能性更大，这就是决策
  - 这个决策过程是有理论依据的，因为通过过去几天的天气作出了粗略分析，认为晴天的概率更大，所以选择了概率较大的决策



问题引入

简例1

- 将明日的天气记作x，将晴天雨天的判断分别记作w_1和w_2，并用$P(w_1)$与$P(w_2)$代表两类的概率，上述的决策规则可以表示为：

  如果 $P(w_1) > P(w_2)$，则$x \in w_1$；反之，则$x \in w_2$

- 在二分类的情况下，$P(w_1) + P(w_2) = 1$。如果决策$x \in w_1$，那么犯错误的概率为$P(error) = 1 - \mathrm{P}(w_1) = \mathrm{P}(w_2)$，反之亦然。很显然，这样做决策的犯错误概率是最小的，也被称为**最小错误准则**

- 这里例子中的概率是在没有对样本进行任何观测情况下的概率，我们称它为**先验概率**

- 如果我们得到了观测天气的某相关指标 y，如何根据该指标进行决策呢？

问题引入

简例1

- 记该指标为y，晴天和雨天的概率可被记为$P(w_1|\,y)$与$P(w_2|\,y)$，这种概率我们称为**后验概率**。这时候的决策规则应该是：

  如果 $P(w_1|\,y) > P(w_2|\,y)$，则$x \in w_1$；反之，则$x \in w_2$

- 根据概率论中的贝叶斯公式，有

$$P(w_i|\,y) = \frac{\mathrm{p}(\mathrm{y},\mathrm{w_i})}{\mathrm{p}(\mathrm{y})} = \frac{p(y|w_i)P(w_i)}{p(y)},$$

  其中，$P(w_i)$是先验概率，$\mathrm{p}(\mathrm{y},w_i)$是联合概率密度，$p(y)$是该指标的概率密度，称为总体密度，$p(y|w_i)$是第 i 种天气对应该指标的概率密度，称为**类条件概率密度**

- 这样，后验概率就转换成了先验概率与类条件密度的乘积，再用总体密度归一化

问题引入

□ 简例1

- 通过贝叶斯公式，我们可以化简刚才的决策规则：

  如果$p(y|w_1)P(w_1) > p(y|w_2)P(w_2)$，则$x \in w_1$；反之，则$x \in w_2$

  其中，先验概率我们可以通过历史的天气数据进行分析和统计，而类条件密度则需要用一定的属于本类的训练样本进行估计

- 这就是贝叶斯决策：在类条件概率密度和先验概率已知（或者可以估计）的条件下，通过贝叶斯公式比较样本属于两类的后验概率，将类别决策为后验概率较大的一类，这样做的目的是为了使决策的总体错误率最小
- 这个例子可以用来直观说明贝叶斯决策的基本思想
- 是否可以再列举一个例子阐述贝叶斯决策的思想？

问题引入

简例2

帅?	性格好?	身高?	上进?	谈朋友
帅	不好	矮	不上进	不谈
不帅	好	矮	上进	不谈
帅	好	矮	上进	谈
不帅	好	高	上进	谈
帅	不好	矮	上进	不谈
帅	不好	矮	上进	不谈
帅	好	高	不上进	谈
不帅	好	中	上进	谈
帅	好	中	上进	谈
不帅	不好	矮	上进	谈
帅	好	矮	不上进	不谈
帅	好	矮	不上进	不谈

§6.2 贝叶斯决策

一、贝叶斯决策理论

二、概率密度函数估计

贝叶斯决策理论

□ 贝叶斯决策理论

- 贝叶斯决策理论也称统计决策理论
- 我们约定样本$x \in R^d$是由 d 维实数特征组成的，即$x = [x_1, x_2, \dots, x_d]^T$
- 假定要研究的分类类别数有 c 个，记作w_i。**类别数c**已知，**各类的先验概率**也已知，各类中样本的分布密度（**即类条件密度$p(x|w_i)$**）也是已知的。我们所要做的决策就是，对于某个未知样x，判断其属于哪一类
- 任一决策都可能会有错误。因此，对于两类问题，在样本x上的错误率为

$$P(\mathrm{e}|x) = \begin{cases} P(w_2|x) & \text{如果决策 } x \in w_1 \\ P(w_1|x) & \text{如果决策 } x \in w_2 \end{cases}$$

- 错误率定义为**所有服从同样分布的独立样本上错误概率的期望**，即

$$P(e) = \int P(\mathrm{e}|x)p(x)\,\mathrm{d}x$$

最小错误率贝叶斯决策

- 在常见的模式识别问题中，我们往往希望尽可能减少分类的错误，即**追求最小错误率**
- 从最小错误率的要求出发，利用概率论中的贝叶斯公式，就能得到使错误率最小的分类决策，我们称之为**最小错误率贝叶斯决策**
- 对于贝叶斯决策，我们定义了错误率：

$$P(e)=\int P(\mathrm{e}|x)p(x)\,\mathrm{d}x$$

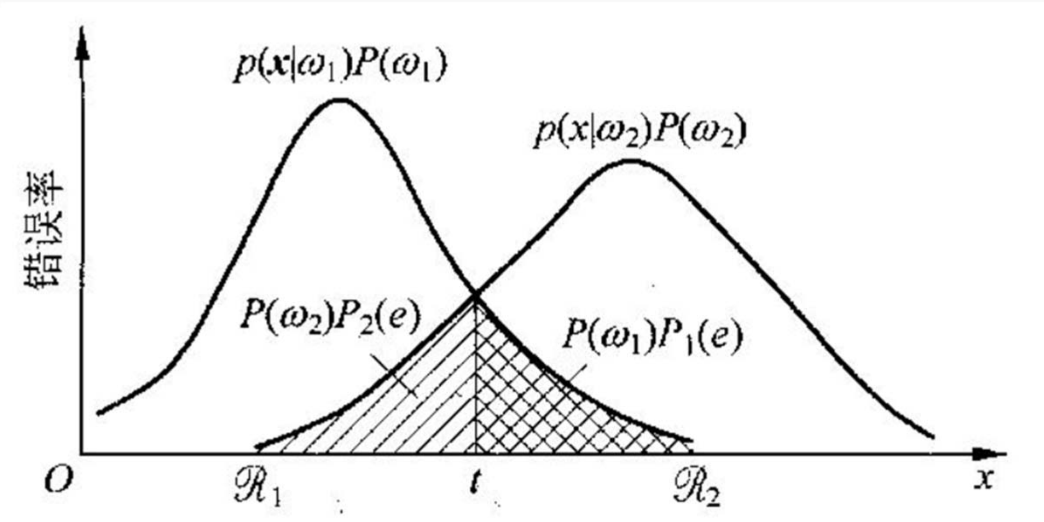
- 而最小错误率贝叶斯决策就是最小化该错误率，优化目标可表示为：

$$min\ \ \mathrm{P}(e)=\int P(\mathrm{e}|x)p(x)\,\mathrm{d}x$$

最小错误率贝叶斯决策

- 而最小错误率贝叶斯决策就是最小化该错误率，优化目标可表示为：

$$min \ \mathrm{P}(e) = \int P(\mathrm{e}|x)p(x)\,\mathrm{d}x$$



- 通过选择合适的分类面（决策面）使总体错误率最小
- 直观地来说，错误率为分类面左侧$p(x|w_2)P(w_2)$曲线下面积与分类面右侧$p(x|w_1)P(w_1)$曲线下面积之和
- 所以在最小错误率条件下，贝叶斯决策的分类面在当x满足$p(x|w_1)P(w_1) = p(x|w_2)P(w_2)$

