第13章 贝叶斯决策

宋有富人，天雨墙坏。其子曰：“不筑，必将有盗”。其邻人之父亦云。暮而果大亡其财。其家甚智其子，而疑邻人之父。

——《韩非子·说难》

北京交通大学《机器学习》课程组

引入“贝叶斯决策论”

宋有富人，天雨墙坏。其子曰：“不筑，必将有盗”。其邻人之父亦云。暮而果大亡其财。其家甚智其子，而疑邻人之父。

——《韩非子·说难》

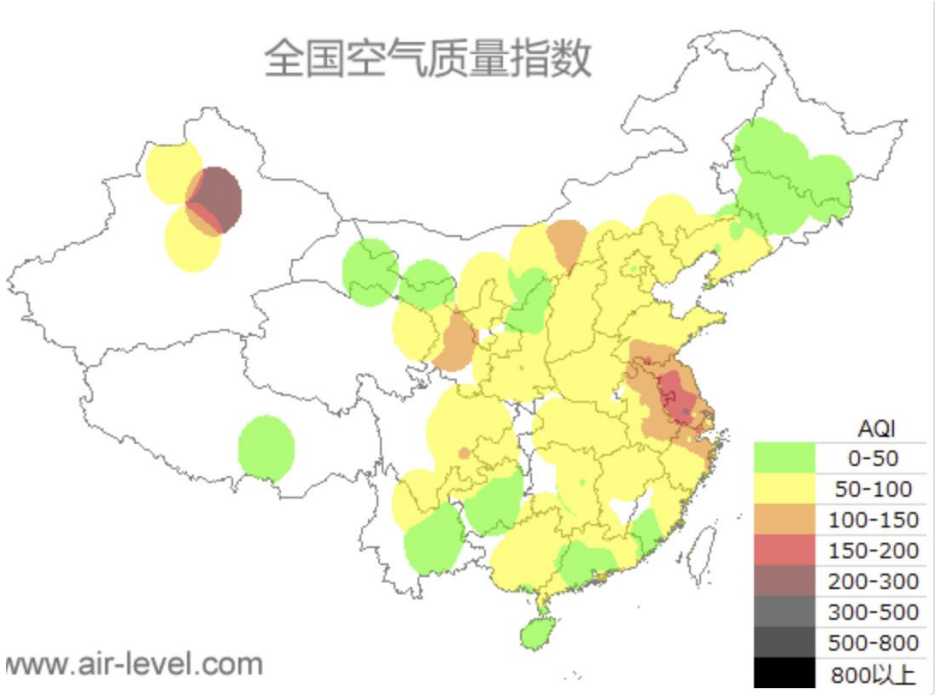
这个故事告诉我们，同样的事实，不同的先验估计，得到的结论会大不相同。

因此，事实非常重要，但是先验也非常重要。

引入“贝叶斯决策论”

