第 5 章　贝叶斯分类算法

5.1　基本概念

5.1.1　主观概率

贝叶斯算法是一种研究不确定性的推理方法。不确定性常用贝叶斯概率表示，它是一种主观概率。通常的经典概率代表事件的物理特性，是不随人意识变化的客观存在。而贝叶斯概率则是人的认识，是个人主观的估计，随个人主观认识的变化而变化。例如，事件的贝叶斯概率只指个人对该事件的置信程度，因此是一种主观概率。

投掷硬币可能出现正反面两种情形，经典概率代表硬币正面朝上的概率，这是一个客观存在；而贝叶斯概率则指个人相信硬币会正面朝上的程度。

同样的例子还有，一个企业家认为“一项新产品在未来市场上销售”的概率是 0.8，这里的 0.8 是根据他多年的经验和当时的一些市场信息综合而成的个人信念。

一个投资者认为“购买某种股票能获得高收益”的概率是 0.6，这里的 0.6 是投资者根据自己多年股票生意经验和当时股票行情综合而成的个人信念。

贝叶斯概率是主观的，对其估计取决于先验知识的正确性和后验知识的丰富和准确度。因此贝叶斯概率常常可能随个人掌握信息的不同而发生变化。

对即将进行的羽毛球单打比赛结果进行预测，不同人对胜负的主观预测都不同。如果对两人的情况和各种现场的分析一无所知，就会认为二者的胜负比例为 1∶1；如果知道其中一人为奥运会羽毛球单打冠军，而另一人只是某省队新队员，则可能给出的概率是奥运会冠军和省队队员的胜负比例为 3∶1；如果进一步知道奥运会冠军刚好在前一场比赛中受过伤，则对他们胜负比例的主观预测可能会下调为 2∶1。所有的预测推断都是主观的，基于后验知识的一种判断，取决于对各种信息的掌握。

经典概率方法强调客观存在，它认为不确定性是客观存在的。在同样的羽毛球单打比赛预测中，从经典概率的角度看，如果认为胜负比例为 1∶1，则意味着在相同的条件下，如果两人进行 100 场比赛，其中一人可能会取得 50 场的胜利，同时丢掉另外 50 场。

主观概率不像经典概率那样强调多次重复，因此在许多不可能出现重复事件的场合能得到很好的应用。上面提到的企业家对未来产品的预测，投资者对股票是否能取得高收益的预测以及羽毛球比赛胜负的预测中，都不可能进行重复的实验，因此，利用主观概率，按照个人对事件的相信程度而对事件做出推断是一种很合理且易于解释的方法。

5.1.2 贝叶斯定理

1. 基础知识

(1) 已知事件 A 发生的条件下，事件 B 发生的概率，叫作事件 B 在事件 A 发生下的条件概率，记为 $P(B|A)$，其中 $P(A)$ 叫作先验概率，$P(B|A)$ 叫作后验概率，计算条件概率的公式为

$$P(B|A)=\frac{P(A\cap B)}{P(A)} \tag{5-1}$$

条件概率公式通过变形得到乘法公式为

$$P(A\cap B)=P(B|A)P(A) \tag{5-2}$$

(2) 设 A，B 为两个随机事件，如果 $P(AB)=P(A)P(B)$ 成立，则称事件 A 和 B 相互独立。此时有 $P(A|B)=P(A)$，$P(AB)=P(A)P(B)$ 成立。

设 $A_1,A_2,\cdots,A_n$ 为 n 个随机事件，如果对其中任意 m $(2\leqslant m\leqslant n)$ 个事件 A_{k_1}，$A_{k_2},\cdots,A_{k_m}$，都有

$$P(A_{k_1},A_{k_2},\cdots,A_{k_m})=P(A_{k_1})P(A_{k_2})\cdots P(A_{k_m}) \tag{5-3}$$