最小错误率贝叶斯决策

- 最小错误率贝叶斯决策采用使后验概率最大的决策，对于两类问题，有如下**决策规则**：

  如果 $P(w_1|\, x) > P(w_2|\, x)$，则$x \in w_1$；反之，则$x \in w_2$

- 后验概率可以用贝叶斯公式求得：

$$P(w_i|x) = \frac{p(x|w_i)P(w_i)}{p(x)} = \frac{p(x|w_i)P(w_i)}{\sum_{j=1}^{2} p(x|w_j)P(w_j)}, \qquad i = 1, 2$$

- 先验概率$P(w_i)$与类条件密度$p(x|w_i)$都已知
- 没有特殊说明下，贝叶斯决策通常指最小错误率贝叶斯决策

最小错误率贝叶斯决策

- 最小错误率贝叶斯决策规则可以表示为多种等价形式，如：
  - 若 $P(w_i|\, x) = max\, \mathrm{P}(w|\, x)$，则 $x \in \mathrm{w_i}$
  - 后验概率对比中，由于总体密度相等，所以决策只需要比较分子

    若 $p(x|w_i)\mathrm{P}(\mathrm{w}_i) = max\, p(x|w)P(w)$，则 $x \in \mathrm{w_i}$
  - 由于先验概率$\mathrm{P}(\mathrm{w_i})$是事先确定的，与当前样本x无关，因此，人们经常把决策规则整理成如下形式，即：

$$l(x) = \frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_2)}{P(w_1)} = \lambda, \text{则} x \in \mathrm{w_1}$$

通过这种方式，可以先算出似然比阈值λ，对每一个样本计算$l(x)$，与λ进行比较，大于阈值的则决策为第一类，小于阈值则决策为第二类

最小错误率贝叶斯决策

- 最小错误率贝叶斯决策规则可以表示为多种等价形式，如：
  - 很多情况下，用对数形式进行计算可能简化计算过程。因此我们定义了对数似然比

$$h(x) = -\ln[\, l(x)] = -\ln\,[p(x|w_1)] + \ln\,[p(x|w_2)]$$

决策规则变为如下形式：

$$h(x) < \ln\frac{P(w_1)}{P(w_2)},\ 则 x \in \mathrm{w}_1$$

概率密度值 $p(x|w_1)$反映了在w_1类中观察到x 的可能性，被称为似然度，$l(x)$ 被称为似然比，$h(x)$被称为对数似然比

最小错误率贝叶斯决策

总结

$$min\ \ \mathrm{P}(e) = \int \mathrm{P}(\mathrm{e}|\boldsymbol{x})p(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}$$ 等价于求解 $$\min_{\boldsymbol{x}} \mathrm{P}(\mathrm{e}|\boldsymbol{x}) \text{ for } \forall \boldsymbol{x}$$

最小错误率贝叶斯决策

- 如果 $\boldsymbol{P(w_1|\,x) > P(w_2|\,x)}$，则$\boldsymbol{x} \in \boldsymbol{w_1}$；反之，则$\boldsymbol{x} \in \boldsymbol{w_2}$

  通过贝叶斯公式得到等价形式

- 若 $\boldsymbol{p(x|w_i)\mathrm{P}(\mathrm{w}_i) = max\, p(x|w)P(w)}$，则 $\boldsymbol{x} \in \mathbf{w_i}$
- 若 $\boldsymbol{l(x) = \frac{p(x|w_1)}{p(x|w_2)} > \frac{P(w_2)}{P(w_1)} = \lambda}$,则$\boldsymbol{x} \in \mathbf{w_1}$

最小风险贝叶斯决策

- 最小错误率默认平等对待每一种决策错误，但是在真实复杂情况下，**不同的决策错误所带来的损失很可能不同**
- 两个例子：**①癌细胞识别**：把正常细胞判定为癌细胞，会带来病人精神上的负担和更多的检查，这是一种损失；但是将癌细胞判定为正常细胞，则给病人带来的损失更大，会导致病人失去最佳治疗时机；**②人脸识别**：把小偷识别成注册用户和把注册用户识别成小偷，对人脸识别所带来的损失也不一样，也需要区别对待
- 所谓最小风险贝叶斯决策，就是**考虑各种错误所带来的损失不同，从而做出一种基于最小风险的最优决策**

最小风险贝叶斯决策

最小风险贝叶斯决策

- 我们约定样本$x \in R^d$ 是由 d 维实数特征组成的，即$x = [x_1, x_2, \ldots, x_d]^T$
- 状态空间Ω由c个可能的状态（c类）组成：$\Omega = \{w_1, w_2, \ldots, w_c\}$
- 对随机向量x可能采取的决策组成了决策空间，它由k个决策组成

$$A = \{a_1, a_2, \ldots, a_k\}$$

- 设对于实际状态为w_j的向量x, 采取决策a_i所带来的损失为：

$$\lambda(a_i, w_j), \quad i = 1, \ldots, k, \quad j = 1, \ldots, c$$

称为**损失函数**

最小风险贝叶斯决策

最小风险贝叶斯决策

- 对于某个样本x，它属于各个状态的后验概率是$P(w_j|x)$,$j=1,\dots,c$，对它采取决策$a_i, i=1,\dots,k$的期望损失是

$$R(a_i|x)=E[\lambda(a_i,w_j)|x]=\sum_{j=1}^{c}\lambda(a_i,w_j)P(w_j|x),\qquad i=1,\dots,k$$