假设类表示独立于样本的抽样分布：

- 希望学到事物的本质特性—类认知表示
- 与训练样本的抽样分布应该无关



对于某些学习任务，希望学习训练集的样本分布：

- 未必知道/不需要知道其本质
- 只需学习事件发生的概率分布（不确定性学习问题）

多类密度估计问题：
样本归类属于不确定性决策问题

目录

目录

13.1 最小错误率贝叶斯决策

Bayes决策理论是

用概率统计方法研究决策问题

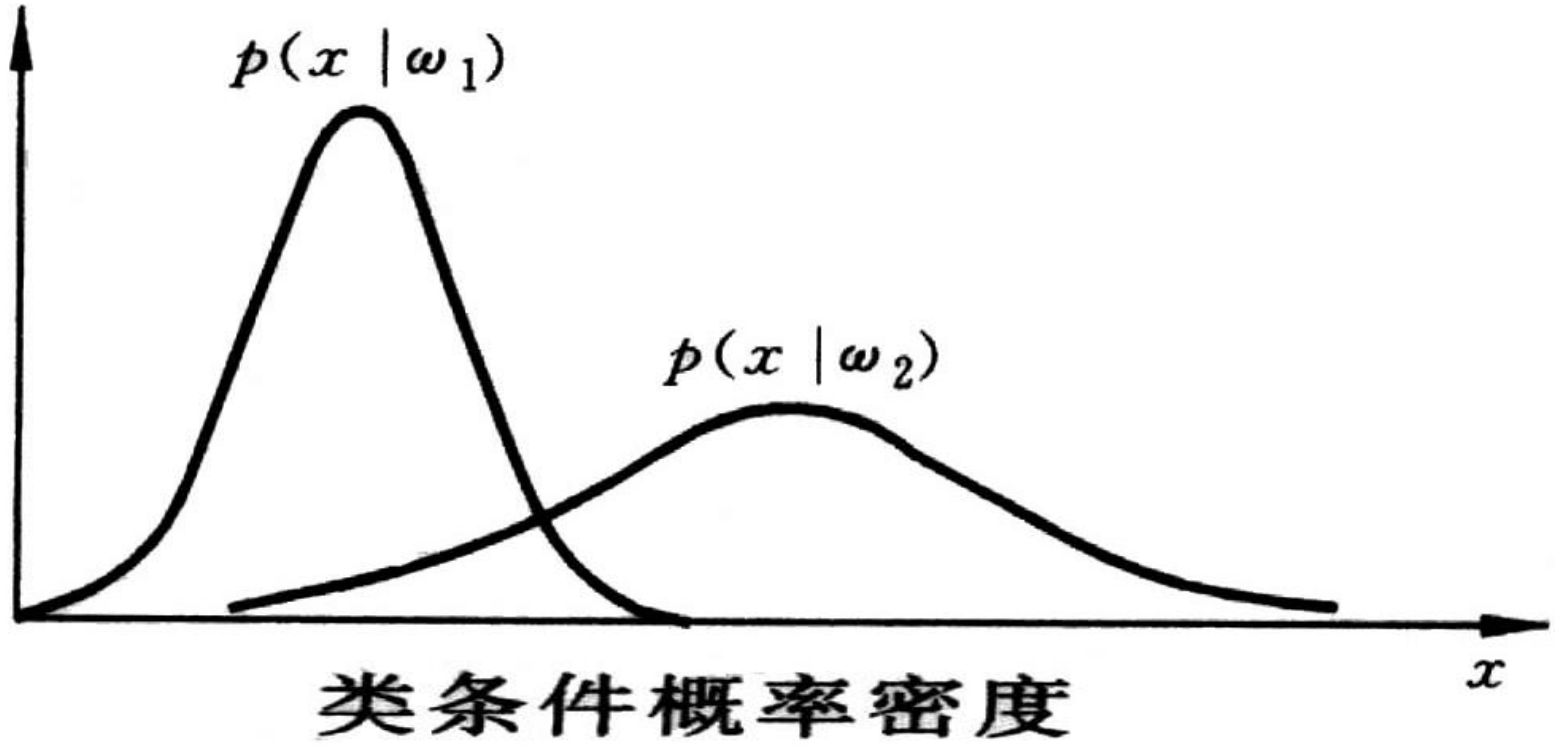
三个重要的概率：先验概率

由样本的先验知识得到先验概率，可从训练集样本中估算出来。

例如，两类10个训练样本，属于w_1为2个，属于w_2为8个，则先验概率$P(w_1) = 0.2$，$P(w_2) = 0.8$。

三个重要的概率：类条件概率密度函数

类条件概率密度函数：样本x在ω_i类条件下，出现的概率密度分布函数。
也称 $p(x|\omega_i)$ 为ω_i 关于x 的似然函数。

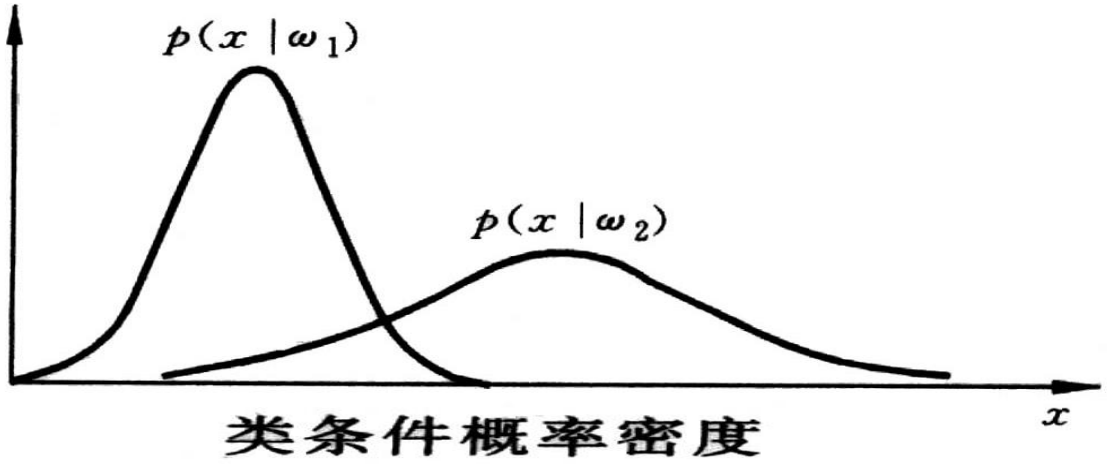


类条件概率密度

三个重要的概率：后验概率

- 后验概率：某个样本 x, 属于ω_i 类的概率, $i=1,\cdots,c$ 。

- 根据样本x的先验概率和类条件概率密度函数$p(x|\omega_i)$ 用Bayes公式重新修正得到后验概率。



类条件概率密度

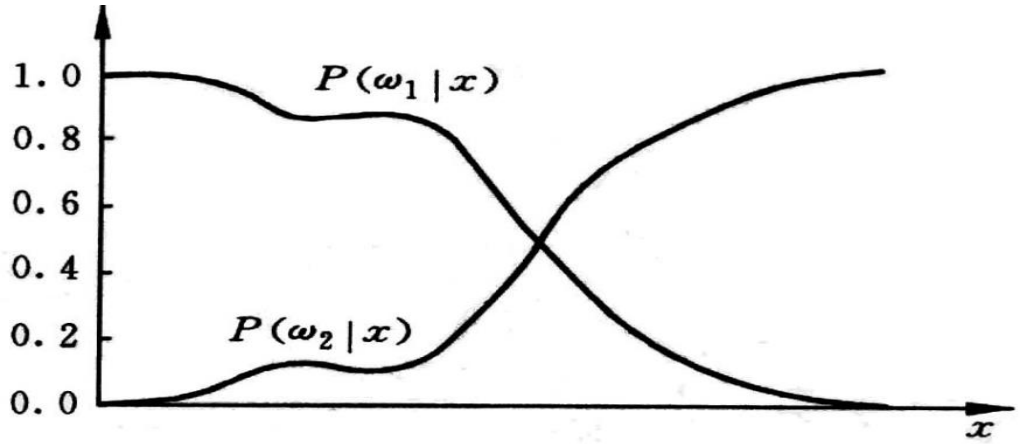


图 2.2 后验概率

贝叶斯公式及最小错误率决策

已知：先验概率 $P(\omega_i)$，类条件概率密度函数 $p(x|\omega_i)$

则 后验概率为 $P(\omega_i|x)=\dfrac{p(x|\omega_i)P(\omega_i)}{p(x)}$

其中，全概率密度 $p(x)=\sum_{i=1}^{c}p(x|\omega_i)P(\omega_i)$

两类$(c=2)$情况下

如果 $P(\omega_1|x)>P(\omega_2|x)$，则$x$属于$\omega_1$类

如果 $P(\omega_1|x)<P(\omega_2|x)$，则$x$属于$\omega_2$类

例子

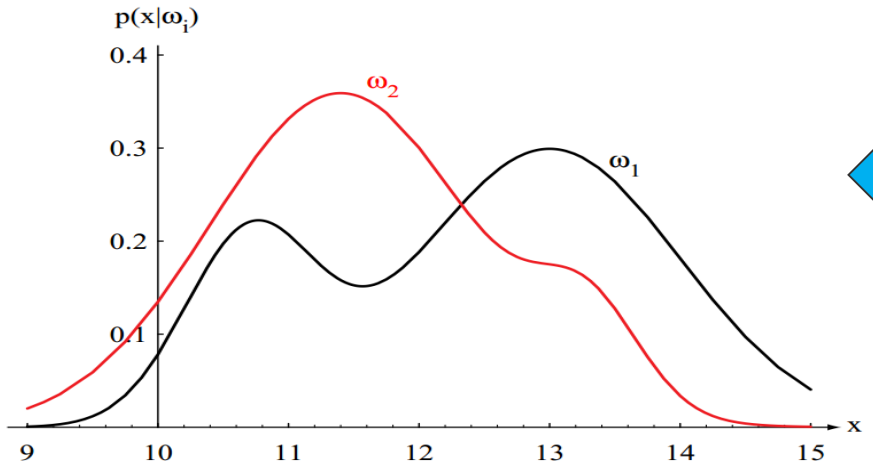
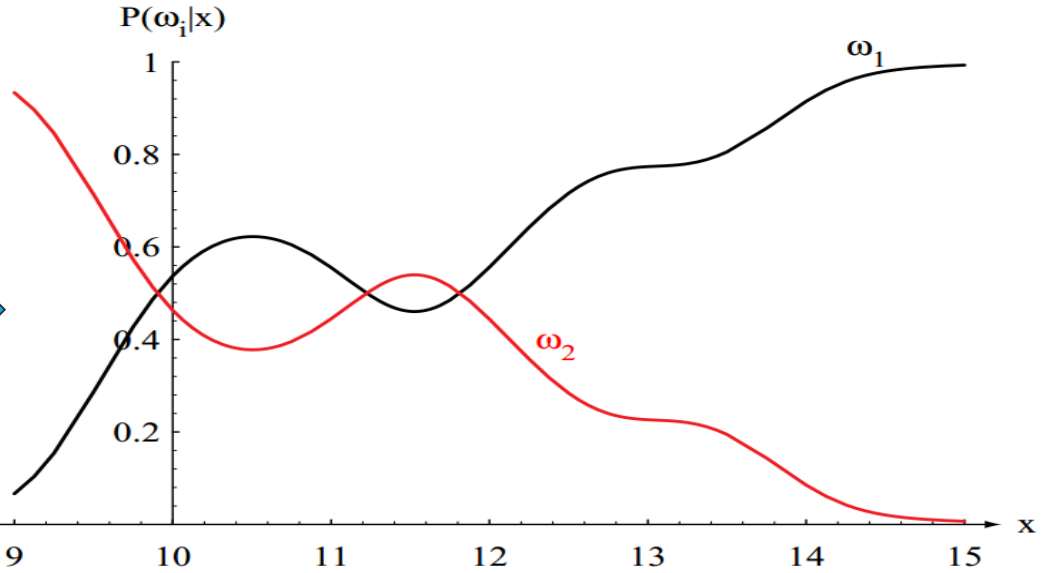


图 2-2　在先验概率 $P(\omega_1)=2/3$，$P(\omega_2)=1/3$ 及图 2-1 给出的类条件概率密度的条件下的后验概率图。此情况下，假定一个模式具有特征值 $x=14$，那么它属于 ω_2 类的概率约为 0.08，属于 ω_1 的概率约为 0.92。在每个 x 处的后验概率之和为 1.0

- 类条件概率密度函数$p(x|\omega_i)$图，显示模式处于ω_i类别时观测某个特定值x的概率密度
- x代表鱼的长度



re 2.2: I
for the c
ı that a
ıtegory
to 1.0.

$$P(\omega_j|x)=\frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

nd $P(\omega_2)=$
in this case,
ability it is
e posteriors

最小错误率贝叶斯决策

■两类情况下贝叶斯分类的三种等价形式

决策：$x \in \omega_1$，否则$x \in \omega_2$

Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; otherwise decide ω_2

Decide ω_1 if $p(x|\omega_1)P(\omega_1) > p(x|\omega_2)P(\omega_2)$; otherwise decide ω_2.

$$l(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{P(\omega_2)}{P(\omega_1)}$$

统计学中 $l(x)$称为似然比，$\frac{P(\omega_2)}{P(\omega_1)}$ 称为似然比阈值

最小错误率贝叶斯决策

- 多类情况下贝叶斯分类

$$P(\omega_i \mid x) = \max_{j=1,\cdots,c} P(\omega_j \mid x), \text{ 则} x \in \omega_i \qquad i, j = 1, 2, \cdots, c$$

问题描述

给定$X_{p\times N}$，$U_{1\times N}$，

要求：训练模型使其能够对未知样本$x_{p\times 1}$进行类别判断

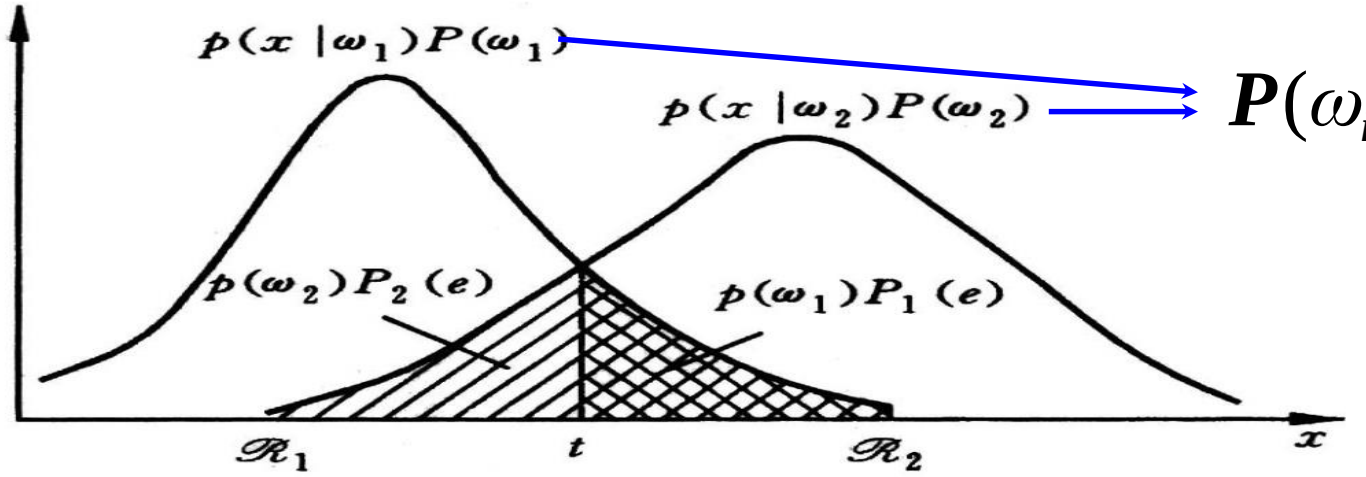
分类规则：将未知样本分给后验概率最大的类

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)}$$

最小错误率贝叶斯决策

为什么这样分类的结果平均错误率最小？

$$P(e)=\int_{-\infty}^{\infty}P(e,x)dx=\int_{-\infty}^{\infty}P(e\mid x)p(x)dx$$



$P(\omega_i\mid x)p(x)$

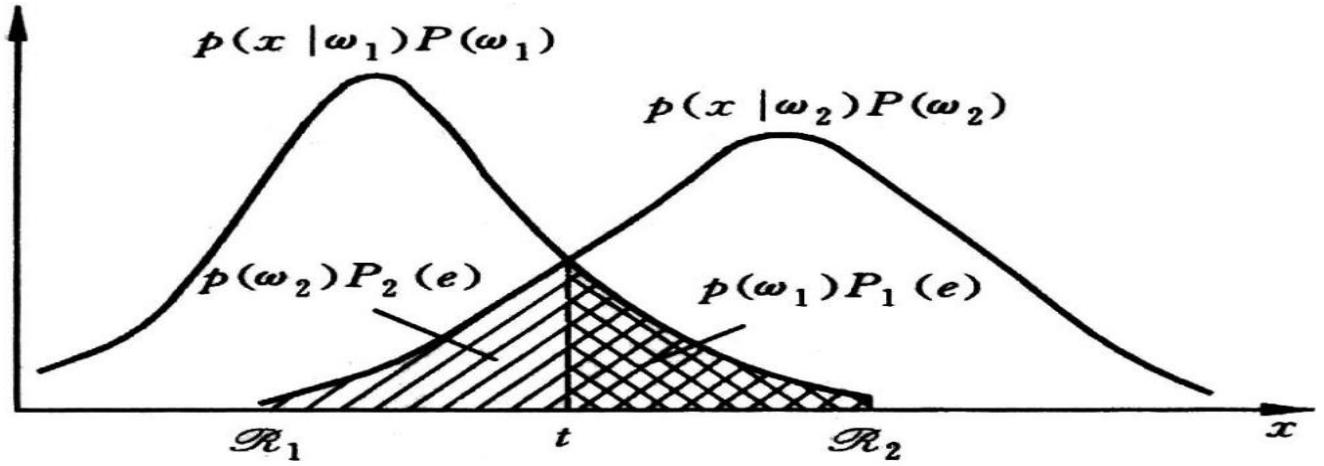
对应贝叶斯决策

判断错误的区域为阴影包围的面积。

$$P(e\mid x)=\begin{cases}P(\omega_1\mid x), & \text{当}P(\omega_2\mid x)>P(\omega_1\mid x)\\ P(\omega_2\mid x), & \text{当}P(\omega_1\mid x)>P(\omega_2\mid x)\end{cases}$$