成立，则称事件 $A_1,A_2,\cdots,A_n$ 相互独立。

(3) 设 $B_1,B_2,\cdots,B_n$ 为互不相容事件，$P(B_i)>0$，$i=1,2,\cdots,n$，且 $\bigcup_{i=1}^{n}B_i=\Omega$，对任意的事件 $A\subset\bigcup_{i=1}^{n}B_i$，计算事件 A 概率的公式为

$$P(A)=\sum_{i=1}^{n}P(B_i)P(A|B_i) \tag{5-4}$$

设 $B_1,B_2,\cdots,B_n$ 为互不相容事件，$P(B_i)>0$，$i=1,2,\cdots,n$，$P(A)>0$，则在事件 A 发生的条件下，事件 B_i 发生的概率为

$$P(B_i|A)=\frac{P(B_iA)}{P(A)}=\frac{P(B_i)P(A|B_i)}{\sum_{i=1}^{n}P(B_i)P(A|B_i)} \tag{5-5}$$

则称该公式为贝叶斯公式。

2. 贝叶斯决策准则

假设 $\Omega=\{C_1,C_2,\cdots,C_m\}$ 是有 m 个不同类别的集合，特征向量 $\boldsymbol{X}$ 是 d 维向量，$P(\boldsymbol{X}|C_i)$ 是特征向量 $\boldsymbol{X}$ 在类别 C_i 状态下的条件概率，$P(C_i)$ 为类别 C_i 的先验概率。根据前面所述的贝叶斯公式，后验概率 $P(C_i|\boldsymbol{X})$ 的计算公式为

$$P(C_i|\boldsymbol{X})=\frac{P(\boldsymbol{X}|C_i)}{P(\boldsymbol{X})}P(C_i) \tag{5-6}$$

其中，$P(\boldsymbol{X})=\sum_{j=1}^{m}P(\boldsymbol{X}|C_j)P(C_j)$。

贝叶斯决策准则：如果对于任意 $i\neq j$，都有 $P(C_i|\boldsymbol{X})>P(C_j|\boldsymbol{X})$ 成立，则样本模式

$\boldsymbol{X}$ 被判定为类别 C_i。

3. 极大后验假设

根据贝叶斯公式可得到一种计算后验概率的方法：在一定假设的条件下，根据先验概率和统计样本数据得到的概率，可以得到后验概率。

令 $P(c)$ 是假设 c 的先验概率，它表示 c 是正确假设的概率，$P(\boldsymbol{X})$ 表示的是训练样本 $\boldsymbol{X}$ 的先验概率，$P(\boldsymbol{X}|c)$ 表示在假设 c 正确的条件下样本 $\boldsymbol{X}$ 发生或出现的概率，根据贝叶斯公式可以得到后验概率的计算公式为

$$P(c|\boldsymbol{X})=\frac{P(\boldsymbol{X}|c)P(c)}{P(\boldsymbol{X})} \tag{5-7}$$

设 C 为类别集合，也就是待选假设集合，在给定未知类别标号样本 $\boldsymbol{X}$ 时，通过计算找到可能性最大的假设 $c\in C$，具有最大可能性的假设或类别被称为极大后验假设(maximum a posteriori)，记作 c_{map}。

$$c_{\text{map}}=\underset{c\in C}{\operatorname{argmax}}P(c|\boldsymbol{X})=\underset{c\in C}{\operatorname{argmax}}\frac{P(\boldsymbol{X}|c)P(c)}{P(\boldsymbol{X})} \tag{5-8}$$

由于 $P(\boldsymbol{X})$ 与假设 c 无关，故上式可变为

$$c_{\text{map}}=\underset{c\in C}{\operatorname{argmax}}P(\boldsymbol{X}|c)P(c) \tag{5-9}$$

当没有给定类别概率的情形下，可做一个简单的假定。假设 C 中每个假设都有相等的先验概率，也就是对于任意的 $c_i,c_j\in C(i\neq j)$，都有 $P(c_i)=P(c_j)$，再做进一步简化，只需计算 $P(\boldsymbol{X}|c)$ 找到使之达到最大的假设。$P(\boldsymbol{X}|c)$ 被称为极大似然假设(maximum likelihood)，记为 c_{ml}。

$$c_{\text{ml}}=\underset{c\in C}{\operatorname{argmax}}P(\boldsymbol{X}|c) \tag{5-10}$$