- 设有某一决策规则$a(x)$，它对特征空间中所有可能的样本x采取决策所造成的期望损失是

$$R(a)=\int R(a(x)|x)p(x)\,\mathrm{d}x$$

$R(a)$称作**平均风险或期望风险**。

最小风险贝叶斯决策就是**最小化这一期望风险**，即：

$$\min_{a} R(a)$$

最小风险贝叶斯决策

最小风险贝叶斯决策

- 对于期望损失公式

$$R(a)=\int R(a(x)|x)p(x)\,\mathrm{d}x$$

其中$R(a(x)|x)$ 和 $p(x)$都是非负的，且$p(x)$是已知的，与决策准则无关。要使积分和最小，就是要对所有x都使$R(a(x)|x)$最小

- 因此，最小风险贝叶斯决策就是

$$\text{若}R(a_i|x)=\min_{j=1,\dots,k} R(a_j|x)\text{，则 } a=a_i$$

最小风险贝叶斯决策

- 对样本x，最小风险贝叶斯决策可以按照如下步骤计算：
  - 利用贝叶斯公式计算后验概率

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{\sum_{i=1}^{c} p(x \mid \omega_i)P(\omega_i)}, \qquad j = 1, \dots, c$$

  - 利用决策表，计算条件风险

$$R(\alpha_i \mid x) = \sum_{j=1}^{c} \lambda(\alpha_i \mid \omega_j)P(\omega_j \mid x), \qquad i = 1, \cdots, k$$

  - 决策：在各种决策中选择风险最小的决策

$$\alpha = \arg \min_{i=1,\cdots,k} R(\alpha_i \mid x)$$

最小风险贝叶斯决策

- 对于两类问题，最小风险贝叶斯决策规则可以总结为：
  - 假设损失函数为：

$$\lambda_{11}, \lambda_{12}, \lambda_{21}, \lambda_{22}$$

  - 最小风险贝叶斯决策为：

$$\lambda_{11}P(\omega_1 \mid x) + \lambda_{12}P(\omega_2 \mid x) \lesseqgtr \lambda_{21}P(\omega_1 \mid x) + \lambda_{22}P(\omega_2 \mid x), \text{则 } x \in \begin{cases} \omega_1 \\ \omega_2 \end{cases}$$

  - 左侧代表将其决策为第一类的损失，右侧代表将其决策为第二类的损失
  - 其中$\lambda_{12} = \lambda(a_1, w_2)$代表将属于第二类的样本决策为第一类时的损失

应用举例

- 假定进行癌细胞识别，已知：
  - 正常细胞$P(\omega_1)=0.9$；癌细胞$P(\omega_2)=0.1$
  - 若观察结果为x，从类条件概率密度曲线上查得，$P(x\mid\omega_1)=0.2$；$P(x\mid\omega_2)=0.4$
  - $\lambda(a_1,w_2)=6$；$\lambda(a_2,w_1)=1$；其余为0
- 如果根据最小错误率准则：

$$P(x\mid\omega_1)P(\omega_1)=0.2*0.9=0.18$$

$$P(x\mid\omega_2)P(\omega_2)=0.4*0.1=0.04$$

  - 决策为第一类

应用举例

- 假定进行癌细胞识别，已知：
  - 正常细胞$P(\omega_1) = 0.9$；癌细胞$P(\omega_2) = 0.1$
  - 若观察结果为x，从类条件概率密度曲线上查得，$P(x \mid \omega_1) = 0.2$；$P(x \mid \omega_2) = 0.4$
  - $\lambda(a_1, w_2) = 6$；$\lambda(a_2, w_1) = 1$；其余为0
- 如果根据最小风险准则：

$$R(\alpha_i \mid x) = \sum_{j=1}^{2} \lambda(\alpha_i \mid \omega_j) P(\omega_j \mid x)$$

有 $R(\alpha_1 \mid x) = \lambda(\alpha_1 \mid \omega_2) P(\omega_2 \mid x) = 1.092$

$R(\alpha_2 \mid x) = \lambda(\alpha_2 \mid \omega_1) P(\omega_1 \mid x) = 0.818$

  - 决策为第二类

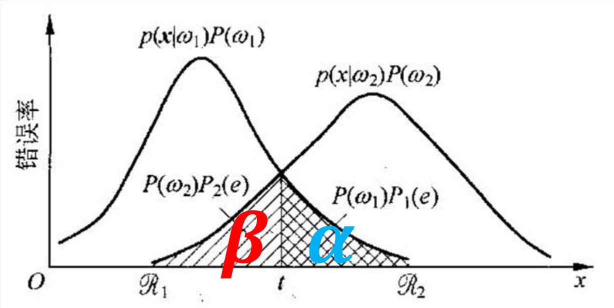
两类错误率

- 如果把$\mathbf{w_1}$看做阴性，$\mathbf{w_2}$看做阳性
  - 我们定义假阳性（FP）为：**真实情况为w_1但被检测为w_2的样本**，以此类推我们有真阳性（TP），假阴性（FN）与真阴性（TF）
  - 我们可以定义
    - 第一类错误率 $\alpha = \frac{FP}{FP+TN}$
    - 第二类错误率 $\beta = \frac{FN}{FN+TP}$
    - 灵敏度 $S_n = \frac{TP}{FN+TP} = 1 - \beta$
    - 特异度 $S_p = \frac{TN}{\mathrm{FP+TN}} = 1 - \alpha$
  - S_n代表了召回率，即在阳性样本中多少比例的样本可以被检测出

决策	真实状态	
	阴性	阳性
检测阳性	FP	TP
检测阴性	TN	FN

两类错误率

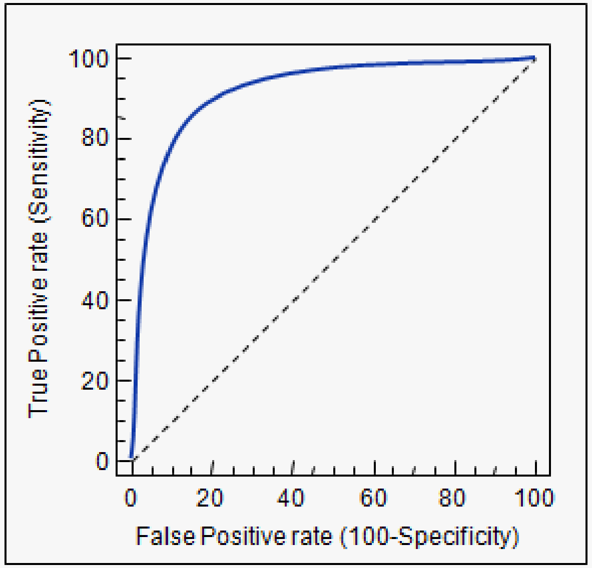
- 如果把w_1看做阴性，w_2看做阳性
  - 第一类错误率 $\alpha = \frac{FP}{FP+TN}$，代表了真实的阴性样本中被错判为阳性的概率，可以表示为右图中的R_2下阴影面积：$P_1(e) = \int_{R_2} p(x \mid \omega_1)dx$
  - 第二类错误率 $\beta = \frac{FN}{FN+TP}$，代表了真实的阳性样本中被错判为阴性的概率，可以表示为右图中的R_1下阴影面积：$P_2(e) = \int_{R_1} p(x \mid \omega_2)dx$



