贝叶斯分类器

赵海涛

haitaozhao@ecust.edu.cn

大纲

- 贝叶斯决策
- 朴素贝叶斯
- 例子

贝叶斯

- **贝叶斯**(约1701-1761) Thomas Bayes，英国数学家。约1701年出生于伦敦，做过神甫。1742年成为英国皇家学会会员。1761年4月7日逝世。
- 贝叶斯在数学方面主要研究概率论。他首先将归纳推理法用于概率论基础理论，并创立了贝叶斯统计理论，对于统计决策函数、统计推断、统计的估算等做出了贡献。他死后，理查德·普莱斯(Richard Price)于1763年将他的著作《机会问题的解法》(An essay towards solving a problem in the doctrine of chances)寄给了英国皇家学会，对于现代概率论和数理统计产生了重要的影响.

贝叶斯决策

- **贝叶斯决策**方法是统计模型决策中的一个基本方法，其基本思想是：
  1. 已知类条件概率密度参数表达式和先验概率。
  2. 利用贝叶斯公式转换成后验概率。
  3. 根据后验概率大小进行决策分类。

贝叶斯网络的应用

最早的PathFinder系统，该系统是淋巴疾病诊断的医学系统，它可以诊断60多种疾病，涉及100多种症状;后来发展起来的Internist－I系统，也是一种医学诊断系统，但它可以诊断多达600多种常见的疾病。

1995年，微软推出了第一个基于贝叶斯网的专家系统，一个用于幼儿保健的网站OnParent (www.onparenting.msn.com)，使父母们可以自行诊断。

贝叶斯网络的应用

(1)故障诊断(diagnose)

(2)专家系统(expert system)

(3)规划(planning)

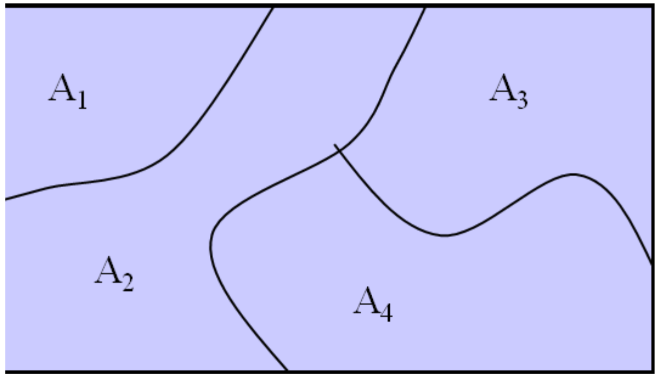
(4)学习(learning)

(5)分类(classifying)

贝叶斯决策理论

几个重要的概率公式

S



- $0 < P(A_i) < 1$
- $P(S) \ = \ 1$ (S 是样本空间)
- 如果 $A_1, A_2, \dots, A_N$ 互斥事件 ($P\left(A_i \cap A_j\right) = \ 0,\ i \neq j$), 则

$$P(A_1 \cup A_2 \cup \cdots \cup A_N) = \sum_{i=1}^{N} P(A_i)$$

几个重要的概率公式

- 条件概率公式

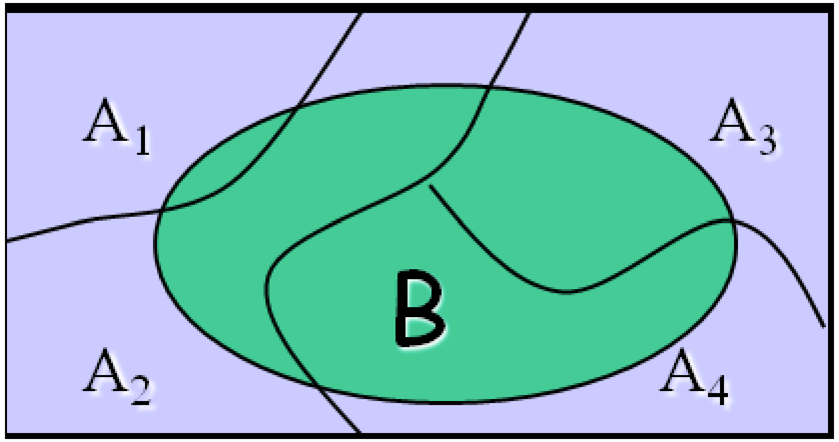
$$P(A \mid B) = \frac{P(A,B)}{P(B)} \qquad P(B \mid A) = \frac{P(A,B)}{P(A)}$$