5.2 贝叶斯分类算法原理

5.2.1 朴素贝叶斯分类模型

贝叶斯分类器诸多算法中朴素贝叶斯分类模型是最早的。它的算法逻辑简单，构造的朴素贝叶斯分类模型结构也比较简单，运算速度比同类算法快很多，分类所需的时间也比较短，并且大多数情况下分类精度也比较高，因而在实际中得到了广泛的应用。该分类器有一个朴素的假定：以属性的类条件独立性假设为前提，即在给定类别状态的条件下，属性之间是相互独立的。朴素贝叶斯分类器的结构示意图如图 5-1 所示。

假设样本空间有 m 个类别 $\{C_1,C_2,\cdots,C_m\}$，数据集有 n 个属性 $A_1,A_2,\cdots,A_n$，给定一未知类别的样本 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$，其中 x_i 表示第 i 个属性的取值，即 $x_i\in A_i$，则可用贝叶斯公式计算样本 $\boldsymbol{X}=(x_1,x_2,\cdots,x_n)$ 属于类别 C_k $(1\leqslant k\leqslant m)$ 的概率。由贝叶斯公式，有 $P(C_k|\boldsymbol{X})=\dfrac{P(C_k)P(\boldsymbol{X}|C_k)}{P(\boldsymbol{X})}\propto P(C_k)P(\boldsymbol{X}|C_k)$，即要得到 $P(C_k|\boldsymbol{X})$ 的值，关键是要计算 $P(\boldsymbol{X}|C_k)$ 和 $P(C_k)$。令 $C(\boldsymbol{X})$ 为 $\boldsymbol{X}$ 所属的类别标签，由贝叶斯分类准则，如

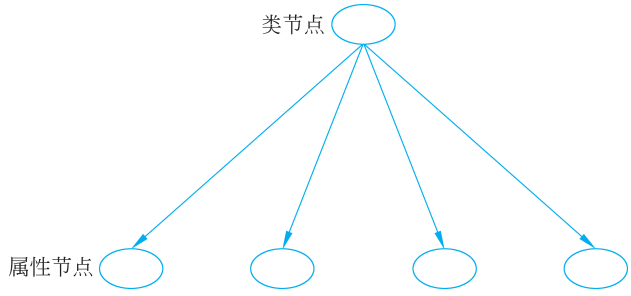

图 5-1 朴素贝叶斯分类器的结构示意图

果对于任意 $i \neq j$ 都有 $P(C_i|\boldsymbol{X}) > P(C_j|\boldsymbol{X})$ 成立，则把未知类别的样本 $\boldsymbol{X}$ 指派给类别 C_i，贝叶斯分类器的计算模型为

$$V(\boldsymbol{X}) = \operatorname{argmax} P(C_i) P(\boldsymbol{X}|C_i) \tag{5-11}$$

由朴素贝叶斯分类器的属性独立性假设，假设各属性 x_i $(i=1,2,\cdots,n)$ 间相互类条件独立，则

$$P(\boldsymbol{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i) \tag{5-12}$$

于是式(5-11)被修改为

$$V(\boldsymbol{X}) = \operatorname*{argmax}_{i} P(C_i) \prod_{k=1}^{n} P(x_k|C_i) \tag{5-13}$$

$P(C_i)$ 为先验概率，可通过 $P(C_i) = d_i/d$ 计算得到。其中，d_i 是属于类别 C_i 的训练样本的个数，d 是训练样本的总数。若属性 A_k 是离散的，则概率可由 $P(x_k|C_i) = d_{ik}/d_i$ 计算得到。其中，d_{ik} 是训练样本集合中属于类 C_i 并且属性 A_k 取值为 x_k 的样本个数，d_i 是属于类 C_i 的训练样本个数。朴素贝叶斯分类的工作过程如下。

(1) 用一个 n 维特征向量 $\boldsymbol{X} = (x_1, x_2, \cdots, x_n)$ 来表示数据样本，描述样本 $\boldsymbol{X}$ 对 n 个属性 $A_1, A_2, \cdots, A_n$ 的量度。

