第3节　朴素贝叶斯分类

—— 贝叶斯理论的应用

3.1 朴素贝叶斯分类

朴素贝叶斯的思想基础是这样的：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。

如：对正常人和病人的识别问题。

假设每个要识别的人有 d 个基本特性的特征，如身高、体重、温度、脉搏……等，从而组成一个 d 维空间的向量 $x=(x_1, x_2, \ldots, x_d)$，识别病人就是要将样本 x 分类成正常人或病人。

朴素贝叶斯分类步骤：

1、如果用 ω 表示人的健康状态，则：

$\omega = \omega_1$ 表示正常人

$\omega = \omega_2$ 表示病人

2、计算 $p(\omega_1 | x)$, $p(\omega_2 | x)$

3、如果 $p(\omega_i | x) = \max\{p(\omega_1 | x), p(\omega_2 | x)\}$，

则 $x \in \omega_i$，$i = 1, 2$。

3.2 两种朴素贝叶斯分类策略

3.2.1 最小错误率的Bayes决策

利用Bayes公式，在模式分类中尽量减少分类的错误，这种分类策略称为最小错误率的Bayes决策。

健康状态分类中，类别状态变量 ω 是一个随机变量，但其概率分布是已知的，即 $p(\omega_1)$、$p(\omega_2)$ 是已知的先验概率，并有： $p(\omega_1)+p(\omega_2)=1$

假设只选择一个特征即温度进行分类，有$d=1$，样本数据的类别条件概率分布如图3.1所示。

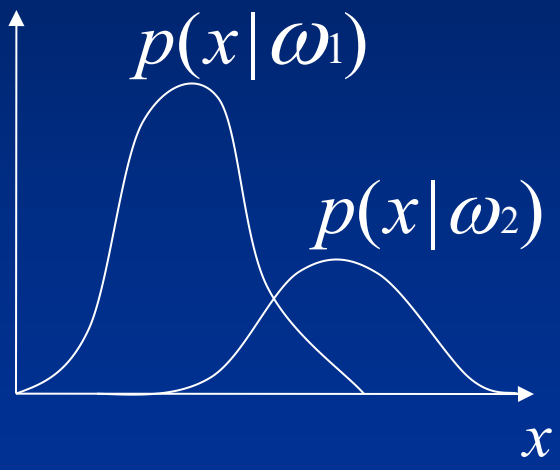

图3.1 样本的条件概率密度

其中，$p(x|\omega_1)$是正常人样本的条件概率；

$p(x|\omega_2)$是病人样本的条件概率密度。

利用Bayes公式，则有状态的后验概率：

$$p(\omega_i|x)=\frac{p(x|\omega_i)\cdot p(\omega_i)}{\sum_{i=1}^{2}p(x|\omega_i)\cdot p(\omega_i)}$$