- 条件概率的链式法则:

$$P(A,B) = P(A|B)P(B) = P(B|A)P(A)$$

全概率公式

S



- 如果 A_1, A_2，… ，A_N 是互斥事件且是对样本空间的一个划分，B 是任意事件，则有

$$P(B) = \sum_{i=1}^{N} P(B \mid A_i) P(A_i)$$

贝叶斯公式

- 贝叶斯公式（贝叶斯准则，贝叶斯定理）：

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

- 如果 $A_1, A_2, \ldots, A_N$ 是互斥事件且是对样本空间的一个划分，B 是任意事件，则贝叶斯公式为:

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A)}{P(B)}$$

其中

$$P(B) = \sum_{i=1}^{N} P(B \mid A_i)P(A_i)$$

独立

- 事件 A 和 B 相互独立，当且仅当：

$$P(A,B) = P(A)P(B)$$

- 由上面的公式，我们可以得到：

$$P(A|B) = P(A), \qquad P(B|A) = P(B)$$

- A 和 B 在给定事件C 的条件下相互独立，当且仅当：

$$P(A|B,C) = P(A|C)$$

独立与条件独立

- $P(AB) \neq P(A)P(B), P(AB|C) = P(A|C)P(B|C)$
- $P(AB) = P(A)P(B), P(AB|C) \neq P(A|C)P(B|C)$

几个例子

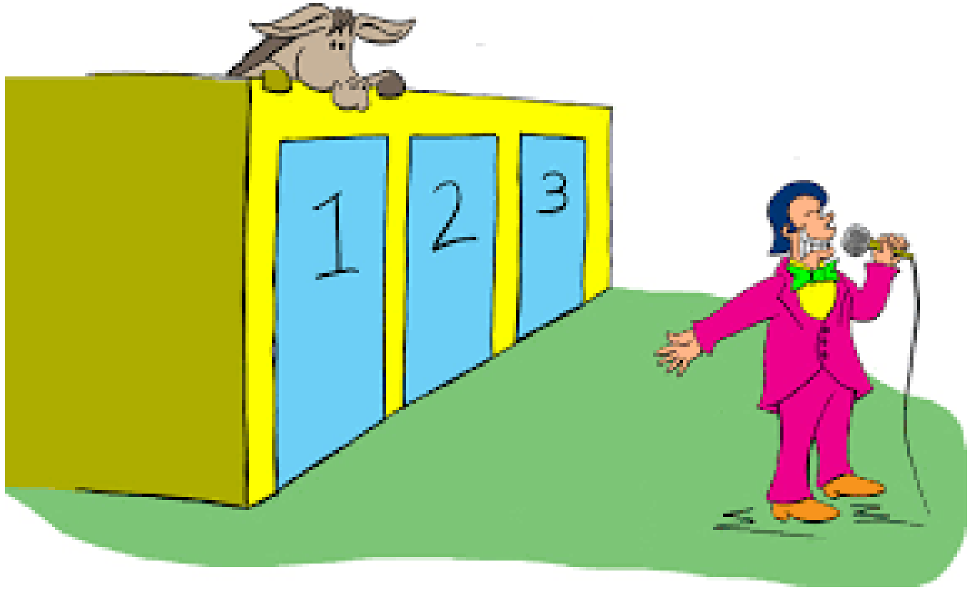
- 假设有三个看起来完全一样的盒子，每个盒子都有确定数量的红色和蓝色的球，它们除了颜色之外完全相同。第i个盒子中红色球和蓝色的数量分别为r_i和b_i，其中$i=1,2,3$。本试验就是随机取一个盒子，然后从该盒子中随机取出一个球。结果是红色球。考虑结果为红色球的基础上，计算球属于一号盒子的概率。