- 如果在某些情况下要在**保证某一类错误率为固定水平**的前提下，使另一类错误率**尽可能低**，该怎么决定决策边界呢？

ROC曲线

- 我们不难发现，随着阈值的连续变化，第一类错误率与第二类错误率也会随着连续变化
  - 将灵敏度（即真阳性率 $1-P_1(e)$ ）与假阳性率（即第二类错误率）当做纵、横坐标，我们可以画出两类错误率随着阈值变化的曲线，如下图
  - 图中每一个对应的曲线上的点，都对应了一个决策阈值，来使得在该阈值下对应的真假阳性率为某值
  - 对于一个决策，总是希望其真阳性率高，假阳性率低
  - **ROC曲线中越靠近左上角的点，说明方法性能越好**
  - 人们也用ROC曲线下的面积AUC来**定量衡量**方法性能



概率密度函数估计

- 为什么需要估计概率密度函数？
  - 在实际生活中，我们不能获得先验概率与类条件件概率，$P(\omega_i)$ 与$p(x \mid \omega_i)$
- 怎么办？
  - 一种很自然的想法是：先根据实际问题中的样本对先验概率与类条件件概率进行估计，即对$p(x \mid \omega_i)$与$P(\omega_i)$进行估计，记为 $\hat{p}(x \mid \omega_i)$与$\hat{P}(\omega_i)$
  - 然后运用估计得到的概率密度设计贝叶斯分类器

基于样本的两步贝叶斯决策

- 当样本数趋近于无穷大时，得到的分类器收敛于理论上的最优解：

$$\hat{p}(x \mid \omega_i) \xrightarrow{N\to\infty} p(x \mid \omega_i)$$

$$\hat{P}(\omega_i) \xrightarrow{N\to\infty} P(\omega_i)$$

概率密度函数的估计

- 如何利用样本集对概率密度函数进行估计？
- 估计概率密度函数的方法有两大类：
  - 参数方法
  - 非参数方法

- 注意，为了保证估计的有效性，我们需要满足两个重要前提：
  - 训练样本的分布能够代表样本的真实分布，即所谓的i.i.d条件
  - 有充足的训练样本

最大似然估计

- **参数估计：**已知概率密度函数的**形式**，只是其中几个参数未知，参数估计的目标是根据样本估计这些参数的值
- 在参数估计中，有几个简单概念：
  - 统计量：样本的某种函数，用来作为对某参数的估计
  - 参数空间：待估计参数的取值空间$\theta \in \Theta$
  - 点估计：统计量$\hat{\theta}(x)$的估计值（根据样本得到的具体值）
  - 区间估计：估计值的区间范围，反映点估计的误差范围
- 最大似然估计则是通过使估计统计量$\hat{\theta}(x)$与样本集的似然函数最大的方式求解参数估计问题

最大似然估计

□ 假设

- 参数θ是确定的未知量（不是随机变量）
- 各类样本集$D_i, i = 1, \ldots, c$中的样本都是从密度为$p(x \mid \omega_i)$的总体中独立抽取出来的（独立同分布，i. i. d. ）
- $p(x \mid \omega_i)$ 具有某种确定的函数形式，只是其参数θ未知
- 各类样本只包含本类分布的信息

- 其中，参数θ通常是向量，比如对一维正态分布$N(\mu_i, \sigma_1^2)$来说，$\theta_i = \begin{bmatrix} \mu_i \\ \sigma_i^2 \end{bmatrix}$，此时$p(x \mid \omega_i)$可写成$p(x \mid \omega_i, \theta_i)$ 或 $p(x \mid \theta_i)$

最大似然估计

- 基于这样的假设，我们可以将**每一类分别计算**，这里只考虑一类样本，记已知样本为：

$$D=\{x_1,x_2,\ldots,x_N\}$$

- 我们定义似然函数（likelihood function）如下：

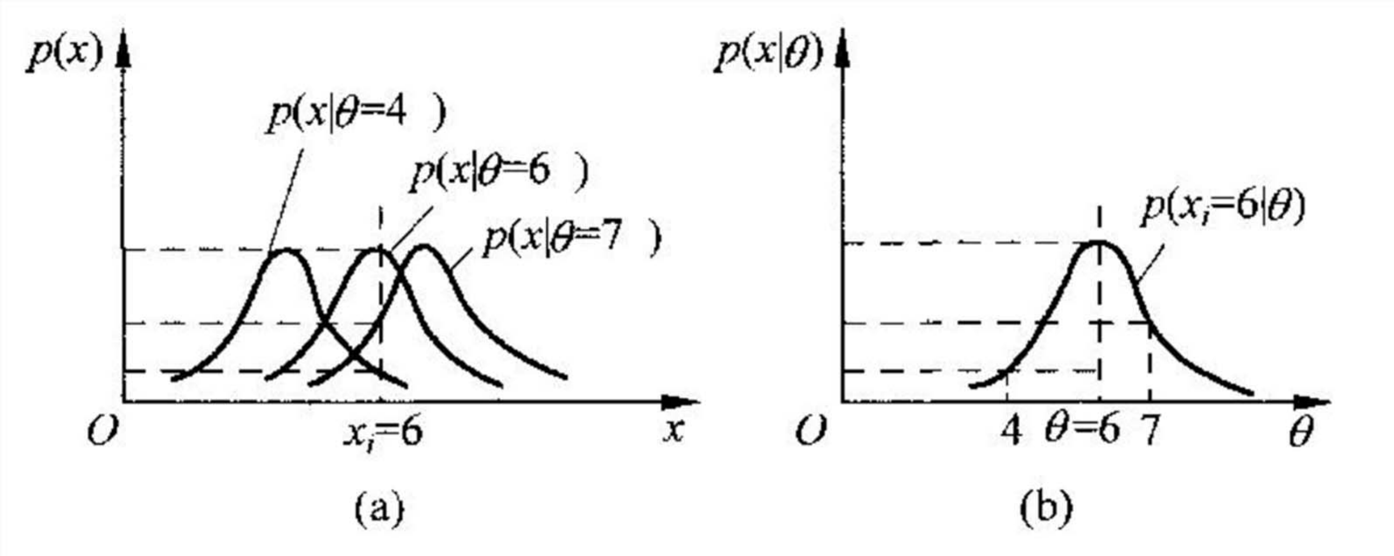
i.i.d

$$l(\theta)=p(D\mid\theta)=p(x_1,x_2,\ldots,x_N\mid\theta)=\prod_{i=1}^{N}p(x_i\mid\theta)$$