(2) 假定样本空间有 m 个类别状态 $C_1, C_2, \cdots, C_m$，对于给定的一个未知类别标号的数据样本 $\boldsymbol{X}$，分类算法将 $\boldsymbol{X}$ 判定为具有最高后验概率的类别，也就是说，朴素贝叶斯分类算法将未知类别的样本 $\boldsymbol{X}$ 分配给类别 C_i，当且仅当对于任意的 j，始终有 $P(C_i|\boldsymbol{X}) > P(C_j|\boldsymbol{X})$ 成立，$1 \leqslant i \leqslant m$，$1 \leqslant j \leqslant m$，$j \neq i$。使 $P(C_i|\boldsymbol{X})$ 取得最大值的类别 C_i 被称为最大后验假定。

(3) 由于 $P(\boldsymbol{X})$ 不依赖类别状态，对于所有类别都是常数，故根据贝叶斯定理，最大化 $P(C_i|\boldsymbol{X})$ 只需要最大化 $P(\boldsymbol{X}|C_i)P(C_i)$ 即可。如果类的先验概率未知，则通常假设这些类别的概率是相等的，即 $P(C_1) = P(C_2) = \cdots = P(C_m)$，所以只需要最大化 $P(\boldsymbol{X}|C_i)$ 即可，否则就要最大化 $P(\boldsymbol{X}|C_i)P(C_i)$。其中，可用频率 S_i/S 对 $P(C_i)$ 进行估计计算，S_i 是给定类别 C_i 中训练样本的个数，S 是训练样本(实例空间)的总数。

(4) 当实例空间中训练样本的属性较多时，计算 $P(\boldsymbol{X}|C_i)$ 可能会比较费时，开销较大，此时可以做类条件独立性的假定：在给定样本类别标号的条件下，假定属性值是相互条件独立的，属性之间不存在任何依赖关系，则下面等式成立：$P(\boldsymbol{X}|C_i) = \prod_{k=1}^{n} P(x_k|C_i)$。

其中,概率 $P(x_1|C_i)$,$P(x_2|C_i)$,…,$P(x_n|C_i)$ 的计算可由样本空间中的训练样本进行估计。实际问题中根据样本属性 A_k 的离散连续性质,考虑下面两种情形。

- 如果属性 A_k 是连续的,则一般假定它服从正态分布,从而计算类条件概率。
- 如果属性 A_k 是离散的,则 $P(x_k|C_i)=S_{ik}/S_i$。其中,S_{ik} 是在实例空间中类别为 C_i 的样本中属性 A_k 上取值为 x_k 的训练样本个数,而 S_i 是属于类别 C_i 的训练样本个数。

(5) 对于未知类别的样本 $\boldsymbol{X}$,对每个类别 C_i 分别计算 $P(\boldsymbol{X}|C_i)P(C_i)$。样本 $\boldsymbol{X}$ 被认为属于类别 C_i,当且仅当 $P(\boldsymbol{X}|C_i)P(C_i)>P(\boldsymbol{X}|C_j)P(C_j)$,$1\leqslant i\leqslant m$,$1\leqslant j\leqslant m$,$j\neq i$,也就是说,样本 $\boldsymbol{X}$ 被指派到使 $P(\boldsymbol{X}|C_i)P(C_i)$ 取得最大值的类别 C_i。

朴素贝叶斯分类模型的算法描述如下。

(1) 对训练样本数据集和测试样本数据集进行离散化处理和缺失值处理。

(2) 扫描训练样本数据集,分别统计训练集中类别 C_i 的个数 d_i 和属于类别 C_i 的样本中属性 A_k 取值为 x_k 的实例样本个数 d_{ik},构成统计表。

(3) 计算先验概率 $P(C_i)=d_i/d$ 和条件概率 $P(A_k=x_k|C_i)=d_{ik}/d_i$,构成概率表。

(4) 构建分类模型 $V(\boldsymbol{X})=\underset{i}{\operatorname{argmax}}P(C_i)P(\boldsymbol{X}|C_i)$。

(5) 扫描待分类的样本数据集,调用已得到的统计表、概率表以及构建好的分类准则,得出分类结果。

5.2.2 贝叶斯信念网络

朴素贝叶斯分类器的条件独立假设似乎太严格了,特别是对那些属性之间有一定相关性的分类问题。下面介绍一种更灵活的类条件概率 $P(X|Y)$ 的建模方法。该方法不要求给定类的所有属性条件独立,而是允许指定哪些属性条件独立。