几个例子

181. 已知某酒鬼有90%的日子都会出去喝酒，喝酒只去固定三家酒吧。今天警察找了其中两家酒吧都没有找到酒鬼。问：酒鬼在第三家酒吧的几率？

[数学天地] 难度:4星

几个例子



题目：假设你参加一个电视游戏节目，节目现场有三扇门，其中一扇门后面是一辆车，另外两扇门后面则是山羊。主持人让你选择其中的一扇门。不妨假设你选择了一号门吧。主持人故意打开了另外一扇门，比如说三号门，让你看见三号门的后面是山羊。然后主持人问你，“你想改变你的选择，换成二号门吗？”这时候，你会怎么做？

The **Monty Hall problem** is a brain teaser, in the form of a probability puzzle, loosely based on the American television game show *Let's Make a Deal* and named after its original host, Monty Hall. The problem was originally posed (and solved) in a letter by Steve Selvin to the *American Statistician* in 1975.[1][2] It became famous as a question from reader Craig F. Whitaker's letter quoted in Marilyn vos Savant's "Ask Marilyn" column in *Parade* magazine in 1990:[3]

术语

- 模式状态 ω(随机变量):
  - ✓ ω_1 表示鲈鱼， ω_2 表示三文鱼
- 概率 $P(\omega_1)$ 和 $P(\omega_2)$ （先验）:
  - ✓ 先验知识：有多大的可能性得到一条鲈鱼或一条三文
- 概率密度函数 $p(x)$ （证据）:
  - ✓ 对模式的某一特征x进行测量，出现的频率值

    （例如，x是亮度测量）

Note: if x and y are different measurements, $p(x)$ and $p(y)$ correspond to different pdfs: $p_X(x)$ and $p_Y(y)$

术语

- 类条件概率密度 $p(x|\omega_j)$ （似然）：
  - 在模式属于 ω_j 类的条件下，对模式的某一特征x进行测量，出现的频率值

右图：三文鱼和鲈鱼的类条件概率密度

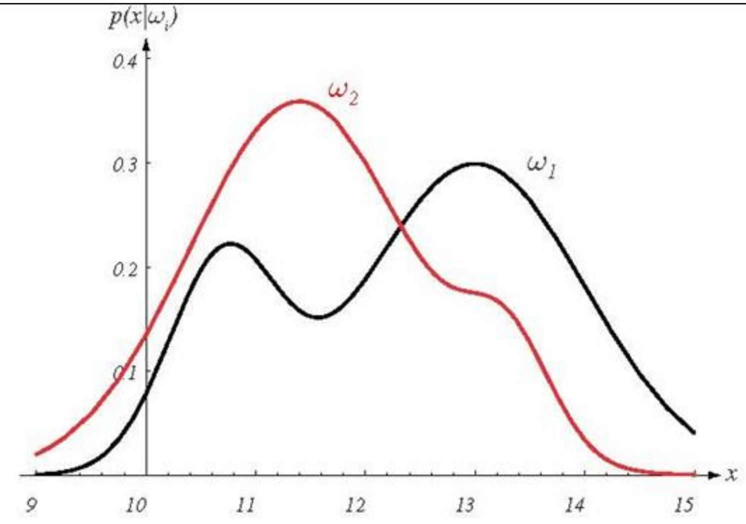


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons,

术语

- 条件概率 $P(\omega_j|x)$（后验）：
  - 在给点给特征x 测量值的条件下，模式属于ω_j类的可能性.

Note: we will be using an uppercase $P(\cdot)$ to denote a probability mass function (pmf) and a lowercase $p(\cdot)$ to denote a probability density function (pdf).

仅使用先验的决策规则

- Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2
- $P(error) = \min[P(\omega_1), P(\omega_2)]$
- 倾向于选择可能出现频率高的类...(在没有其它信息的条件下最优).
- 总是得到相同的决策!
- 只做一次决策是有一定道理的...