最小错误率贝叶斯决策

- t将x轴分成两个区域R_1和R_2， R_1为$(-\infty, t)$； R_2为(t,∞)



- 平均错误率$P(e)$

$$P(e) = \int_{R_1} P(\omega_2|x)p(x)dx + \int_{R_2} P(\omega_1|x)p(x)dx$$

$$P(e) = P(\omega_2)\int_{R_1} p(x|\omega_2)dx + P(\omega_1)\int_{R_2} p(x|\omega_1)dx$$

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

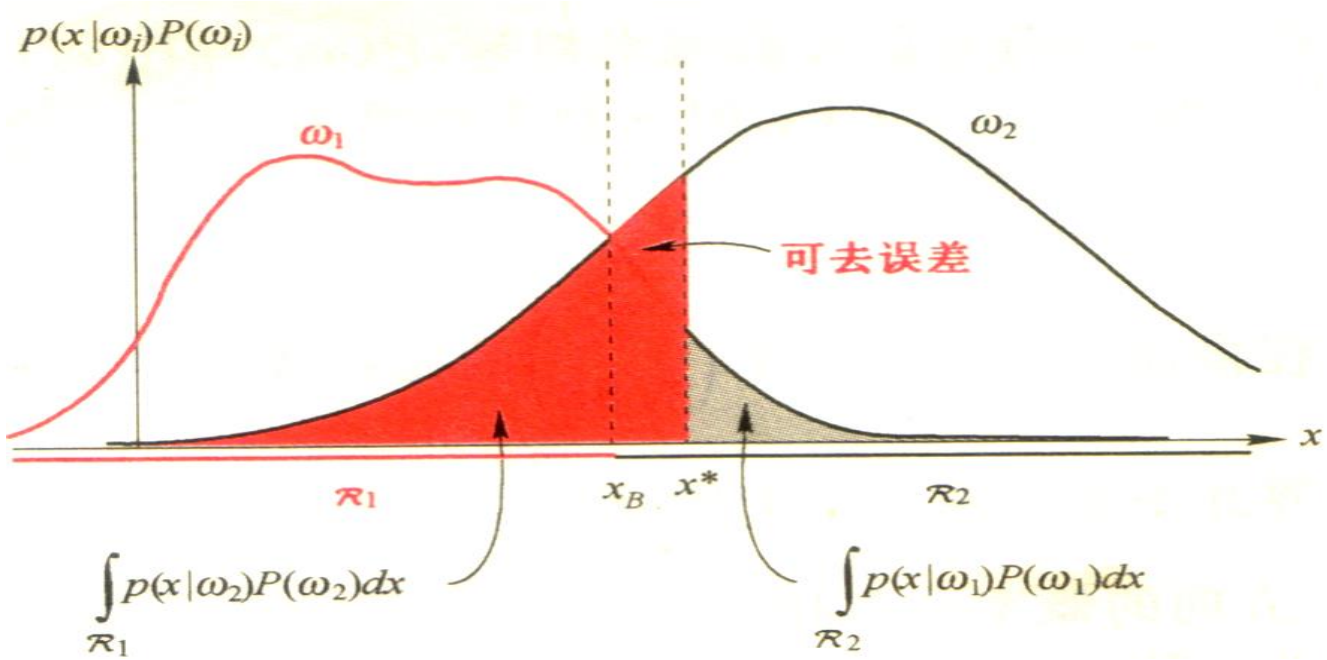
贝叶斯公式

最小错误率贝叶斯决策

- 决策规则实际上对每个x都使 $p(e|x)$取小者

$$P(e)=\int_{-\infty}^{\infty}P(e,x)dx=\int_{-\infty}^{\infty}P(e|x)p(x)dx$$

- 移动决策面 t 都会使错误区域增大，因此贝叶斯决策的平均错误率最小。



最小错误率贝叶斯决策

应用实例： 细胞识别

已知：正常类$P(w_1)=0.9$； 异常类$P(w_2)=0.1$待识别细胞x，从类条件概率密度曲线上查得$p(x|w_1)=0.2$； $p(x|w_2)=0.4$

解：利用Bayes公式分别计算ω_1和ω_2的后验概率

$$P(\omega_1 \mid x)=\frac{p(x \mid \omega_1)P(\omega_1)}{\sum_{j=1}^{2} p(x \mid \omega_j)P(\omega_j)}=0.818$$

$$P(\omega_2 \mid x)=1-P(\omega_1 \mid x)=0.182$$

因此 $P(\omega_1 \mid x)=0.818>P(\omega_2 \mid x)=0.182 \qquad x \in \omega_1$

这种规则后验概率起决定作用。这里**没有考虑错误分类带来的损失**。

目录

13.1 贝叶斯分类器

- **输入类表示：**
  - **输入类认知表示**

    $\underline{X_i} = p_i(x) = p(x|X_i)$

  - **输入类相似映射函数**

    $\mathrm{Sim}_X(x, \underline{X_i}) = a_i p_i(x)$

$p_i(x)$**表示第**i**输入类中**x**发生的概率**$p_i(x) = p(x|X_i)$

a_i**表示第**i**类发生的概率**$a_i = P(X_i)$

$X = \{x_1, x_2, \cdots, x_N\}$

类标集$U = \{u_1, u_2, \cdots, u_N\}$, $u_k = [u_{1k}, u_{2k}, \cdots, u_{ck}]^T$,

$$\begin{cases} u_{ik} = 1, x_k \in X_i \\ u_{ik} = 0, x_k \notin X_i \end{cases} \qquad \sum_{i=1}^{c} u_{ik} = 1$$

- **输出类表示：**
  - **输出类认知表示**

    $\underline{Y_i} = \widehat{p_i(y)} = p(y|Y_i)$

  - **输出类相似映射函数**

    $\mathrm{Sim}_X(y, \underline{Y_i}) = \hat{a}_i \widehat{p_i(y)}$

$\widehat{p_i(y)}$**表示第**i**输出类中**y**发生的概率**$\widehat{p_i(y)} = p(y|Y_i)$

$\hat{a}_i$**表示第**i**输出类发生的概率**$\hat{a}_i = P(Y_i)$

$Y = \{y_1, y_2, \cdots, y_N\}$

类标集$V = \{v_1, v_2, \cdots, v_N\}^T$**，隶属度**$v_k = [v_{1k}, v_{2k}, \cdots, v_{ck}]^T$**，**

$v_{ik} = P(Y_i|y_k)$　　$\sum_{i=1}^{c} v_{ik} = 1$

■ 归类表示：

	输入端	输出端
四元组	$(\boldsymbol{X}, \boldsymbol{U}, \underline{X}, Sim_X)$	$(Y, V, \boldsymbol{\underline{Y}}, \boldsymbol{Sim_Y})$
特性表示	$[x_{\tau k}]_{p\times N}$	$[y_{\tau k}]_{d\times N}$
类标集	$U = [u_{ik}]_{c\times N}$（硬划分）	$\mathrm{V} = [v_{ik}]_{c\times N}$ $v_{ik} = P(Y_i\|y_k)$且$\sum_{i=1}^{c} v_{ik} = 1$（软划分） $i = argmax_i P(Y_i\|y_k)$，对象划分为第i类
类认知表示	$\underline{X_i} = p_i(x) = p(x\|X_i)$ 是r.v. x的密度函数 $p_i(x) = p(x\|X_i)$第i输入类中x发生的概率	$\underline{Y_i} = \widehat{p_i(y)} = p(y\|Y_i)$ 是r.v. y的密度函数 $\widehat{p_i(y)} = p(y\|Y_i)$第$i$输出类中y发生的概率
相似性映射	$Sim_X\left(x, \underline{X_i}\right) = a_i p_i(x)$ $= P(X_i)p(x\|X_i)$ $a_i = P(X_i)$第i输入类发生的概率；	$Sim_Y\left(\mathrm{y}, \underline{Y_i}\right) = \hat{a}_i \widehat{p_i(y)}$ $= P(Y_i)p(y\|Y_i)$ $\hat{a}_i = P(Y_i)$第i输出类发生的概率；

贝叶斯分类器

最一般的贝叶斯分类器

如果 $i = \arg\max_j P(Y_j \mid y)$，则判断对象*o*属于*i*类

贝叶斯定理

$$P(Y_i \mid y_k) = \frac{p(y_k \mid Y_i)P(Y_i)}{p(y_k)}$$

样本相对于类标记的“类条件概率”,或称“似然”