记为在参数θ下观测到样本集D的概率联合分布密度

最大似然估计

最大似然函数示意图

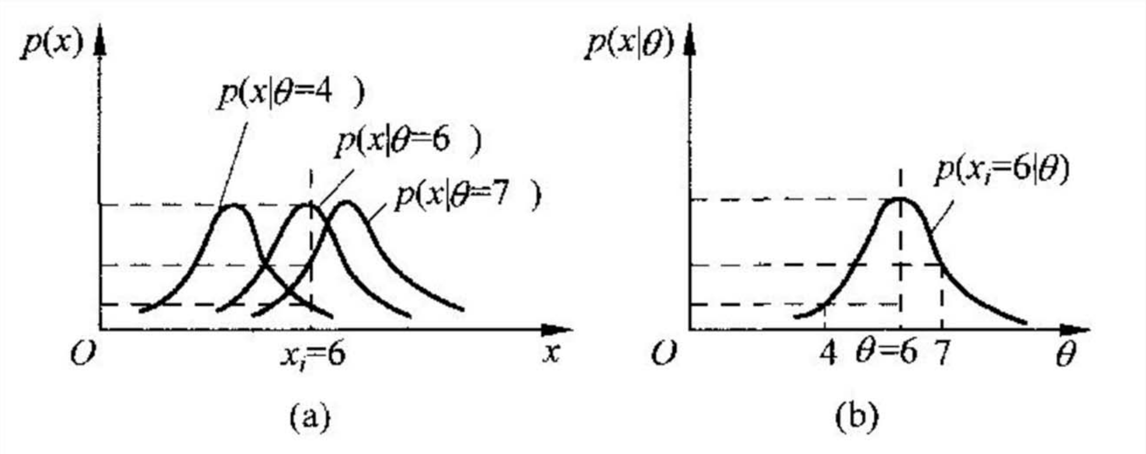


- 左图为不同参数下的概率分布，$x_1 = 6$ 在不同参数下对应的似然函数值不同，选择最大情况下的参数值作为估计值

最大似然估计

最大似然函数示意图

- 基本思想：如果在$\theta=\hat{\theta}$下$l(\theta)$最大，则$\hat{\theta}$应该是“最可能”的参数值
- 对于整个样本集$D=\{x_1,x_2,\dots,x_N\}$，我们可以定义对于D的函数，记作$\hat{\theta}=\arg\max_{\theta} p(D\mid\theta)$，称作最大似然估计量。为了计算方便，我们还可以定义对数似然函数：$H(\theta)=\ln l(\theta)=\sum_{i=1}^{N}\ln p(x_i\mid\theta)$



最大似然估计

最大似然估计的求解

- 若似然函数满足连续、可微的条件，则最大似然估计量就是方程

$$dl(\theta)/d\theta = 0 \text{ 或 } dH(\theta)/d\theta = 0$$

的解（必要条件）

- 若未知参数不止一个，即$\theta = [\theta_1, \theta_2, \dots, \theta_s]^T$，记梯度算子：

$$\nabla_\theta = \left[\frac{\partial}{\partial\theta_1}, \frac{\partial}{\partial\theta_2}, \cdots, \frac{\partial}{\partial\theta_s}\right]^T$$

- 则最大似然估计量的必要条件由s个方程组成

$$\nabla_\theta l(\theta) = 0$$

例子

- 最大似然估计求解正态分布：
  - 若以单变量正态分布为例：$\theta = [\theta_1, \theta_2]^T \quad \theta_1 = \mu \quad \theta_2 = \sigma^2$

$$p(x \mid \theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right]$$

- 最大似然估计求解均匀分布：

$$p(x \mid \theta) = \begin{cases} \dfrac{1}{\theta_2 - \theta_1} & \theta_1 < x < \theta_2 \\ 0 & \text{others} \end{cases}$$

最大似然估计

讨论：

- 如果$l(\theta)$或$H(\theta)$连续、可微，存在最大值且上述必要条件方程组有唯一解，则其解就是最大似然估计量（比如多元正态分布）。
- 如果必要条件有多解，则需从中求似然函数最大者
- 若不满足条件，则无一般性方法，用其它方法求最大（例如均匀分布）
- 对分布的前提假设要正确

- 思考：如果由均匀分布产生的数据用高斯分布做估计会如何？

贝叶斯估计

□ 贝叶斯估计

- 把概率密度函数的参数估计问题当成贝叶斯决策问题
- 思路与贝叶斯决策类似，只是离散的决策状态变成了连续的估计

□ 基本思想

- 把待估计参数θ_i看作具有先验分布$p(\theta_i)$的随机变量其取值与样本集有关，根据样本集D估计θ_i

$$p(\theta \mid D) = \frac{p(D \mid \theta)p(\theta)}{\int_{\Theta} p(D \mid \theta)p(\theta)d\theta} = \frac{p(D \mid \theta)p(\theta)}{p(D)}$$

贝叶斯估计

贝叶斯估计

- 损失函数：把θ估计为$\hat{\theta}$ 所造成的损失，记为 $\lambda(\hat{\theta},\theta)$
- 期望风险：$\mathbf{R} = \int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)\mathbf{p}(\mathbf{x},\theta)\mathbf{d}\theta\mathbf{dx}$

$$= \int_{E^d}\int_{\Theta}\lambda(\hat{\theta},\theta)\mathbf{p}(\theta \mid x)\mathbf{p}(x)\mathbf{d}\theta\mathbf{dx}$$

$$= \int_{E^d}\mathbf{R}(\hat{\theta} \mid x)\mathbf{p}(x)\mathbf{dx}$$

其中， $\mathbf{x} \in E^d$, $\theta \in \Theta$

- 其中条件风险：$\mathbf{R}(\hat{\theta} \mid \mathbf{x}) = \int_{\Theta}\lambda(\hat{\theta},\theta)\mathbf{p}(\theta \mid \mathbf{x})\mathbf{d}\theta \ \mathbf{x} \in E^d$
- 最小化期望风险 ➞ 最小化条件风险（对所有可能的x）
- 有限样本集D下，最小化经验风险： $\mathbf{R}(\hat{\theta} \mid \mathbf{D}) = \int_{\Theta}\lambda(\hat{\theta},\theta)\mathbf{p}(\theta \mid \mathbf{D})\mathbf{d}\theta$