其结果如图3.2所示：

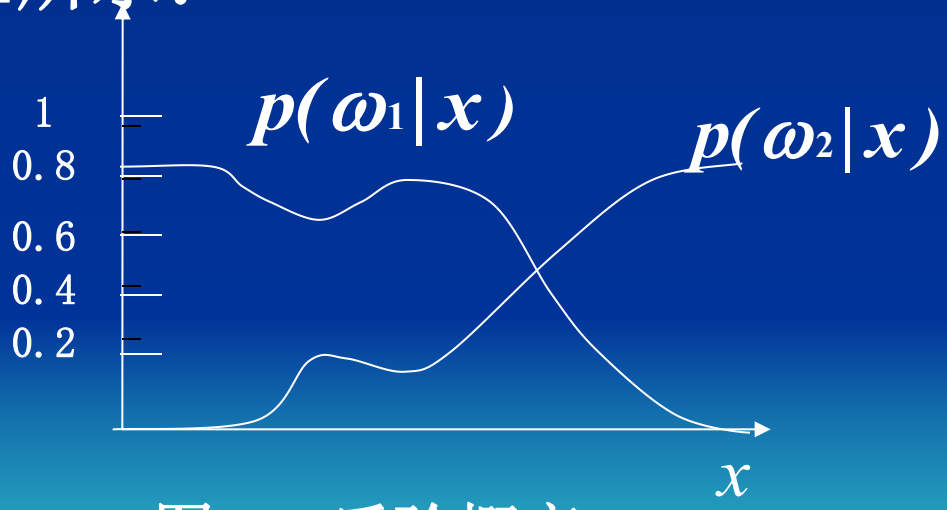

图3.2 后验概率

采用最小错误率进行Bayes决策：

如果 $p(\omega_1 | x) > p(\omega_2 | x)$，则把测试样本 x 归类于 ω_1；反之，归类于 ω_2。

则决策规则可简写为：

$$p(\omega_i | x) = \max_{j=1,2} p(\omega_j | x), \text{则} x \in \omega_i$$

最小错误率分析

(1) 错误率

错误率是指平均错误率，以 $p(e)$ 表示，有定义：

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

其中，$\int_{-\infty}^{+\infty}(\ )dx$ 表示在整个 d 维空间上积分。

(2) 条件错误概率：

根据Bayes决策规则：

当 $p(\omega_1|x) > p(\omega_2|x)$ 时，决策结果为 ω_1；

当 $p(\omega_1|x) < p(\omega_2|x)$ 时，决策结果为 ω_2。

显然，在作出决策 ω_1 时，条件错误概率 $p(e|x)$ 为：

$$p(\omega_2|x) \qquad x \in \omega_2$$

在作出决策 ω_2 时，条件错误概率 $p(e|x)$ 为：

$$p(\omega_1|x) \qquad x \in \omega_1$$

从而有条件错误概率：

$$p(e\mid x)=\begin{cases} p(\omega_2\mid x) & \text{当} p(\omega_1\mid x)>p(\omega_2\mid x)\text{时} \\ p(\omega_1\mid x) & \text{当} p(\omega_1\mid x)<p(\omega_2\mid x)\text{时} \end{cases}$$

$$=\int_{-\infty}^{t} p(\omega_2\mid x)\cdot p(x)dx+\int_{t}^{+\infty} p(\omega_1\mid x)\cdot p(x)dx$$

$$=\int_{-\infty}^{t} p(x\mid\omega_2)\cdot p(\omega_2)dx+\int_{t}^{+\infty} p(x\mid\omega_1)\cdot p(\omega_1)dx$$

$$=p(\omega_2)\cdot\int_{-\infty}^{t} p(x\mid\omega_2)dx+p(\omega_1)\cdot\int_{t}^{+\infty} p(x\mid\omega_1)dx$$