先验概率

“证据”（evidence）因子，与类标记无关

贝叶斯分类器

■ 归类等价公理：

- 输出隶属度

  $v_{ik} = P(Y_i|y_k)$

- 输出类相似映射

  $Sim_Y\left(\mathrm{y}, \underline{Y_i}\right) = \hat{a}_i \widehat{p_i(y)} = P(Y_i)p(y|Y_i)$

$$\tilde{y} = \arg\max_j \mathrm{Sim}_Y\left(\mathrm{y}, \underline{Y_j}\right) = \mathrm{argmax}_j\, P(Y_j)p(y|Y_j)$$

$$= \arg\max_j \frac{P(Y_j)p(y|Y_j)}{p(y)} = \mathrm{argmax}_j P(Y_j|y)$$

$$= \arg\max_j\ v_{jk} = \vec{y}$$

所以，归类等价公理对于贝叶斯分类器成立。

贝叶斯分类器

假设$X = Y$，则有$y = x$，$\forall k, y_k = x_k$，$\forall i, \widehat{p_i(y)} = \widehat{p_i(x)}$，
$\widehat{a_i}(= P(Y_i))$是对于$a_i(= P(X_i))$的估计，$\widehat{p_i(x)}$是对$p_i(x)$的估计。

需指出$\forall i, X_i = Y_i$一般不成立，继续使用Y_i来代表第i个输出类。
(X, U)为训练输入，$(\underline{Y}, Sim_Y)$为待学习的分类器。

值不是0或者1

由于是不确定决策，V不是硬划分，也需要学习。
如果学习到 $(\underline{Y}, Sim_Y)$ ，V可以计算得到。

$$v_{ik} = P(Y_i \mid y_k)$$

贝叶斯分类器

对于贝叶斯分类，重要的是得到：

$$\hat{p}_i(x) \Rightarrow p_i(x) = p(x \mid X_i)$$ **类条件概率密度函数**

$$P(Y_i) \Rightarrow P(X_i)$$ **先验概率**

从而得到

$$P(Y_j \mid x) \Rightarrow P(X_j \mid x)$$ **后验概率**

属于密度估计问题，采用不同的估计$p_i(x)$、$P(X_i)$或者$P(X_j|x)$ 的方法，可以得到不同的贝叶斯分类器。

首先讲最简单的贝叶斯分类器——朴素贝叶斯分类

13.1 贝叶斯分类器

- 隶属度v_{ik}为

$$P(Y_i|\boldsymbol{x}) = \frac{p(\boldsymbol{x},Y_i)}{p(\boldsymbol{x})} = \frac{P(Y_i)p(\boldsymbol{x}|Y_i)}{p(\boldsymbol{x})} \tag{7.7}$$

$$P(Y_i|x_1,\cdots,x_d) = \frac{P(Y_i)\,p(x_1,\cdots,x_d|Y_i)}{p(x_1,\cdots,x_d)} \tag{7.8}$$

“类认知表示$Y_i = p(y|Y_i)$”
第i个输出类中样本 $\boldsymbol{x}$发生的概率，直接计算较为困难

“类先验概率”
第i个输出类发生的概率

“证据”（evidence）因子，与类标记无关

目录

13.2 朴素贝叶斯分类

假设输入特征空间中的每个特征只取有限离散值，则可以通过参数密度估计得到

$$\hat{P}_i(x) \Rightarrow P_i(x) = P(x \mid X_i)$$ 类条件概率或似然

$$P(Y_i) \Rightarrow P(X_i)$$ 先验概率

输入空间维数很高时，训练集中的样本过于稀疏。很多特征空间的元素并没有训练集中的样本落入。直接进行密度估计，许多 x对应的$\hat{P}_i(x)$为零，估计偏差太大。

假设不同维的特征彼此独立于类标，分别对每一维特征进行分布估计，然后根据独立性条件，将每一维的分布估计相乘得到分布估计$P_i(x)$。这就是朴素贝叶斯分类算法。

朴素贝叶斯分类

假设给定X_i时，每个特征之间是独立的，则

$$P(x \mid Y_i) = \prod_{r=1}^{p} P((x)_r \mid Y_i)$$

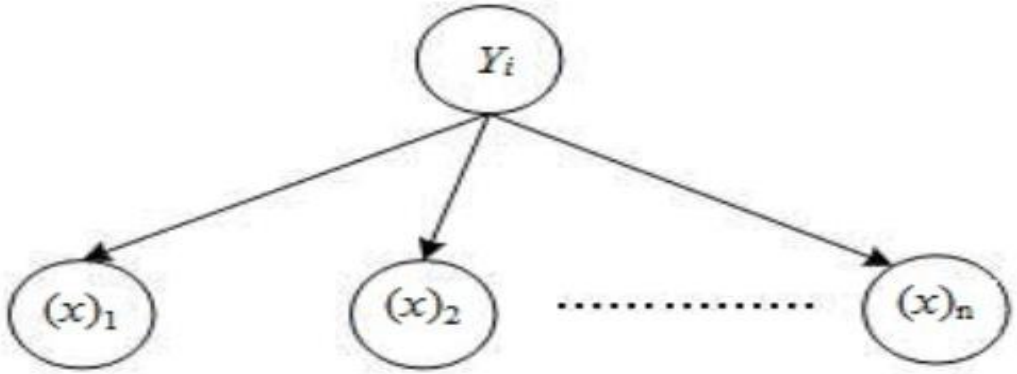
类条件概率

$$\text{sim}_Y(x, Y_i) = P(Y_i)P(x \mid Y_i) = P(Y_i)\prod_{r=1}^{p} P((x)_r \mid Y_i)$$

$(x)_r$表示 x 的第r个特征的特征值。

朴素贝叶斯分类

特征条件独立性假设指数据的所有特征向量都条件独立于类变量，即每一个特征变量都以类标号变量作为唯一父节点。



朴素贝叶斯分类模型结构

随机向量 $x=((x)_1,(x)_2,\cdots,(x)_p)^T$，$(x)_1,(x)_2,\cdots,(x)_p$ 是p个不同特征，可以看作p个随机变量，假设它们相互独立,则

$$P(x\,|\,Y_i)=P((x)_1,(x)_2,...,(x)_p\,|\,Y_i)=\prod_{r=1}^{p}P((x)_r\,|\,Y_i)$$

朴素贝叶斯分类

朴素贝叶斯分类忽略了特征间的条件依赖关系，故大大提高了运算效率和计算的可行性

假设第r个特征$(x)_r$的特征值集合是$\{a_{r1}, a_{r2}, \cdots, a_{rS_r}\}$

朴素贝叶斯分类的关键是估计

$P(Y_i)$ **先验概率**

$P(a_{rl}|Y_i) = P((x)_r = a_{rl}|Y_i)$ **类条件概率或似然**

$r = 1, 2, \ldots, p, l = 1, 2, \ldots, S_r, i = 1, \ldots, c$

参数估计

- 概率模型的训练过程就是参数估计过程，统计学界的两个学派提供了不同的解决方案：
  - **频率主义学派** (Frequentist)：
    - ✓认为参数虽然未知，但却客观存在**固定值**，因此可通过优化似然函数等准则来确定参数值；
    - ✓极大似然估计（Maximum Likelyhood Estiamation,MLE）就源自这个学派，根据数据采样来估计概率分布参数的经典方法；
  - **贝叶斯学派** (Bayesian)：
    - ✓认为参数是未观察到的随机变量，其本身也可有分布，因此可假定参数服从一个先验分布，然后基于观测到的数据来计算参数的后验分布。

13.2.1 最大似然估计

类紧致性准则希望得到具有最大类内相似性的类表示：

$$\max_{\underline{Y}} \prod_{k=1}^{N} Sim_Y(x_k, \underline{Y_{\vec{x}_k}}) = \max_{\underline{Y}} \prod_{k=1}^{N}\prod_{i=1}^{c} Sim_Y(x_k, \underline{Y_i})^{u_{ik}}$$

$$= \max_{\underline{Y}} \prod_{k=1}^{N}\prod_{i=1}^{c} (P(x_k \mid Y_i)\, p(Y_i))^{u_{ik}}$$

$$= \max_{\underline{Y}} \prod_{k=1}^{N}\prod_{i=1}^{c} (p(Y_i)\prod_{r=1}^{p} P(x_{rk} \mid Y_i))^{u_{ik}}$$

$$= \max_{\underline{Y}} \prod_{k=1}^{N}\prod_{i=1}^{c} (p(Y_i)\prod_{r=1}^{p}\prod_{l=1}^{S_r} P(a_{rl} \mid Y_i)^{\delta(x_{rk}-a_{rl})})^{u_{ik}}$$

$\delta()$是Kronecker函数

$$\sum_{i=1}^{c} p(Y_i) = 1, \sum_{l=1}^{S_r} P(a_{rl} \mid Y_i) = 1$$

约束条件

最大化此目标函数就是最大似然估计方法

最大似然估计方法是类紧致性准则的特例。

最大似然估计

- 朴素贝叶斯分类器的训练过程就是基于训练集估计类先验概率 $P(Y_i)$，并为每个属性估计类条件概率 $P(a_{rl}|Y_i)$
  - 若有充足的独立同分布样本，则可容易地估计出类先验概率

$$P(Y_i)=\frac{\sum_{k=1}^{N}u_{ik}}{N},\qquad i=1,2,\ldots,c$$

  - 对离散属性而言，统计对应属性值的样本数量，类条件概率可估计为

$$P(a_{rl}\,|\,Y_i)=P((x)_r=a_{rl}\,|\,Y_i)=\frac{\sum_{k=1}^{N}\delta(x_{rk}-a_{rl})u_{ik}}{\sum_{k=1}^{N}u_{ik}}$$

$$r=1,2,\ldots,p;\quad l=1,2,\ldots,S_r;\quad i=1,2,\ldots c$$

$\delta()$是Kronecker函数

当$n=0$时，$\delta(n)=1$；　$n\neq 0$时$\delta(n)=0$。

朴素贝叶斯分类例子-离散

举例 表13.1是一个名词性数据集，每一个数据包括五个特征和一个类标记。特征F1有三个属性值$\{s,o,r\}$，特征F2有三个属性值$\{h,m,c\}$，特征F3有两个属性值$\{h,n\}$，特征F4有两个属性值$\{t,f\}$，特征F5有两个属性值$\{d,r\}$,类标Class有两种$\{L_1,L_2\}$。假设五个特征是**相互独立的**，由该数据集训练一个朴素贝叶斯分类器并确定$x=\{s,m,h,t,d\}$的类标。

例 13.1

$$P((x)_1=s|L1)=\frac{3}{13},P((x)_1=o|L1)=\frac{6}{13}$$

$$P((x)_1=r|L1)=\frac{4}{13},P((x)_2=h|L1)=\frac{3}{13}$$

.....