运用条件概率进行决策

- 运用贝叶斯公式，后验概率可表示为:

$$P(\omega_j \mid x) = \frac{p(x \mid \omega_j)P(\omega_j)}{p(x)} = \frac{\text{likelihood} \times \text{prior}}{\text{evidence}}$$

其中

$$p(x) = \sum_{i=1}^{2} p(x \mid \omega_j)P(\omega_j)$$

运用条件概率进行决策

- $P(\omega_1) = \frac{2}{3}, \ P(\omega_2) = \frac{1}{3}$

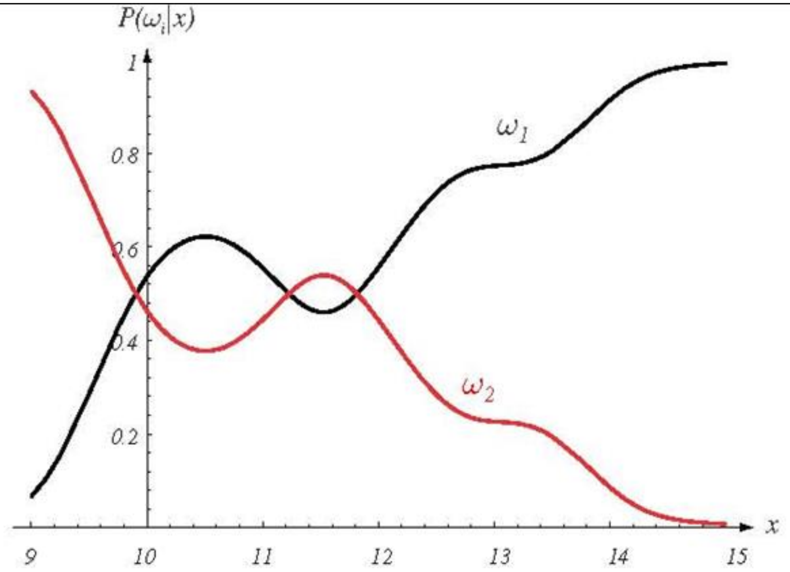
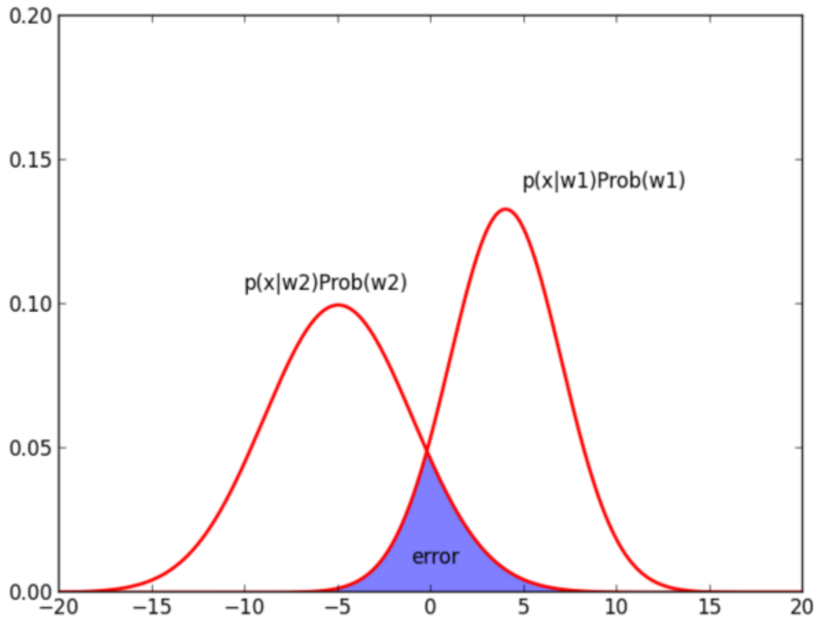


FIGURE 2.2. Posterior probabilities for the particular priors $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ for the class-conditional probability densities shown in Fig. 2.1. Thus in this case, given that a pattern is measured to have feature value $x = 14$, the probability it is in category ω_2 is roughly 0.08, and that it is in ω_1 is 0.92. At every x, the posteriors sum to 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons, Inc.

运用条件概率进行决策

- 错误概率:

$$P(\text{ error } \mid x) = \begin{cases} P(\omega_1 \mid x) & \text{if we decide } \omega_2 \\ P(\omega_2 \mid x) & \text{if we decide } \omega_1 \end{cases}$$



- 平均错误概率:

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

- 运用贝叶斯准则可得到最优决策，既使平均错误概率最小化，因为

$$P(\text{ error } |x) = \min\left[P(\omega_1|x), P(\omega_2|x)\right]$$

概率或密度函数如何得到?