贝叶斯估计

贝叶斯估计量

- 在样本集下，使条件风险最小的估计量
- 常用的损失函数：平方损失函数$\lambda(\hat{\theta}, \theta) = (\theta - \hat{\theta})^2$
- 可以证明，采用平方误差损失函数，则θ的贝叶斯估计量$\hat{\theta}$是在给定x时θ的条件期望，即$\hat{\theta} = E[\theta \mid x] = \int_{\Theta} \theta p(\theta \mid x) d\theta$
- 同理可得，在给定样本集D下，θ的贝叶斯估计是：

$$\hat{\theta} = E[\theta \mid D] = \int_{\Theta} \theta p(\theta \mid D) d\theta$$

贝叶斯估计

总结

- 求解贝叶斯估计的方法（平方误差损失）
  - 确定θ的先验分布： $p(\theta)$
  - 求样本集的联合分布： $p(\mathrm{D} \mid \theta)=\prod_{i-1}^{N} p(x_i \mid \theta)$
  - 求θ的后验概率分布： $p(\theta \mid D)=\frac{p(D|\theta)p(\theta)}{\int_{\Theta} p(D|\theta)p(\theta)d\theta}$
  - 求θ的贝叶斯估计量： $\hat{\theta}=\int_{\Theta} \theta p(\theta \mid D) d\theta$

贝叶斯估计

贝叶斯决策与贝叶斯估计的比较

贝叶斯决策	贝叶斯估计
样本$\boldsymbol{x}$	样本集合 $D=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N\}$
类的先验概率$p(\omega_i)$	参数先验分布 $p(\theta_i)$
真实状态 ω_i	真实参数 θ_i
决策α_i	估计 $\hat{\theta}_i$
类别状态：离散	分布参数：连续
损失函数表（决策表）	损失函数

参数估计

最大似然估计 vs 贝叶斯估计

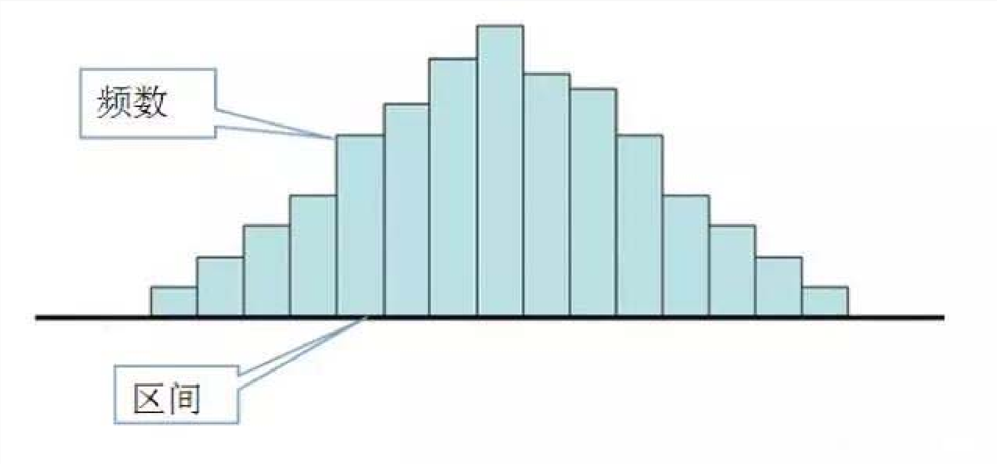
- 最大似然估计简单直观
- 当训练样本数无穷多的时候, 最大似然估计和贝叶斯估计的结果是一样的
- 贝叶斯估计由于使用了先验概率，利用了更多的信息。如果这些信息是可靠的, 那么有理由认为贝叶斯估计比最大似然估计的结果更准确

有时候先验概率很难设计，怎么办?

- 在没有特别先验知识的时候，取先验概率是这个区域中的均匀分布（无信息先验）
- 这种情况下最大似然估计结果和贝叶斯估计结果相似

非参数估计

- 直方图方法：非参数概率密度估计的最简单方法
  - 把x的每个分量分成k个等间隔小窗若 $x \in R^d$，则形成k^d 个小舱
  - 统计落入各个小舱内的样本数q_i
  - 相应小舱的概率密度为 q_i/NV（N：样本总数，V：小舱体积）



无需考虑用参数方法表达概率密度

非参数估计

非参数估计的基本原理

- 问题：已知样本集$D=\{x_1,\ldots,x_N\}$，其中样本都是从服从$p(x)$的分布中独立抽取，求$\hat{p}(x)$来近似$p(x)$
- 考虑随机变量x落入区域$\Re$的概率　$P_{\Re}=\int_{\Re} p(x)dx$
- D中有k个样本落入区域$\Re$的概率　$P_k=C_N^k P_{\Re}^k(1-P_{\Re})^{N-k}$
- k的期望值　$E[k]=NP_{\Re}$
- 因此取 $P_{\Re}$ 的估计　$\hat{P}_{\Re}=\frac{k}{N}$　（k:实际落入$\Re$中的样本数）
- 设$p(x)$，且$\Re$足够小，$\Re$的体积为V, 则有$P_{\Re}=\int_{\Re} p(x)dx=p(x)V$，$x\in\Re$
- 因此$\hat{p}(x)=\frac{k}{NV}$