$$=p(\omega_2)\cdot p(e_2)+p(\omega_1)\cdot p(e_1)$$

图3.3给出了一维条件下错误率 $p(e|x)$ 分布图示。以上讨论不难推广到 n 维特征空间的情况。

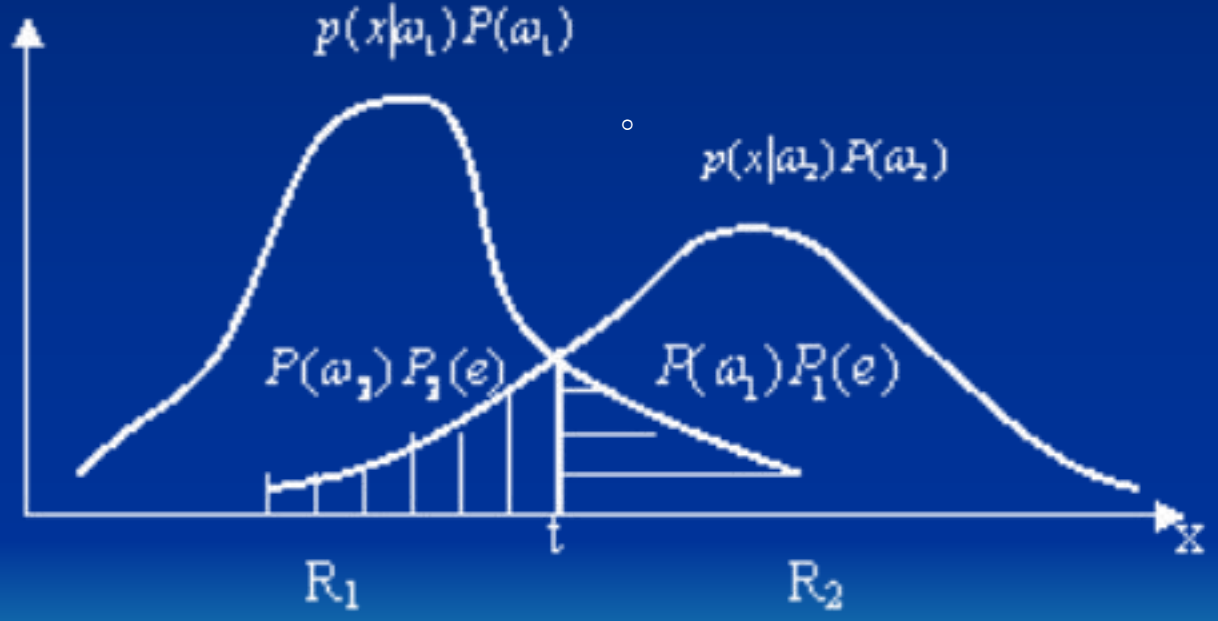

图3.3 分类错误率的条件概率密度

(3) 最小错误率Bayes决策的本质

实际上是对每个 $p(e|x)$ 都取最小者，从而积分：

$$\int_{-\infty}^{+\infty} p(e|x)\cdot p(x)dx$$

也必然达到最小。也即 $p(e)$ 达到最小。

图3.3说明最小错误率Bayes决策规则确实使错误率最小。这种方法可以推广到多类的分类决策之中。

例1：

- 假设在某个局部地区细胞识别中正常ω_1和异常ω_2两类的先验概率分别为：
  - 正常状态：$p(\omega_1) = 0.9$
  - 异常状态：$p(\omega_2) = 0.1$

- 现有一待识别的细胞，根据其观察值 x，从条件概率密度分布曲线上查得

$$p(x \mid \omega_1) = 0.2, \quad p(x \mid \omega_2) = 0.4$$

试对该细胞 x 进行分类。

解：

利用贝叶斯公式，分别计算出其分属ω_1及ω_2类别的后验概率：

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

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

根据最小错误贝叶斯决策，

$$p(\omega_1 \mid x) = 0.818 > p(\omega_2 \mid x) = 0.182$$

有：合理的分类应是将 x 归入正常状态。

3.2.2 最小风险的Bayes决策

在决策中有时需要考虑一个比错误率更广泛的概念——风险。对事物的分类，第一考虑的是尽可能正确的分类。第二要考虑错误判断分类带来的后果。

例如，把健康人归类为病人，或把病人归类为健康人。这会带来精神压力或延误病情，即会引起损失。

最小风险的Bayes决策是考虑各种错误造成损失不同而提出的决策，需要考虑损失函数。

1. 基本概念

(1) 决策空间 A

假设决策空间 A 由 α 个决策组成：

$$A = \{ a_1, a_2, \ldots\ldots, a_\alpha \}$$

按一般理解，如果有 C 个类别，则 C 个决策即够了。但实际上，对于 C 个类别有 C 种不同决策之外，还允许有其他决策，例如“拒绝”决策，则这时就有：

$$\alpha = C + 1$$

(2) 损失函数

损失函数 λ 表示为：

$$\lambda(a_i, \omega_j) \quad i=1, 2, \ldots, \alpha; \quad j=1, 2, \ldots, C$$

其中，$\lambda(a_i, \omega_j)$ 表示为状态为 ω_j，而决策为 a_i 时所带来的损失。根据状态 ω_j，损失函数 λ 和决策 a_i 则可得一般决策表，如表3.1所示。

表3.1 损失决策表

损失 / 状态 / 决策	自然状态			
	ω_1	ω_2	·····	ω_c
a_1	$\lambda(a_1,\omega_1)$	$\lambda(a_1,\omega_2)$	·····	$\lambda(a_1,\omega_c)$
a_2	$\lambda(a_2,\omega_1)$	$\lambda(a_2,\omega_2)$	·····	$\lambda(a_2,\omega_c)$
⋮	⋮	⋮	⋮	⋮
a_α	$\lambda(a_\alpha,\omega_1)$	$\lambda(a_\alpha,\omega_2)$	·····	$\lambda(a_\alpha,\omega_c)$

2. 最小风险Bayes决策

(1) 已知先验概率： $p(\omega_j)$, 以及样本的条件概率密度： $p(x \mid \omega_j)$, $j=1, 2, \ldots, C$