1. 模型表示

贝叶斯信念网络(Bayesian Belief Network,BBN),简称贝叶斯网络,用图形表示一组随机变量之间的概率关系。贝叶斯网络有以下两个主要成分。

(1) 一个有向无环图(Directed Acyclic Graph,DAG),表示变量之间的依赖关系。

(2) 一个概率表,把各节点和它的直接父节点关联起来。

考虑3个随机变量 A、B 和 C,其中 A 和 B 相互独立,并且都直接影响第3个变量 C。3个变量之间的关系可以用图5-2(a)中的有向无环图概括。图中每个节点表示一个变量,每条弧表示变量之间的依赖关系。如果从 X 到 Y 有一条有向弧,则 X 是 Y 的父母,Y 是 X 的子女。另外,如果网络中存在一条从 X 到 Z 的有向路径,则 X 是 Z 的祖先,而 Z 是 X 的后代。例如,在图5-2(b)中,A 是 D 的后代,D 是 B 的祖先,而且 B 和 D 都不是 A 的后代节点。贝叶斯网络的重要性质:贝叶斯网络中的一个节点,如果它的父节点已知,则它条件独立于其所有的非后代节点。图5-2(b)中给定 C,A 条件独立于 B 和 D,

因为 B 和 D 都是 A 的非后代节点。朴素贝叶斯分类器中的条件独立假设也可以用贝叶斯网络来表示。如图 5-2(c)所示，其中 Y 是目标类，$\{X_1, X_2, \cdots, X_5\}$ 是属性集。

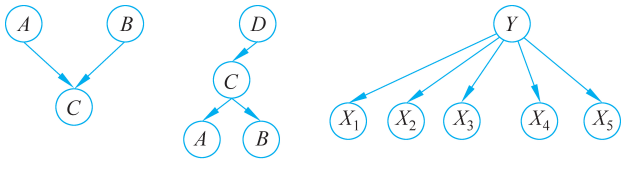

图 5-2 贝叶斯网络

在贝叶斯网络中，除了网络拓扑结构要求的条件独立性外，每个节点还关联一个概率表。如果节点 X 没有父节点，则表中只包含先验概率 $P(X)$；如果节点 X 只有一个父节点 Y，则表中包含条件概率 $P(X|Y)$；如果节点 X 有多个父节点 $\{Y_1, Y_2, \cdots, Y_k\}$，则表中包含条件概率 $P(X|Y_1, Y_2, \cdots, Y_k)$。

图 5-3 是贝叶斯网络的一个例子，对心脏病或心口痛患者建模。假设图中每个变量都是二值的。心脏病节点(HD)的父节点对应于影响该疾病的危险因素，如锻炼(E)和饮食(D)等。心脏病节点的子节点对应该病的症状，如胸痛(CP)和高血压(BP)等。如图 5-3 所示，心口痛(HB)可能源于不健康的饮食，同时又可能导致胸痛。