- 如果概率已知，贝叶斯准则是最优的
- 有两种方式可以得到贝叶斯准则所需的概率:

1. 相对频率方法 (客观). 概率只能通过实验得到
2. 贝叶斯方法 (主观). 概率值可以反映某种程度的信念，可以基于实验也可以基于某种观点

例子

- 对鲈鱼和三文鱼进行分类
- 特征 x：鱼的亮度
- 根据贝叶斯准则，我们需要计算

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

- 需要

$$p(x \mid \omega_j) \text{ 和 } P(\omega_j), \qquad j = 1,2$$

例子

通过收集的数据确定先验概率：对两类雨的数量进行计数并计算.

例如，1000个样本： $\#\omega_1 = 900$，$\#\omega_2 = 100$

$$P(\omega_1) = \frac{900}{1000} = 0.9$$

$$P(\omega_2) = \frac{100}{1000} = 0.1$$

例子

- 确定类条件概率密度（似然）$p(x \mid \omega_j)$ $(j = 1,2)$
  - 离散化鱼的亮度值，使其落入某个小的区间，并使用归一化的直方图

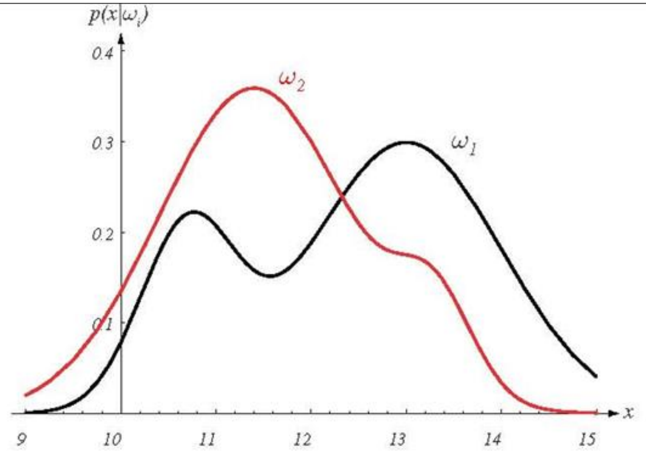


FIGURE 2.1. Hypothetical class-conditional probability density functions show the probability density of measuring a particular feature value x given the pattern is in category ω_i. If x represents the lightness of a fish, the two curves might describe the difference in lightness of populations of two types of fish. Density functions are normalized, and thus the area under each curve is 1.0. From: Richard O. Duda, Peter E. Hart, and David G. Stork, *Pattern Classification*. Copyright © 2001 by John Wiley & Sons,

如对某个x，可得到 $p(x|\omega_1) = 0.2$ 和 $p(x|\omega_2) = 0.4$

例子

- 计算后验概率

$$P(\omega_1 \mid x) = \frac{p(x \mid \omega_1)P(\omega_1)}{\sum_{j=1}^{2} p(x \mid \omega_j)P(\omega_j)} = \frac{0.2 \times 0.9}{0.2 \times 0.9 + 0.4 \times 0.1} = 0.818$$

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

朴素贝叶斯

朴素贝叶斯基本方法

- 训练数据集：$T=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_N,y_N)\}$
- 由X和Y的联合概率分布$P(X,Y)$独立同分布产生

- 朴素贝叶斯通过训练数据集学习联合概率分布$P(X,Y)$，
  - 即先验概率分布：$P(Y=c_k),\ k=1,2,\cdots,K$
  - 及条件概率分布：$P(X=\boldsymbol{x}\mid Y=c_k)=P\left(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\mid Y=c_k\right),\ k=1,2,\cdots,K$

于是学习到了联合分布概率

- 注意：条件概率为指数级别的参数：$K\prod_{j=1}^{n}S_j$

朴素贝叶斯基本方法

- 条件独立性假设:

$$P(X=\boldsymbol{x} \mid Y=c_k) \quad = P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)} \mid Y=c_k)=\prod_{j=1}^{n} P(X^{(j)}=x^{(j)} \mid Y=c_k)$$