表 13.1 只具有名称性特征的数据集

NO.	F1	F2	F3	F4	F5	Class
1	s	h	h	f	d	L2
2	s	h	h	t	d	L2
3	o	h	h	f	d	L1
4	r	m	h	f	d	L1
5	r	c	n	f	d	L1
6	r	c	n	t	d	L2
7	o	c	n	t	d	L1
8	s	m	h	f	d	L2
9	s	c	n	f	d	L1
10	r	m	n	f	d	L1
11	s	m	n	t	r	L1
12	o	m	h	t	r	L1
13	o	h	n	f	r	L1
14	r	m	h	t	r	L2
15	r	m	n	f	r	L1
16	s	m	n	t	r	L1
17	o	m	h	t	r	L1
18	o	h	n	f	r	L1
19	r	m	h	t	r	L2
20	r	c	n	t	r	L2

朴素贝叶斯分类例子-离散

对于给定$x=\{s,m,h,t,d\}$计算：

$$Sim_Y(x,\underline{Y_1}) = P(L1)\times P((x)_1=s|L1)\times P((x)_2=m|L1)\times P((x)_3=h|L1)\times P((x)_4=t|L1)\times P((x)_5=d|L1)$$

$$=\frac{13}{20}\times\frac{3}{13}\times\frac{7}{13}\times\frac{4}{13}\times\frac{5}{13}\times\frac{6}{13}=\frac{126}{13^4}$$

$$Sim_Y(x,\underline{Y_2}) = P(L2)\times P((x)_1=s|L2)\times P((x)_2=m|L2)\times P((x)_3=h|L2)\times P((x)_4=t|L2)\times P((x)_5=d|L2)$$

$$=\frac{7}{20}\times\frac{3}{7}\times\frac{3}{7}\times\frac{5}{7}\times\frac{5}{7}\times\frac{4}{7}=\frac{45}{7^4}$$

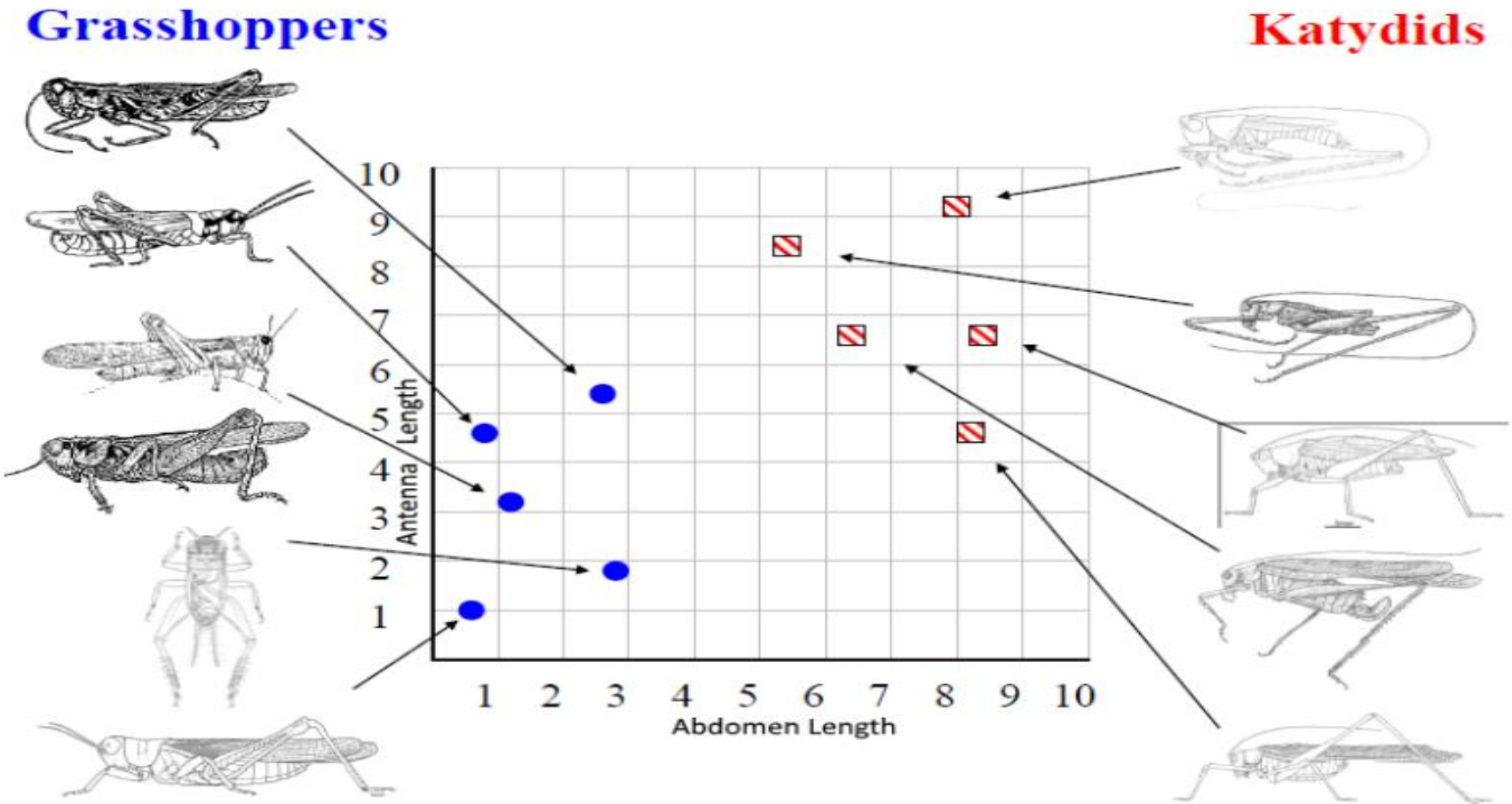
例 13.1

表 13.1　只具有名称性特征的数据集

NO.	F1	F2	F3	F4	F5	Class
1	s	h	h	f	d	L2
2	s	h	h	t	d	L2
3	o	h	h	f	d	L1
4	r	m	h	f	d	L1
5	r	c	n	f	d	L1
6	r	c	n	t	d	L2
7	o	c	n	t	d	L1
8	s	m	h	f	d	L2
9	s	c	n	f	d	L1
10	r	m	n	f	d	L1
11	s	m	n	t	r	L1
12	o	m	h	t	r	L1
13	o	h	n	f	r	L1
14	r	m	h	t	r	L2
15	r	m	n	f	r	L1
16	s	m	n	t	r	L1
17	o	m	h	t	r	L1
18	o	h	n	f	r	L1
19	r	m	h	t	r	L2
20	r	c	n	t	r	L2