(2) 求后验概率：

由Bayes公式求出：

$$p(\omega_j \mid x)=\frac{p(x \mid \omega_j)\cdot p(\omega_j)}{\sum_{j=1}^{C} p(x \mid \omega_j)\cdot p(\omega_j)}$$

(3) 条件损失 (条件风险)

综合损失函数$\lambda(a_i, \omega_j)$和后验概率 $p(\omega_j | x)$，得到风险最小的决策。

在决策 a_i 情况下的条件损失为：

$$R(a_i | x) = \sum_{j=1}^{C} \lambda(a_i, \omega_j) \cdot p(\omega_j | x)$$

条件损失也称为条件风险，反映了采用某一决策 a_i 所带来的风险。

(4) 最小风险Bayes决策规则

最小风险Bayes决策规则表示为：

$$R(a_k|x)=\min_{i=1,2,\ldots,\alpha} R(a_i|x)$$

在求实际问题决策时，最小风险Bayes决策步骤如下：

i）以Bayes公式求取样本 x 的后验概率 $p(\omega_j|x)$ $j=1,2,\ldots,C$

ii) 综合后验概率及损失决策表，按条件损失 $R(a_i|x)$ 的式子：

$$R(a_i|x)=\sum_{j=1}^{C}\lambda(a_i,\omega_j)\cdot p(\omega_j|x)$$

求取条件风险。

iii) 对求出的α个条件风险值 $R(a_i|x)$, i=1, 2, …, α 进行比较，找出使条件风险最小的决策a_k，即

$$R(a_k|x)=\min_{i=1,2,\ldots,\alpha}R(a_i|x)$$

则 a_k 就是最小风险Bayes决策。

例2：

假设在某个地区人们的健康状态 ω_1 和患病状态 ω_2 的先验概率分别为：

健康状态：$p(\omega_1) = 0.9$

患病状态：$p(\omega_2) = 0.1$

现有一个待识别的个体，其观察值为 x，从条件概率密度分布曲线上查得：

$$p(x \mid \omega_1) = 0.2$$

$$p(x \mid \omega_2) = 0.4$$

试按最小风险Bayes决策进行分类。

解：(1) 损失决策表如下表所示：

损失		状态	
		ω_1	ω_2
决策	a_1	$\lambda_{11}=0$	$\lambda_{12}=6$
	a_2	$\lambda_{21}=1$	$\lambda_{22}=0$

(2) 用Bayes公式，分别求后验概率：

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

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

(3) 根据条件损失公式求出条件风险

$$R(a_1|x)=\lambda_{11}\times p(\omega_1|x)+\lambda_{12}\times p(\omega_2|x)=0+6\times 0.182$$
$$=1.092$$

$$R(a_2|x)=\lambda_{21}\times p(\omega_1|x)+\lambda_{22}\times p(\omega_2|x)=1\times 0.818+0$$
$$=0.818$$

(4) 最小风险决策

由于 $R(a_1|x)>R(a_2|x)$，说明决策 a_2 的风险小于决策 a_1，故取a_2 为决策。

最后识别的结果为：样本 x 属于患病状态 ω_2 类。

3.2.3 最小错误率与最小风险Bayes决策的关系

定义：如果损失函数取下面形式

$$\lambda(a_i,\omega_j)=\begin{cases}0, & \text{当} i=j\text{时}\\ 1, & \text{当} i\neq j\text{时}\end{cases}\qquad i,j=1,2,...,C$$

则称其为 0 - 1 损失函数。

即对于正确决策 (当 $i=j$ 时) 没有损失，而对任何错误决策其损失为1。

用0-1 损失函数时的最小风险Bayes决策：

$$R(a_i \mid x) = \sum_{\substack{j=1 \\ i \neq j}}^{C} p(\omega_j \mid x)$$

则最小风险Bayes决策为：

$$R(a_k \mid x) = \min_{i=1,2,\ldots,\alpha} R(a_i \mid x)$$

即最小错误率的Bayes决策是在0-1 损失函数下的最小风险的Bayes决策。

3.3 朴素贝叶斯分类实例

- 对于SNS社区来说，不真实的账号（使用虚假身份或用户的小号）是一个普遍存在的问题，作为SNS社区的运营商，希望可以检测出这些不真实账号，从而在一些运营分析报告中避免这些账号的干扰，亦可以加强对SNS社区的了解与监管。