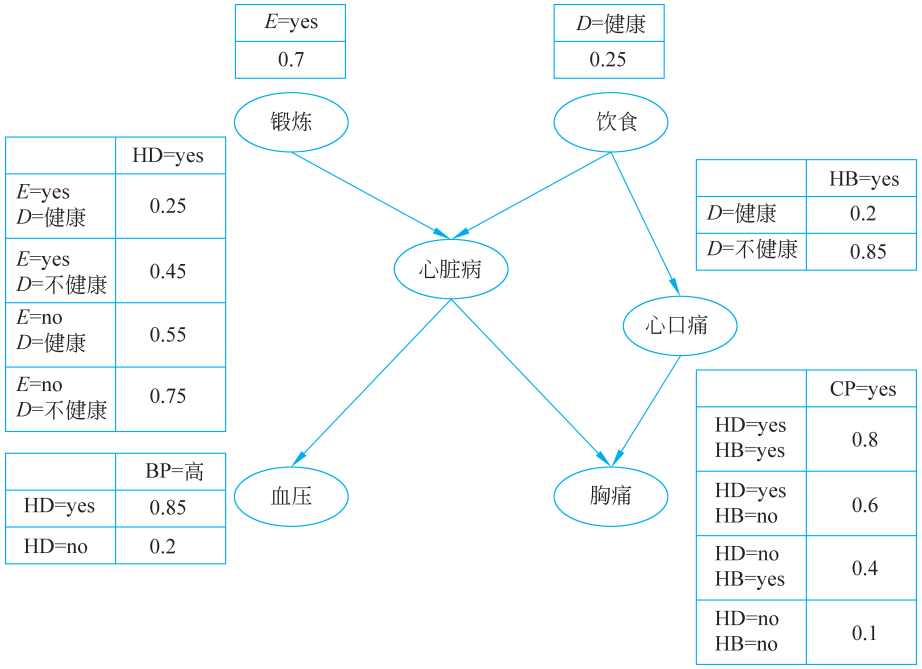

图 5-3 发现心脏病和心口痛病人的贝叶斯网络

影响疾病的危险因素对应的节点只包含先验概率，而心脏病、心口痛以及它们的相应症状所对应的节点都包含条件概率。为了节省空间，图中省略了一些概率。注意 $P(X=\bar{x})=1-P(X=x)$，$P(X=\bar{x}|Y)=1-P(X=x|Y)$，其中 $\bar{x}$ 表示与 x 相反的结

果。因此,省略的概率可以很容易求得。例如,条件概率

$$\begin{aligned}P(\text{心脏病}=\text{no}\mid\text{锻炼}=\text{no},\text{饮食}=\text{健康}) &= 1-P(\text{心脏病}=\text{yes}\mid\text{锻炼}=\text{no},\text{饮食}=\text{健康})\\ &=1-0.55\\ &=0.45\end{aligned}$$

2. 模型建立

贝叶斯网络的建模包括两个步骤:创建网络结构以及估计每个节点的概率表中的概率值。网络拓扑结构可以通过对主观的领域专家知识编码获得,算法 5.1 给出了归纳贝叶斯网络拓扑结构的一个系统过程。

算法 5.1 贝叶斯网络拓扑结构的生成算法。

(1) 设 $T=(X_1,X_2,\cdots,X_d)$ 表示变量的一个总体次序。

(2) FOR $j=1$ to d DO。

(3) 令 $X_T(j)$ 表示 T 中第 j 个次序最高的变量。

(4) 令 $\pi(X_T(j))=\{X_1,X_2,\cdots,X_T(j-1)\}$ 表示排在 $X_T(j)$ 前面的变量的集合。

(5) 从 $\pi(X_T(j))$ 中去掉对 X_j 没有影响的变量(使用先验知识)。

(6) 在 $X_T(j)$ 和 $\pi(X_T(j))$ 中剩余的变量之间画弧。

(7) END FOR。

以图 5-3 为例解释上述步骤,执行步骤(1)后,设变量次序为(E,D,HD,HB,CP,BP),从变量 D 开始,经过步骤(2)~(7),得到以下条件概率。

- $P(D\mid E)$ 化简为 $P(D)$。
- $P(\text{HD}\mid E,D)$ 不能化简。
- $P(\text{HB}\mid \text{HD},E,D)$ 化简为 $P(\text{HB}\mid D)$。
- $P(\text{CP}\mid \text{HB},\text{HD},E,D)$ 化简为 $P(\text{CP}\mid \text{HB},\text{HD})$。
- $P(\text{BP}\mid \text{CP},\text{HB},\text{HD},E,D)$ 化简为 $P(\text{BP}\mid \text{HD})$。

基于以上条件概率,创建节点之间的弧 (E,HD)、(D,HD)、(D,HB)、(HD,CP)、(HB,CP) 和 (HD,BP)。这些弧构成了如图 5-3 所示的网络结构。

算法 5.1 保证生成的拓扑结构不包括环。这一点的证明也很简单。如果存在环,至少有一条弧从低序节点指向高序节点,并且至少存在另一条弧从高序节点指向低序节点。由于算法 5.1 不允许从低序节点到高序节点的弧存在,因此拓扑结构中不存在环。