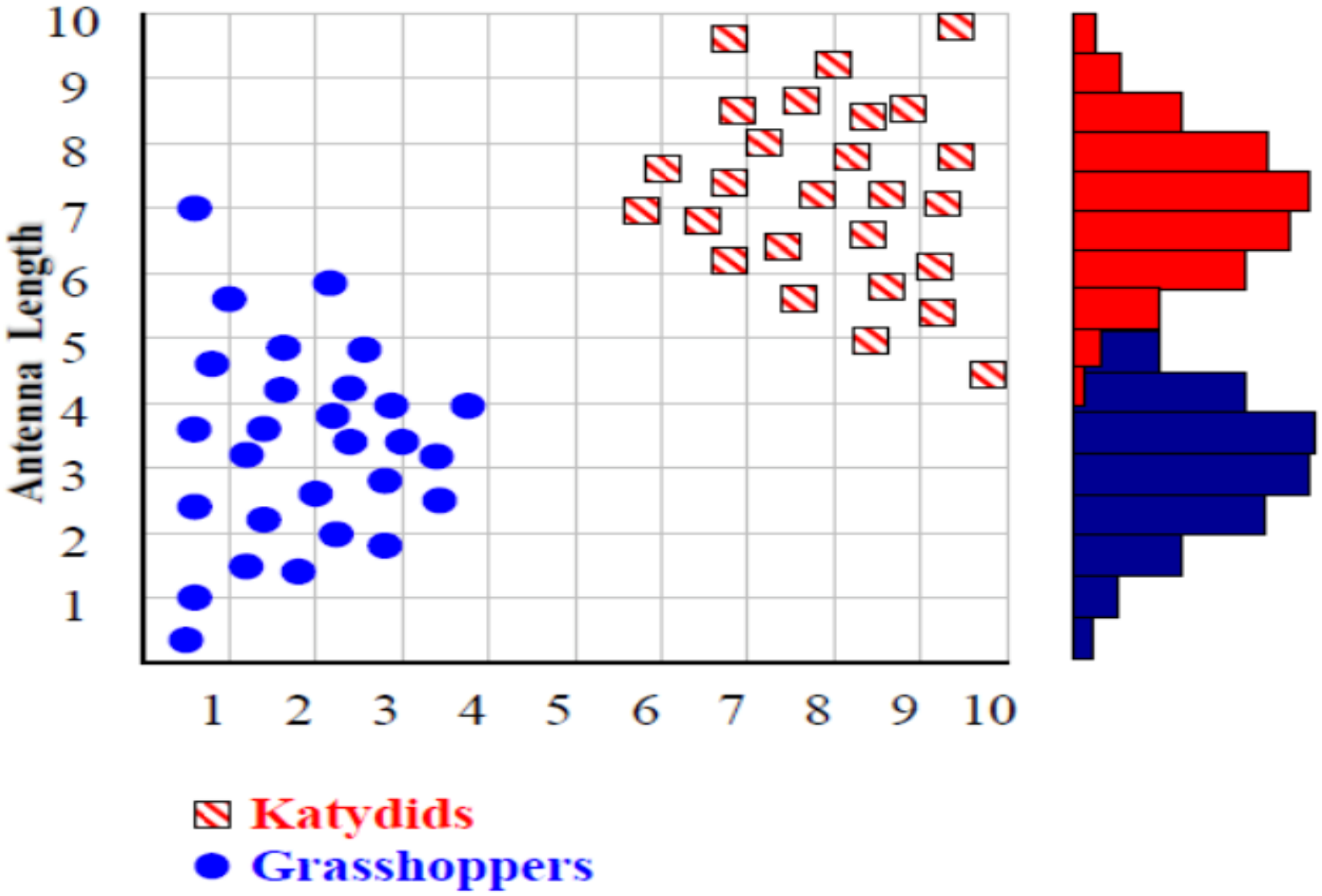
可知$x=\{s,m,h,t,d\}$的类标是L2。

举例 – 连续属性值



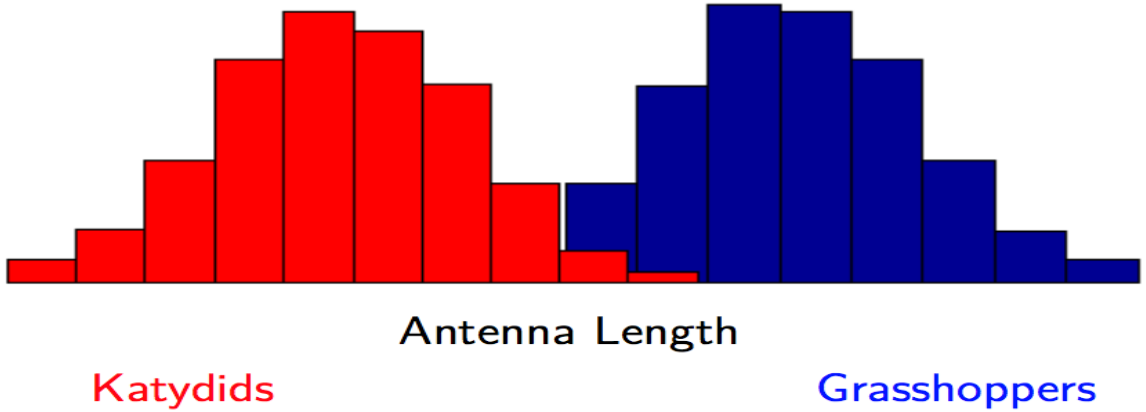
Adapted from http://www.cs.ucr.edu/~eamonn/CE/Bayesian%20Classification%20withInsect examples.pdf

举例 – 连续属性值

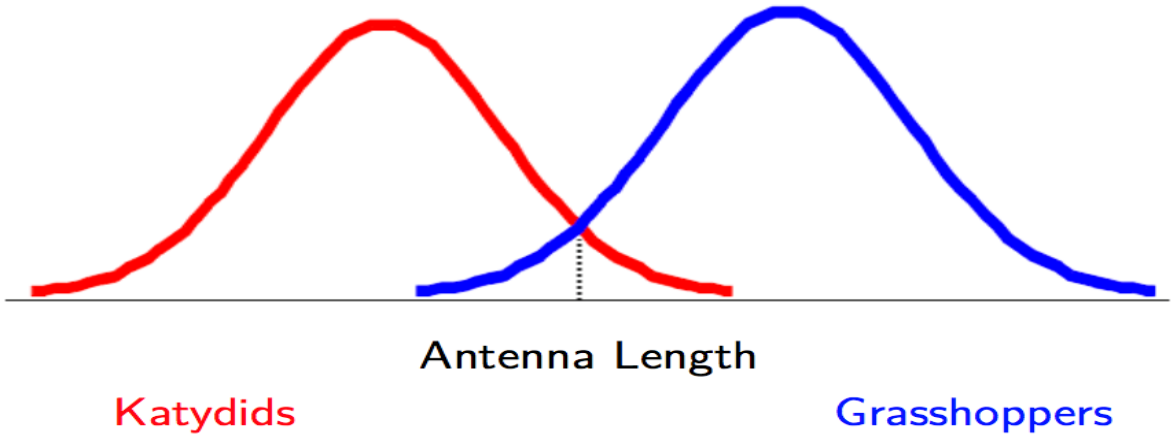

Antenna Length
10
9
8
7
6
5
4
3
2
1
1 2 3 4 5 6 7 8 9 10
Katydids
Grasshoppers

举例 – 连续属性值

- 各类物种触角长度直方图:

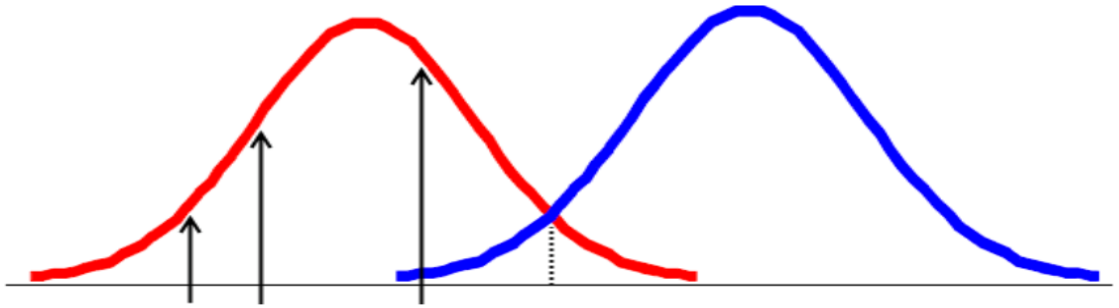


- 分布P(antenna_length | grasshopper) and P(antenna_length|katydids) :

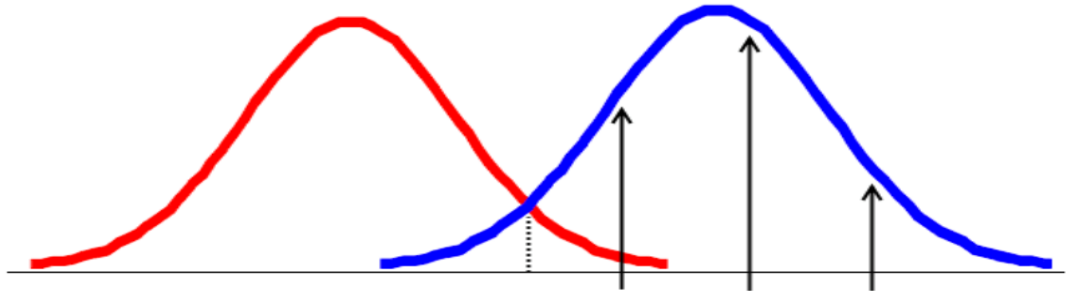


举例 – 连续属性值

- 给定触角长度, 可以判定物种的类别(grasshopper or katydid)



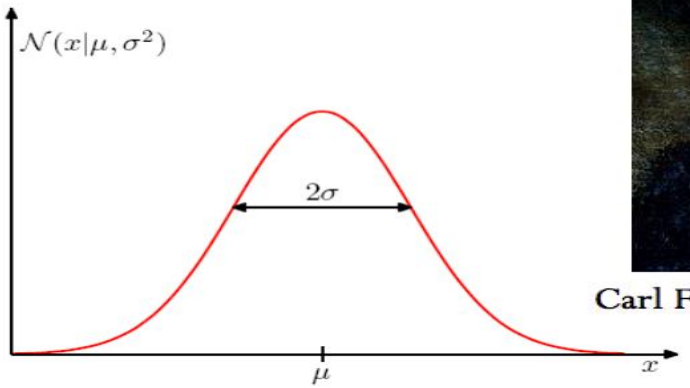
Katydid is more likely!



Grasshopper is more likely!

- 对于连续属性考虑概率密度函数，如高斯分布

$$p(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Carl F. Gauss (1777 – 1855)

朴素贝叶斯分类例子-连续

- 一组人类身体特征的统计资料。

性别	身高（英尺）	体重（磅）	脚掌（英寸）
男	6	180	12
男	5.92	190	11
男	5.58	170	12
男	5.92	165	10
女	5	100	6
女	5.5	150	8
女	5.42	130	7
女	5.75	150	9

已知某人身高6英尺、体重130磅，脚掌8英寸，请问该人是男是女？根据朴素贝叶斯分类器，计算下面这个式子的值。

P(身高|性别) x P(体重|性别) x P(脚掌|性别) x P(性别)

朴素贝叶斯分类例子-连续

- 可以假设男性和女性的身高、体重、脚掌都是正态分布，通过样本计算出均值和方差，也就是得到正态分布的密度函数。
- 可以把值代入密度函数，算出某一点的密度函数的值。
- 男性的身高是均值5.855、方差0.035的正态分布。所以，男性的身高为6英尺的概率的相对值等于1.5789
- （大于1并没有关系，因为这里是密度函数的值，只用来反映各个值的相对可能性）
  。

$$p(\text{height}|\text{male}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(\frac{-(6-\mu)^2}{2\sigma^2}\right) \approx 1.5789$$

朴素贝叶斯分类例子-连续

P(身高=6|男) x P(体重=130|男) x P(脚掌=8|男) x P(男) = 6.1984 x e-9

P(身高=6|女) x P(体重=130|女) x P(脚掌=8|女) x P(女) = 5.3778 x e-4

- 女性的概率比男性要高出将近10000倍，所以判断该人为女性。

13.2.2 贝叶斯估计

- 如果训练集中某些特征值 a_{rl}始终没有出现在样例中，用最大似然估计容易出现所要估计的概率值 $\forall i, P(a_{rl}|Y_i) = 0$，因此使得$\forall i, \mathrm{Sim}_Y(x, Y_i) = 0$。
  - 未出现的特征值会“抹去”其他的特征值对分类的影响，使分类产生严重偏差。

- 拉普拉斯最早提出：在分子分母同时加入一个正常数，在一定程度上削弱训练集中特征值缺失的影响，称为**拉普拉斯平滑方法**。
  - 这种方法的本质思想是假设未出现的特征值以一个特定的先验概率出现，