- 首先设$C=0$表示真实账号，$C=1$表示账号不真实。
- 接着，选择样本x的三个特征属性：

x_1：日志数量/注册天数；

x_2：好友数量/注册天数；

x_3：是否使用真实头像。

在SNS社区中这三项都是可以直接从数据库里得到或计算出来的。

- 并给出属性值的划分：

  x_1：$\{x_1 \le 0.05,\ 0.05 < x_1 < 0.2,\ x_1 \ge 0.2\}$；

  x_2：$\{x_2 \le 0.1,\ 0.1 < x_2 < 0.8,\ x_2 \ge 0.8\}$；

  x_3：$\{x_3 = 0$ (不是)，$x_3 = 1$ (是)$\}$。

- **获取训练样本：**使用运维人员曾经人工检测过的1万个账号作为训练样本。
- **计算样本每个类别的先验概率：**用样本中真实账号和不真实账号数量分别除以1万，得到：

$$p(C=0)=8900/10000=0.89$$

$$p(C=1)=1100/10000=0.11$$

计算样本 x 各个分量在不同类别下的条件概率：

$$p(x_1 \le 0.05 \mid C = 0) = 0.3$$

$$p(0.05 < x_1 < 0.2 \mid C = 0) = 0.5$$

$$p(x_1 \ge 0.2 \mid C = 0) = 0.2$$

$$p(x_1 \le 0.05 \mid C = 1) = 0.8$$

$$p(0.05 < x_1 < 0.2 \mid C = 1) = 0.1$$

$$p(x_1 \ge 0.2 \mid C = 1) = 0.1$$

$$p(x_2 \le 0.1 \mid C = 0) = 0.1$$
$$p(0.1 < x_2 < 0.8 \mid C = 0) = 0.7$$
$$p(x_2 \ge 0.8 \mid C = 0) = 0.2$$
$$p(x_2 \le 0.1 \mid C = 1) = 0.7$$
$$p(0.1 < x_2 < 0.8 \mid C = 1) = 0.2$$
$$p(x_2 \ge 0.8 \mid C = 1) = 0.1$$

$$p(x_3 = 0 \mid C = 0) = 0.2$$
$$p(x_3 = 1 \mid C = 0) = 0.8$$
$$p(x_3 = 0 \mid C = 1) = 0.9$$
$$p(x_3 = 1 \mid C = 1) = 0.1$$

假定现在有一个账号 x，使用非真实头像(即x_3= 0)，日志数量与注册天数的比率为0.1(即 x_1= 0.1)，好友数与注册天数的比率为0.2(即 x_2= 0.2)。得到样本 x = {0.1, 0.2, 0}。

则 x 的三个分量取值分别位于以下区间：

$$0.05 < x_1 < 0.2$$

$$0.1 < x_2 < 0.8$$

$$x_3 = 0$$

$$p(C=0)\cdot p(x\mid C=0)=0.89\times 0.5\times 0.7\times 0.2=0.0623$$

$$p(C=1)\cdot p(x\mid C=1)=0.11\times 0.1\times 0.2\times 0.9=0.00198$$

由于　$p(C=0)\cdot p(x|C=0) > p(C=1)\cdot p(x|C=1)$

最后决策：　$x \in \{ C=0 \}$

可以看到，虽然这个用户没有使用真实头像，但是通过分类器的鉴别，更倾向于将此账号归入真实账号类别。

SNS社区的朴素贝叶斯分类解决方案中，做了如下假设：

1. 真实账号比非真实账号平均具有更大的日志密度、各大的好友密度以及更多的使用真实头像。

2. 日志密度、好友密度和是否使用真实头像在账号真实性给定的条件下是独立的。

- 但是，上述第二条假设很可能并不成立。一般来说，好友密度除了与账号是否真实有关，还与是否有真实头像有关，因为真实的头像会吸引更多人加其为好友。

为了获取更准确的分类，可以将假设修改如下：

- 真实账号比非真实账号平均具有更大的日志密度、各大的好友密度以及更多的使用真实头像。
- 日志密度与好友密度、日志密度与是否使用真实头像在账号真实性给定的条件下是独立的。
- 使用真实头像的用户比使用非真实头像的用户平均有更大的好友密度。