非参数估计

非参数估计的基本原理

- 如何选择V?
  - 过大，估计偏差很大
  - 过小，某些区域中缺少样本

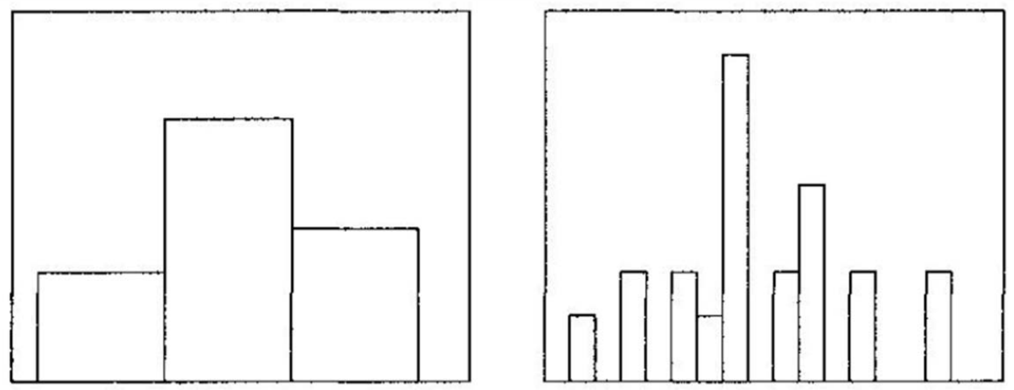


图 3-3 小窗宽度对直方图估计的影响示意

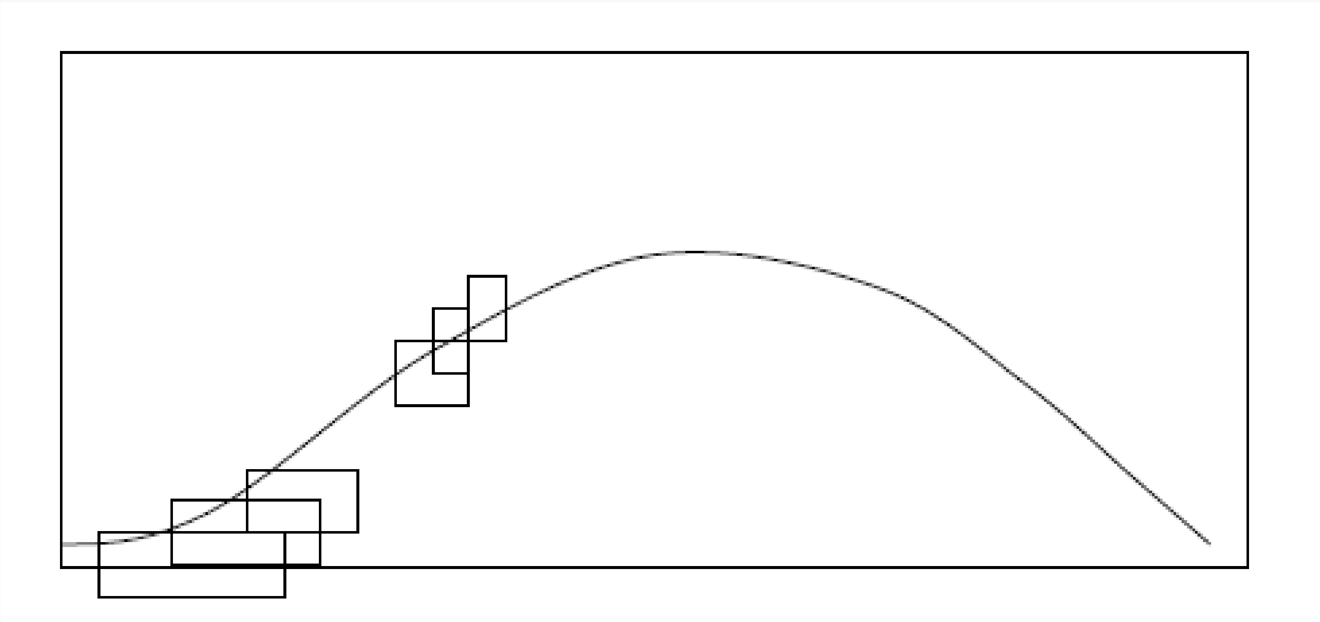
- 选择V：Parzen 窗法
- 选择k，V为正好包含x的k_N个近邻：k_N近邻估计

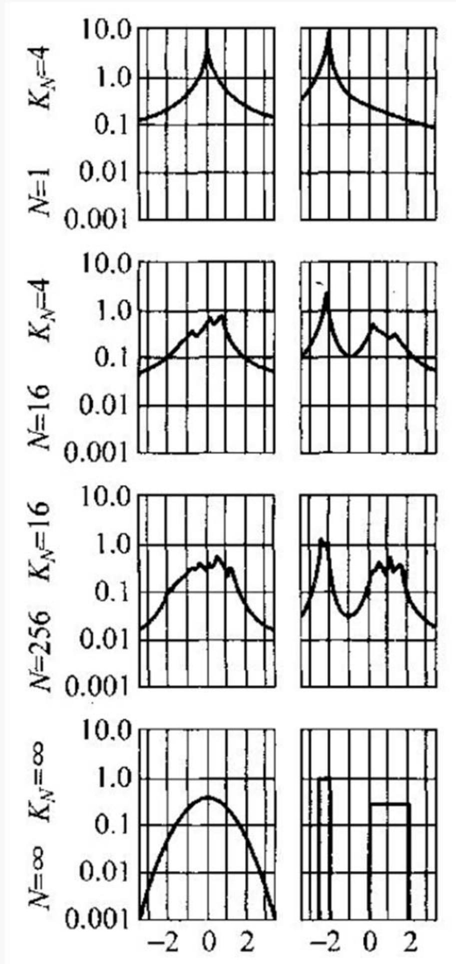
k_N近邻估计

k_n近邻估计

$$\hat{p}_n(x) = \frac{k_n/N}{V_n}$$

通过确认每一个小区域内样本数k_N来确定区域大小

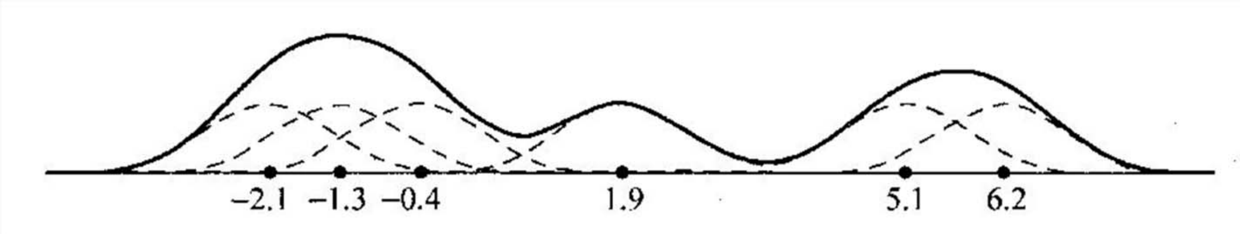




Parzen窗法

Parzen窗法

$$\hat{p}(x) = \frac{1}{N}\sum_{i=1}^{N} k(x, x_i)$$



窗函数(核函数)：

$k(x, x_i)$:反映x_i对 $p(x)$ 的贡献，实现小区域选择

满足条件：

$$k(x, x_i) \geq 0$$
$$\int k(x, x_i)dx = 1$$

常用窗函数

超立方体窗（方窗）

$$k(x,x_i)=\begin{cases}\dfrac{1}{h^d} & \text{if} \quad \left|x^i-x_i^j\right|\le h/2, j=1,2,\cdots,d\\ 0 & \text{otherwise}\end{cases}$$

h为超立方体棱长， $V=h^d$

正态窗（高斯窗）

$$k(x,x_i)=\frac{1}{\sqrt{(2\pi)^d\rho^{2d}|Q|}}\exp\left\{-\frac{1}{2}\frac{(x-x_i)^T Q^{-1}(x-x_i)}{\rho^2}\right\}\ \left(\Sigma=\rho^2 Q\right)$$