- 将这种思想加以普遍化，就是所谓的**贝叶斯估计**。

考虑类一致性

- 假设人们对于类输入认知表示$\underline{X}$有先验估计$\underline{\boldsymbol{X}}_{@}$。
- 由“类一致性准则”，期望$\underline{Y}$越接近$\underline{\boldsymbol{X}}_{@}$越好，或者$\underline{Y}$与$\underline{\boldsymbol{X}}_{@}$越相似越好（$\boldsymbol{Sim}(\underline{\boldsymbol{Y}}, \underline{\boldsymbol{X}}_{@})$越大越好）

Dirichlet分布是分布之上的分布

$$\mathbf{Sim}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta}_0) = \boldsymbol{p}(\hat{\boldsymbol{\theta}} | \boldsymbol{\theta}_0) = \frac{\boldsymbol{\Gamma}(\boldsymbol{\alpha}_0)}{\boldsymbol{\Gamma}(\boldsymbol{\alpha}_1) \ldots \boldsymbol{\Gamma}(\boldsymbol{\alpha}_c)} \prod_{\boldsymbol{i}=\mathbf{1}}^{\boldsymbol{c}} \hat{\boldsymbol{p}}_{\boldsymbol{i}}^{\boldsymbol{\alpha}_{\boldsymbol{i}}-\mathbf{1}}$$

$$Sim(\underline{Y}, \underline{X}_{@}) = p(\underline{Y} \mid \underline{X}_{@})$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_c)} \prod_{i=1}^{c} P(Y_i)^{\alpha_i - 1} \prod_{r=1}^{p} \frac{\Gamma(\alpha_{ri})}{\Gamma(\alpha_{rli})\ldots\Gamma(\alpha_{rS_r i})} \times \prod_{l=1}^{S_r} P(a_{rl} \mid Y_i)^{\alpha_{rli}-1}$$

gamma函数

$$\Gamma(x) = \int_0^{\infty} t^{x-1} e^{-t} dt$$

$$\forall i, \theta_i = \frac{1-\alpha_i}{c-\alpha_0}, \alpha_0 = \sum_{i=1}^{c} \alpha_i, \forall l, \theta_{rli} = \frac{1-\alpha_{rli}}{S_r - \alpha_{ri}}, \alpha_{ri} = \sum_{l=1}^{S_r} \alpha_{rli}$$

考虑类一致性和紧致性的目标函数

- **满足紧致性准则**

$$\prod_{k=1}^{N} Sim_Y(x_k, \underline{Y}_{\underline{x}_k}) = \max_{\underline{Y}} \prod_{k=1}^{N} \prod_{i=1}^{c} \left(p(Y_i) \prod_{r=1}^{p} \prod_{l=1}^{S_r} P(a_{rl} \mid Y_i)^{\delta(x_{rk} - a_{rl})} \right)^{u_{ik}}$$

- **同时考虑“类一致性”+“类紧致性”，最大化下边的目标函数就是贝叶斯估计方法。**

$$\mathrm{Sim}(\underline{Y}, X_@) \prod_{k=1}^{N} \mathrm{Sim}_Y(x_k, \underline{Y}_{\underline{x}_k})$$

$$= \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\ldots\Gamma(\alpha_c)} \prod_{i=1}^{c} P(Y_i)^{\alpha_i - 1} \prod_{r=1}^{p} \frac{\Gamma(\alpha_{ri})}{\Gamma(\alpha_{rli})\ldots\Gamma(\alpha_{rS_r i})} \prod_{l=1}^{S_r} P(a_{rl} \mid Y_i)^{\alpha_{rli} - 1}$$

$$\times \prod_{k=1}^{N} \prod_{i=1}^{c} \left(P(Y_i) \prod_{r=1}^{p} \prod_{l=1}^{S} P(a_{rl} \mid Y_i)^{\delta(x_{rk} - a_{rl})} \right)^{u_{ik}}$$

$$\text{s.t.} \quad \sum_{i=1}^{c} p(Y_i) = 1$$

$$\sum_{l=1}^{S_r} P(a_{rl} \mid Y_i) = 1$$

贝叶斯估计

拉格朗日乘子法

- **先验概率**

$$P(Y_i)=\frac{\alpha_i-1+\sum_{k=1}^{N}u_{ik}}{N+\alpha_0-c},\qquad i=1,2,...,c$$

- **条件概率**

$$P(a_{rl}\mid Y_i)=P((x)_r=a_{rl}\mid Y_i)=\frac{\alpha_{rli}-1+\sum_{k=1}^{N}\delta(x_{rk}-a_{rl})u_{ik}}{\alpha_{ri}-S_r+\sum_{k=1}^{N}u_{ik}}$$

$$r=1,2,...,p;\quad l=1,2,...,S_r;\quad i=1,2,...c$$

常用的贝叶斯估计公式

$$\forall i \forall r \forall l, \alpha_i - 1 = \alpha_{rli} - 1 = \lambda$$

- **先验概率**

$$P_\lambda(Y_i) = \frac{\lambda + \sum_{k=1}^{N} u_{ik}}{N + c\lambda}, \qquad i = 1, 2, \ldots, c$$

- **条件概率**

$$P_\lambda(a_{rl} \mid Y_i) = P_\lambda((x)_r = a_{rl} \mid Y_i) = \frac{\lambda + \sum_{k=1}^{N} \delta(x_{rk} - a_{rl}) u_{ik}}{S_r \lambda + \sum_{k=1}^{N} u_{ik}}$$

$$r = 1, 2, \ldots, p; \quad l = 1, 2, \ldots, S_r; \quad i = 1, 2, \ldots c \qquad \lambda \geq 0$$

式中$\lambda \geq 0$，**当$\lambda = 0$时贝叶斯估计退化为最大似然估计**

当$\lambda = 1$时，贝叶斯估计变为拉普拉斯平滑估计

小结

- 朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。
- 首先基于特征条件独立假设学习输入/输出的联合概率分布；然后基于此模型，对给定的输入x，利用贝叶斯定理求出后验概率，将x归为最大后验概率所对应的类。
- 实现简单，学习与预测的效率都很高，是一种常用的方法。
- 朴素贝叶斯法将实例分到后验概率最大的类中，等价于期望风险最小化。

目录

13.3 最小风险贝叶斯决策

把分类错误引起的“损失”加入到决策中去。

决策论中：采取的决策称为动作，用a_i表示；

每个动作带来的损失，用λ表示。

状态空间Ω由c个自然状态(c类)组成

$$\Omega = \{\omega_1, \omega_2, \cdots, \omega_c\}$$

决策空间A由a个决策a_i组成，　$i = 1, 2, \cdots, a$

$$\mathbf{A} = \{a_1, a_2, \cdots a_a\}$$

损失函数$\lambda(a_i, \omega_j)$，　$i = 1, 2, \cdots, a, j = 1, 2, \cdots, c$

$\lambda(a_i, \omega_j)$表示：当真实状态为ω_j时采取决策a_i的损失。

最小风险的叶斯决策

一般用**决策表**或**损失矩阵**表示上述三者关系。

决策表**表示各种状态下的决策损失**，如下表：

状态 / 损失 / 决策	自然状态 ω_1	ω_2	…	ω_j	…	ω_c
α_1	$\lambda(\alpha_1,\omega_1)$	$\lambda(\alpha_1,\omega_2)$	…	$\lambda(\alpha_1,\omega_j)$	…	$\lambda(\alpha_1,\omega_c)$
α_2	$\lambda(\alpha_2,\omega_1)$	$\lambda(\alpha_2,\omega_2)$	…	$\lambda(\alpha_2,\omega_j)$	…	$\lambda(\alpha_2,\omega_c)$
⋮	⋮	⋮	… … …	⋮	⋮	⋮
α_i	$\lambda(\alpha_i,\omega_1)$	$\lambda(\alpha_i,\omega_2)$	…	$\lambda(\alpha_i,\omega_j)$	…	$\lambda(\alpha_i,\omega_c)$
⋮	⋮	⋮	… … …	⋮	⋮	⋮
α_a	$\lambda(\alpha_a,\omega_1)$	$\lambda(\alpha_a,\omega_2)$	…	$\lambda(\alpha_a,\omega_j)$	…	$\lambda(\alpha_a,\omega_c)$

最小风险贝叶斯决策

由于引入了“损失”的概念，不能只根据后验概率来决策，必须考虑所采取的决策是否使损失最小。

- 对于给定的x，决策a_i，此时的条件期望损失：

$$R(a_i|x)=E[\lambda(a_i,\omega_j)]=\sum_{j=1}^{c}\lambda(a_i,\omega_j)P(\omega_j|x)\quad i=1,2,\cdots,a$$

- 在决策论中条件期望损失称为条件风险，即x被判为i类时损失的均值。

最小风险贝叶斯决策

- 决策a可看成随机向量x的函数，记为$a(x)$，它本身也是一个随机变量。定义**期望风险**R

$$R = \int R(a(x)|x)p(x)dx$$

dx是d维特征空间的体积元，积分在整个特征空间。

- **期望风险**R反映对整个特征空间上所有x的取值都采取相应的决策$a(x)$所带来的**平均风险**；

- **条件风险**$R(a_i|x)$只反映观察到**某一x**的条件下采取决策a_i所带来的风险。

最小风险贝叶斯决策

$$R = \int R(a(x)|x)p(x)dx$$

如果采取每个决策行动a_i使条件风险$R(a_i|x)$最小，则对所有的x作出决策时，其期望风险R也必然最小。

这就是**最小风险Bayes决策**。

最小风险贝叶斯决策

最小风险Bayes决策步骤：

①已知$P(\omega_j), p(x|\omega_j)$，根据待识别的$x$，由Bayes公式计算后验概率$P(\omega_j|x)$；

②利用决策表，计算出采取a_i决策的条件风险$R(a_i|x)$

$$R(a_i|x) = \sum_{i=1}^{c} \lambda(a_i|\omega_j)\, P(\omega_j|x),\ \ i = 1,2,\cdots,\mathbf{c}$$