- “朴素”贝叶斯名字由来，牺牲分类准确性

- 贝叶斯定理：$P(Y=c_k \mid X=\boldsymbol{x})=\dfrac{P(X=\boldsymbol{x}|Y=c_k)P(Y=c_k)}{\sum_k P(X=\boldsymbol{x}|Y=c_k)P(Y=c_k)}$

- 代入上式：$P(Y=c_k \mid X=\boldsymbol{x})=\dfrac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}|Y=c_k)}$

朴素贝叶斯基本方法

- 贝叶斯分类器：

$$y = f(\boldsymbol{x}) = \arg\max_{c_k} \frac{P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)}{\sum_k P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)}$$

- 分母对所有c_k都相同：

$$y = \arg\max_{c_k} P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)} \mid Y = c_k)$$

后验概率最大化的含义：

- 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，假设选择$0-1$损失函数：$f(X)$为决策函数

$$L(Y,f(X))=\begin{cases}1, & Y\neq f(X)\\ 0, & Y=f(X)\end{cases}$$

- 期望风险函数：$R_{\exp}(f)=E[L(Y,f(X))]$
- 取条件期望：$R_{\exp}(f)=E_X\sum_{k=1}^{K}[L(c_k,f(X))]P(c_k\mid X)$

后验概率最大化的含义：

- 只需对$X=x$逐个极小化，得：

$$
\begin{aligned}
f(\boldsymbol{x}) &= \arg\min_{y\in\mathcal{Y}} \sum_{k=1}^{K} L(c_k, y) P(c_k \mid X=\boldsymbol{x}) \\
&= \arg\min_{y\in\mathcal{Y}} \sum_{k=1}^{K} P(y \neq c_k \mid X=\boldsymbol{x}) \\
&= \arg\min_{y\in\mathcal{Y}} (1 - P(y = c_k \mid X=\boldsymbol{x})) \\
&= \arg\max_{y\in\mathcal{Y}} P(y = c_k \mid X=\boldsymbol{x})
\end{aligned}
$$

- 推导出后验概率最大化准则：$f(x) = \arg\max_{c_k} P(c_k \mid X=\boldsymbol{x})$

朴素贝叶斯法的参数估计

应用极大似然估计法估计相应的概率

- 先验概率$P(Y=c_k)$的极大似然估计是：$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N}, k=1,2,\cdots,K$

- 设第j个特征$x^{(j)}$可能取值的集合为：$\{a_{j1},a_{j2},\cdots,a_{js_j}\}$

- 条件概率的极大似然估计：$P(X^{(j)}=a_{jl}\mid Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)}$

$$j=1,2,\cdots,n;\ \ l=1,2,\cdots,S_j;\ \ k=1,2,\cdots,K$$

朴素贝叶斯法的参数估计

朴素贝叶斯法：

- 输入：
  - 训练数据集$T = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_N, y_N)\}$
  - $x_i^{(j)}$第i个样本的第j个特征: $x_i = (x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)})^T$
  - a_{jl}第j个特征可能取的第l个值$x_i^{(j)} \in \left\{a_{j1}, a_{j2}, \cdots, a_{js_j}\right\}$

- 输出：
  - $\boldsymbol{x}$的分类$y_i \in \{c_1, c_2, \cdots, c_K\}$

朴素贝叶斯法的参数估计

步骤

1、计算先验概率和条件概率

$$P(Y=c_k)=\frac{\sum_{i=1}^{N} I(y_i=c_k)}{N}, k=1,2,\cdots,K$$

$$P(X^{(j)}=a_{jl} \mid Y=c_k)=\frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl}, y_i=c_k)}{\sum_{i=1}^{N} I(y_i=c_k)}$$

$$j=1,2,\cdots,n; \quad l=1,2,\cdots,S_j; \quad k=1,2,\cdots,K$$

朴素贝叶斯法的参数估计

步骤

2、对于给定的实例$\boldsymbol{x} = (x^{(1)}, x^{(2)}, \cdots, x^{(n)})^T$

计算$P(Y = c_k)\prod_{j=1}^{n} P(X^{(j)} = x^{(j)} \mid Y = c_k), k = 1,2,\cdots,K$