超球窗

$$k(x,x_i)=\begin{cases}V^{-1} & \text{if} \quad \|x-x_i\|\le\rho\\ 0 & \text{otherwise}\end{cases}$$

V为超球体积， ρ为半径

§6.3 朴素贝叶斯分类器

一、基本概念

二、应用举例

三、拉普拉斯修正

朴素贝叶斯分类器

- 贝叶斯公式估计后验概率的困难

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

  - 类条件概率$P(x|c)$：所有属性的联合概率
  - 难以从有限的训练样本直接估计得到

- 朴素贝叶斯分类器

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

  - 属性条件独立性假设
  - 对于已知类别，假设所有属性互相独立

朴素贝叶斯分类器

朴素贝叶斯分类器

- 后验概率估计

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

- 贝叶斯判定准则

$$h^*(\boldsymbol{x}) = \underset{c\in\boldsymbol{y}}{\operatorname{argmax}}\, P(c|\boldsymbol{x}) = \underset{c\in\boldsymbol{y}}{\operatorname{argmax}}\, P(c)\prod_{i=1}^{d} P(x_i|c)$$

- 朴素贝叶斯分类器的训练过程
  - 基于训练集D估计类先验概率$P(c)$
  - 为每个属性估计条件概率$P(x_i|c)$

朴素贝叶斯分类器

朴素贝叶斯分类器

- 基于训练集$\boldsymbol{D}$估计类先验概率$\boldsymbol{P(c)}$

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

- 为每个属性估计条件概率$\boldsymbol{P(x_i|c)}$
  - 离散属性

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

  - 连续属性：假定$\boldsymbol{p(x_i|c) \sim N(\mu_{c,i}, \sigma_{c,i}^2)}$

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

应用举例

- 应用西瓜数据集3.0训练一个朴素贝叶斯分类器
  - 估计类先验概率$P(c)$

$$P(\text{好瓜}=\text{是})=\frac{8}{17}\approx 0.471\,,$$
$$P(\text{好瓜}=\text{否})=\frac{9}{17}\approx 0.529\,.$$

  - 每个属性条件概率$\boldsymbol{P(x_i|c)}$

$$P_{\text{青绿}|\text{是}}=P(\text{色泽}=\text{青绿}\mid\text{好瓜}=\text{是})=\frac{3}{8}=0.375$$
$$P_{\text{青绿}|\text{否}}=P(\text{色泽}=\text{青绿}\mid\text{好瓜}=\text{否})=\frac{3}{9}\approx 0.333$$

表 4.3 西瓜数据集 3.0

编号	色泽	根蒂	敲声	纹理	脐部	触感	密度	含糖率	好瓜
1	青绿	蜷缩	浊响	清晰	凹陷	硬滑	0.697	0.460	是
2	乌黑	蜷缩	沉闷	清晰	凹陷	硬滑	0.774	0.376	是
3	乌黑	蜷缩	浊响	清晰	凹陷	硬滑	0.634	0.264	是
4	青绿	蜷缩	沉闷	清晰	凹陷	硬滑	0.608	0.318	是
5	浅白	蜷缩	浊响	清晰	凹陷	硬滑	0.556	0.215	是
6	青绿	稍蜷	浊响	清晰	稍凹	软粘	0.403	0.237	是
7	乌黑	稍蜷	浊响	稍糊	稍凹	软粘	0.481	0.149	是
8	乌黑	稍蜷	浊响	清晰	稍凹	硬滑	0.437	0.211	是
9	乌黑	稍蜷	沉闷	稍糊	稍凹	硬滑	0.666	0.091	否
10	青绿	硬挺	清脆	清晰	平坦	软粘	0.243	0.267	否
11	浅白	硬挺	清脆	模糊	平坦	硬滑	0.245	0.057	否
12	浅白	蜷缩	浊响	模糊	平坦	软粘	0.343	0.099	否
13	青绿	稍蜷	浊响	稍糊	凹陷	硬滑	0.639	0.161	否
14	浅白	稍蜷	沉闷	稍糊	凹陷	硬滑	0.657	0.198	否
15	乌黑	稍蜷	浊响	清晰	稍凹	软粘	0.360	0.370	否
16	浅白	蜷缩	浊响	模糊	平坦	硬滑	0.593	0.042	否
17	青绿	蜷缩	沉闷	稍糊	稍凹	硬滑	0.719	0.103	否

应用举例

- 估计每个属性条件概率$P(x_i|c)$

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

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

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

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

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

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

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

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

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

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

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

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

$$p_{\text{密度: 0.697}|\text{是}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi} \cdot 0.129} \exp\left(-\frac{(0.697 - 0.574)^2}{2 \cdot 0.129^2}\right) \approx 1.959$$

$$p_{\text{密度: 0.697}|\text{否}} = p(\text{密度} = 0.697 \mid \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi} \cdot 0.195} \exp\left(-\frac{(0.697 - 0.496)^2}{2 \cdot 0.195^2}\right) \approx 1.203$$

$$p_{\text{含糖: 0.460}|\text{是}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{是}) = \frac{1}{\sqrt{2\pi} \cdot 0.101} \exp\left(-\frac{(0.460 - 0.279)^2}{2 \cdot 0.101^2}\right) \approx 0.788$$

$$p_{\text{含糖: 0.460}|\text{否}} = p(\text{含糖率} = 0.460 \mid \text{好瓜} = \text{否}) = \frac{1}{\sqrt{2\pi} \cdot 0.108} \exp\left(-\frac{(0.460 - 0.154)^2}{2 \cdot 0.108^2}\right) \approx 0.066$$

应用举例

对测试样例1进行分类测试

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	?

$$h^*(\boldsymbol{x}) = \underset{c\in\boldsymbol{y}}{\operatorname{argmax}}\, P(c)\prod_{i=1}^{d} P(x_i|c)$$

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

好瓜！

清華大学

朴素贝叶斯分类器

拉普拉斯修正

- 特殊情况：某个属性值在训练集中没有与某个类同时出现过

$$P_{清脆|是} = P(敲声=清脆 \mid 好瓜=是) = \frac{0}{8} = 0$$

- 拉普拉斯修正

$$\hat{P}(c) = \frac{|D_c| + 1}{|D| + N} \qquad \hat{P}(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$$

$$\hat{P}_{清脆|是} = \hat{P}(敲声=清脆 \mid 好瓜=是) = \frac{0+1}{8+3} \approx 0.091$$

- 意义：避免了因为训练集样本不充分导致概率估值为零的问题
- 先验的影响：训练集变大时，修正过程引入的先验(N_i, N)影响可忽略，估值趋向于实际概率值