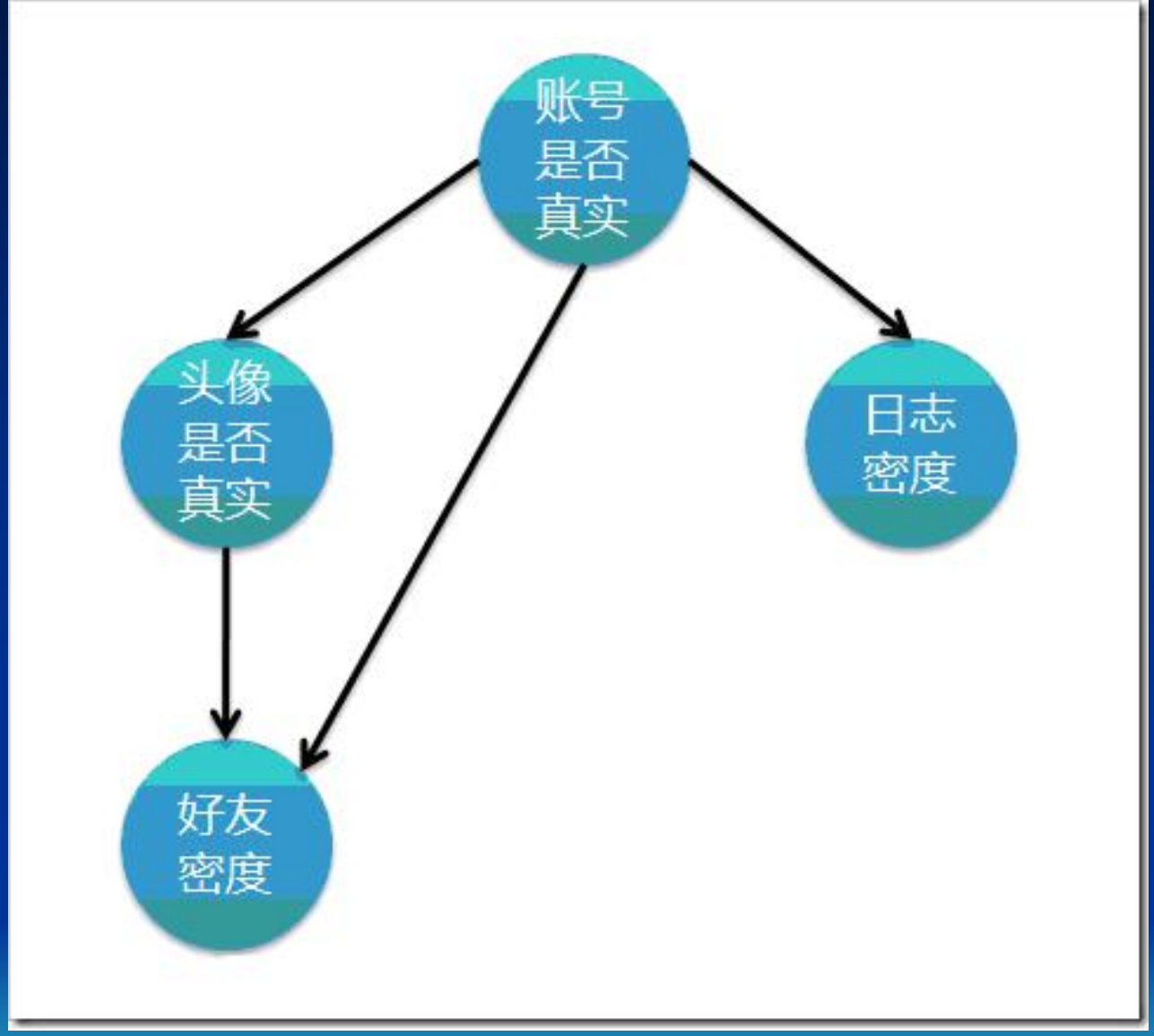

图3.4 特征属性之间的关联关系

朴素贝叶斯分类的局限性

- 朴素贝叶斯分类有一个限制条件，就是特征属性必须有条件独立或基本独立（实际上在现实应用中几乎不可能做到完全独立）。当这个条件成立时，朴素贝叶斯分类法的准确率是最高的，但不幸的是，现实中各个特征属性间往往并不条件独立，而是具有较强的相关性，这样就限制了朴素贝叶斯分类的能力。