③上式得到的a个条件风险值$R(a_i|x)$，$i = 1,2,\cdots,\mathbf{c}$

找出使条件风险最小的决策a_k

即 $R(a_k|x) = \min\limits_{i=1,2,\cdots,\mathbf{c}} R(a_i|x)$

则 a_k是最小风险Bayes决策

最小风险贝叶斯决策

两类情况的最小风险贝叶斯决策

$$R(a_1 \mid x) = \lambda_{11} P(\omega_1 \mid x) + \lambda_{12} P(\omega_2 \mid x)$$
$$R(a_2 \mid x) = \lambda_{21} P(\omega_1 \mid x) + \lambda_{22} P(\omega_2 \mid x)$$

这时最小风险的Bayes决策法则为：

如果$R(a_1/x) < R(a_2/x)$，则x的真实状态w_1，否则w_2。

■**两类时最小风险Bayes决策规则的另两种形式：**

- 如果 $(\lambda_{21} - \lambda_{11}) P(\omega_1 \mid x) > (\lambda_{12} - \lambda_{22}) P(\omega_2 \mid x)$,
  则决策ω_1；否则ω_2
- 如果 $l(x) = \dfrac{p(x|\omega_1)}{p(x|\omega_2)} > \dfrac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \dfrac{P(\omega_2)}{P(\omega_1)}$,
  则决策ω_1；否则ω_2

最小风险贝叶斯决策

应用实例： 细胞识别

已知：正常类$P(w_1)=0.9$； 异常

细胞 x， 从类条件概率密度曲线上

$p(x|w_2)=0.4$

解：利用Bayes公式分别计算ω_1和

$$P(\omega_1|x)=\frac{p(x|\omega_1)P(\omega_1)}{\sum_{j=1}^{2}p(x|\omega_j)P(\omega_j)}$$

$$P(\omega_2|x)=1-P(\omega_1|x)=0.182$$

因此 $P(\omega_1|x)=0.818>P(\omega_2|x)=0.182$ $x\in\omega_1$

这里决策与最小错误率决策结论相反，损失起了主导作用。

λ不易确定，要与有关专家商定。

$$\lambda_{11}=0,\ \lambda_{12}=6,\ \lambda_{21}=1,\ \lambda_{22}=0$$

条件风险 $R(a_1|x)=\sum_{j=1}^{2}\lambda_{1j}P(\omega_j|x)=\lambda_{12}P(\omega_2|x)=1.092$

$$R(a_2|x)=\lambda_{21}P(\omega_1|x)=0.818$$

由于$R(a_1|x)>R(a_2|x)$, 所以$x\in\omega_2$

最小化风险分类

- 分类问题中，类唯一性公理通常不成立，一定会存在被分错的样本。样本分到各类的错误代价有所不同，设计相似性(相异性)映射时，需考虑错分成本，使错分成本最小的类相似性最大。
  - 设输入实际属于X_j却属于Y_i而导致的损失或者成本为λ_{ji}，样本被指派到Y_i的风险定义为：

$$R(Y_i \mid x) = \sum_{j=1}^{c} \lambda_{ji} P(X_j \mid x)$$

  - 样本x与类认知表示$\underline{Y_i}$的类相异性映射定义为期望风险：

$$\mathrm{Ds}_Y(y, \underline{Y_i}) = \mathrm{Ds}_Y(x, \underline{Y_i}) = R(Y_i \mid x)$$

  - 每个样本被指派到相异性最小的类中，样本x的预测函数为：

$$\arg\min_i \mathrm{Ds}_Y(x, \underline{Y_i}) = \arg\min_i R(Y_i \mid x)$$

0-1损失下最小化风险分类

- **假设采用0-1损失，定义损失为：**

$$\lambda_{ji} = \begin{cases} 0, & i = j \\ 1, & i \neq j \end{cases}$$

$$R(Y_i \mid x) = \sum_{j=1}^{c} \lambda_{ji} P(\underline{Y_j} \mid x)$$

$$= \sum_{j \neq i} P(\underline{Y_j} \mid x)$$

$$= 1 - P(\underline{Y_i} \mid x)$$

此时，最小化风险等价于最大后验的贝叶斯分类器

目录

13.4 效用最大化分类

- **样本分类错误成本很低，一旦分类正确，收益很大。比如，购买彩票。设计类相似性映射函数时，需考虑正确分类的收益（或效用）。**
  - 设输入实际属于X_j却属于Y_i而导致的效用或者收益为U_{ji}，样本被指派到Y_i的效用（或者收益）为：

$$U(Y_i \mid x) = \sum_{j=1}^{c} U_{ji} P(\underline{Y_j} \mid x)$$

  - 样本x与类认知表示$\underline{Y_i}$的类相似性映射定义为期望效用：

$$\mathrm{Sim}_Y(y, \underline{Y_i}) = \mathrm{Sim}_Y(x, \underline{Y_i}) = U(Y_i \mid x)$$

  - 效用最大化分类将类相似性最大的类作为样本的指派，类预测函数：

$$\arg\min_i \mathrm{Sim}_Y(x, \underline{Y_i}) = \arg\max_i U(Y_i \mid x)$$

分类方法

查表方法

- **基于训练集，事先计算朴素贝叶斯分类器所涉及的所有估值，对预测数据进行查表式的判别（预测速度较高要求）**

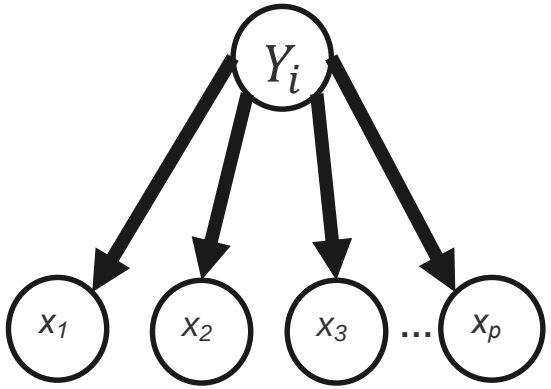
懒惰学习方法

- **不进行任何训练，待收到预测请求时再根据当前数据集进行概率估值**

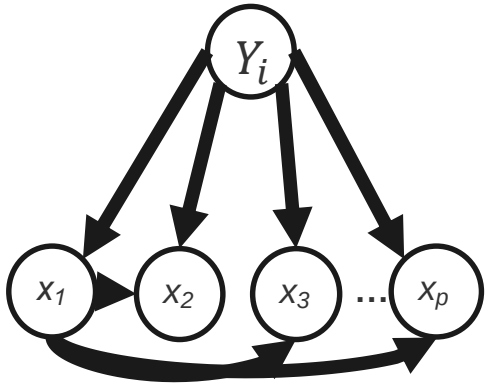
增量学习方法

- **数据不断增加时，可在现有估值基础上仅对新增样本的属性值所涉及的概率估值进行计数修正**

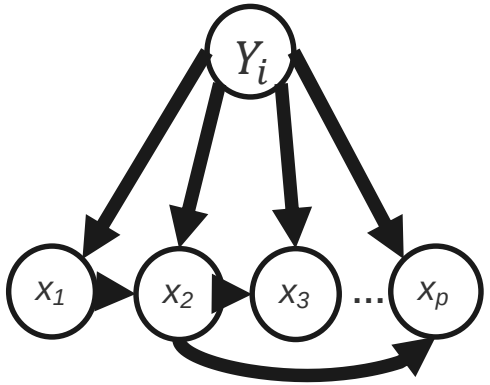
扩展：半朴素贝叶斯分类器



Naïve Bayes分类器



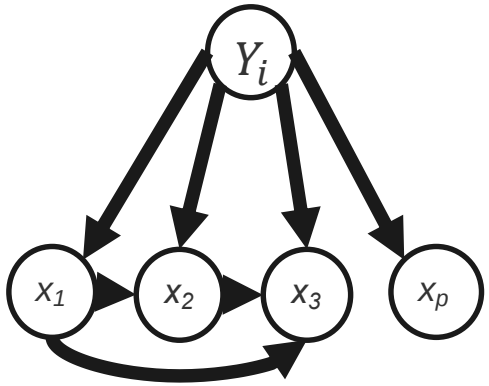
SPODE分类器



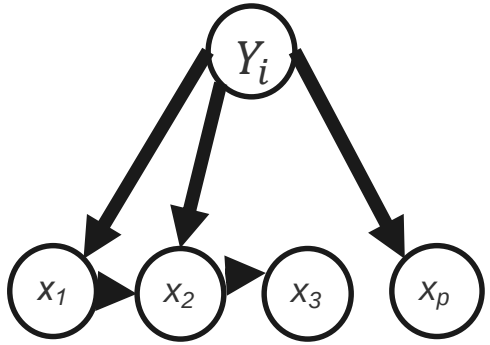
TAN分类器

- 朴素贝叶斯分期器：所有特征变量以类标号变量为唯一父节点。
- **SPODE(Super-Parent ODE)**分类器：假设所有特征变量都依赖于同一个特征变量x_1（“超父”特征变量），可通过交叉验证等模型选择方法来确定超父特征变量。
- **TAN**分类器：其中部分特征变量之间具有强相关性，使用条件互信息刻画两个特征变量之间的相关性。

扩展：半朴素贝叶斯分类器



BAN分类器



GBN分类器

- BAN(BN Augmented Naive Bayes)分类器：进一步扩展TAN分类器，允许各特征变量所对应的结点之间的关系构成一个图，而不只是树。
- GBN(General Bayesian Network)分类器：是一种无约束的贝叶斯网络分类器。和其他贝叶斯分类器的区别是：在其他分类器中均将类标号变量所对应的结点作为一个特殊的结点（即是各特征变量的父结点），而GBN中将类标号变量作为一普通结点。

BJTU “Machine Learning” Group

于　剑：jianyu@bjtu.edu.cn；

景丽萍：lpjing@bjtu.edu.cn；

田丽霞：lxtian@bjtu.edu.cn；

黄惠芳：hfhuang@bjtu.edu.cn；

李晓龙：hlli@bjtu.edu.cn；

吴　丹：wudan@bjtu.edu.cn；

万怀宇：hywan@bjtu.edu.cn；

王　晶：wj@bjtu.edu.cn.