3、确定$\boldsymbol{x}$的类别

$$y = \arg\max_{c_k} P(Y = c_k)\prod_{j=1}^{n} P(X^{(j)} = x^{(j)} \mid Y = c_k)$$

例子

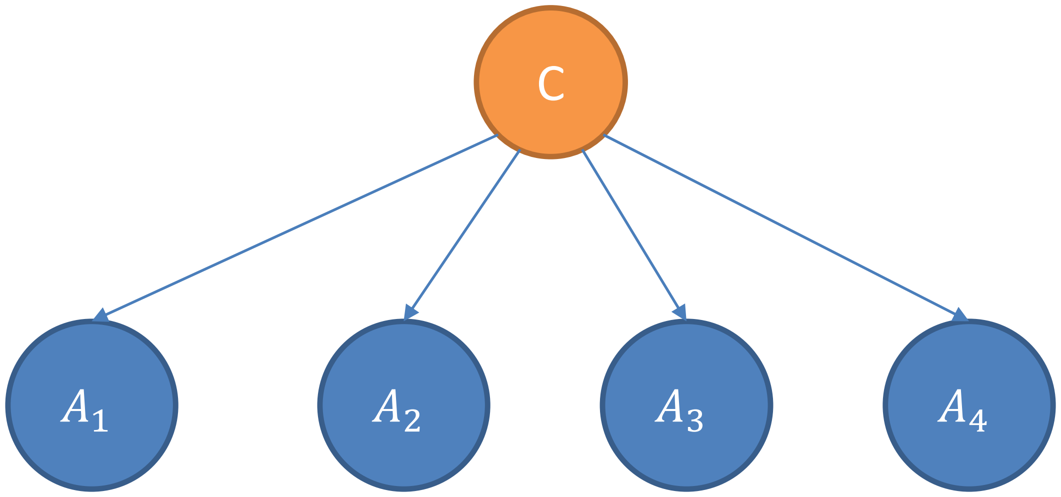
Day	Outlook	emperature	Humidity	Wind	PlayTennis
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
1	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild·	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes·
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

例子

测试

<Outlook=sunny, Temperature=cool, Humidity=high, Wind=strong>

$$c(x) = \underset{c \in \{yes, no\}}{\arg\max}\, P(c)P(\text{sunny} \mid c)P(\text{cool} \mid c)P(\text{high} \mid c)P(\text{strong} \mid c)$$



例子

$$P(yes) = (9+1)/(14+2) = 10/16$$
$$P(sunny \mid yes) = (2+1)/(9+3) = 3/12$$
$$P(cool \mid yes) = (3+1)/(9+3) = 4/12$$
$$P(high \mid yes) = (3+1)/(9+2) = 4/11$$
$$P(strong \mid yes) = (3+1)/(9+2) = 4/11$$

$$P(no) = (5+1)/(14+2) = 6/16$$
$$P(sunny \mid no) = (3+1)/(5+3) = 4/8$$
$$P(cool \mid no) = (1+1)/(5+3) = 2/8$$
$$P(high \mid no) = (4+1)/(5+2) = 5/7$$
$$P(strong \mid no) = (3+1)/(5+2) = 4/7$$

$$P(yes)P(sunny \mid yes)P(cool \mid yes)P(high \mid yes)P(strong \mid yes) = 0.0069$$
$$P(no)P(sunny \mid no)P(cool \mid no)P(high \mid no)P(strong \mid no) = 0.0191$$

贝叶斯估计

考虑用极大似然估计可能会出现所要估计的概率值为0的情况，这时会影响到后验概率的计算结果，使分类产生偏差。解决这一问题的方法是采用贝叶斯估计。

- 条件概率的贝叶斯估计： $P_\lambda(X^{(j)} = a_{jl} \mid Y = c_k) = \dfrac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl}, y_i=c_k)+\lambda}{\sum_{i=1}^{N} I(y_i=c_k)+S_j\lambda}$

- 先验概率的贝叶斯估计： $P_\lambda(Y = c_k) = \dfrac{\sum_{i=1}^{N} I(y_i=c_k)+\lambda}{N+K\lambda}$

谢谢各位同学！