然而,如果对变量采用不同的排序方案,得到的网络拓扑结构可能会有变化。某些拓扑结构可能质量很差,因为它在不同的节点对之间产生了很多条弧。从理论上讲,可能需要检查所有 $d!$ 种可能的排序才能确定最佳的拓扑结构,这是一项计算开销很大的任务。一种替代的方法是把变量分为原因变量和结果变量,然后从各原因变量向其对应的结果变量画弧。这种方法简化了贝叶斯网络结构的建立。一旦找到了合适的拓扑结构,与各节点关联的概率表就确定了。对这些概率的估计比较容易,与朴素贝叶斯分类器中所用的方法类似。

5.3 贝叶斯算法实例分析


贝叶斯算法实例分析

5.3.1 朴素贝叶斯分类器

【例 5.1】 应用朴素贝叶斯分类器来解决这样一个分类问题：根据天气状况来判断某天是否适合打网球。给定如表 5-1 所示的 14 个训练实例，其中每天由属性 outlook，temperature，humidity 和 windy 来表征，类属性为 play tennis。

表 5-1 14 个训练实例

day	outlook	temperature	humidity	windy	play tennis
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
7	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

现有一测试实例 x：＜outlook＝sunny，temperature＝cool，humidity＝high，windy＝strong＞，问这一天是否适合打网球？显然，我们的任务就是要预测此新实例的类属性 play tennis 的取值(yes 或 no)，为此，我们构建了如图 5-4 所示的朴素贝叶斯分类器。

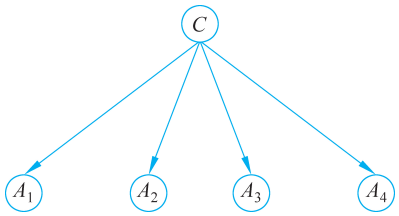

图 5-4 朴素贝叶斯分类器的结构

图中的类节点 C 表示类属性 play tennis，其他 4 个节点 A_1，A_2，A_3，A_4 分别代表 4 个属性 outlook，temperature，humidity，windy，类节点 C 是所有属性节点的父节点，属性节点和属性节点之间没有任何的依赖关系。根据公式有

$$V(x)=\underset{c\in\{\text{yes,no}\}}{\operatorname{argmax}} P(c)\,P(\text{sunny}\,|\,c)\,P(\text{cool}\,|\,\text{c})\,P(\text{high}\,|\,\text{c})\,P(\text{strong}\,|\,\text{c})$$

为计算 $V(x)$，需要从如表 5-1 所示的 14 个训练实例中估计出概率：$P(\text{yes})$，$P(\text{sunny}|\text{yes})$，$P(\text{cool}|\text{yes})$，$P(\text{high}|\text{yes})$，$P(\text{strong}|\text{yes})$，$P(\text{no})$，$P(\text{sunny}|\text{no})$，$P(\text{cool}|\text{no})$，$P(\text{high}|\text{no})$，$P(\text{strong}|\text{no})$。具体的计算如下：

$$P(\text{yes})=9/14$$
$$P(\text{sunny}|\text{yes})=2/9$$
$$P(\text{cool}|\text{yes})=3/9$$
$$P(\text{high}|\text{yes})=3/9$$
$$P(\text{strong}|\text{yes})=3/9$$
$$P(\text{no})=5/14$$
$$P(\text{sunny}|\text{no})=3/5$$
$$P(\text{cool}|\text{no})=1/5$$
$$P(\text{high}|\text{no})=4/5$$
$$P(\text{strong}|\text{no})=3/5$$

所以有

$$P(\text{yes})\,P(\text{sunny}|\text{yes})\,P(\text{cool}|\text{yes})\,P(\text{high}|\text{yes})\,P(\text{strong}|\text{yes})=0.005\,291$$
$$P(\text{no})\,P(\text{sunny}|\text{no})\,P(\text{cool}|\text{no})\,P(\text{high}|\text{no})\,P(\text{strong}|\text{no})=0.020\,570\,4$$

可见，朴素贝叶斯分类器将此实例分类为 no。

【例 5.2】 应用朴素贝叶斯分类器来解决这样一个分类问题：给出一个商场顾客数据库(训练样本集合)，判断某一顾客是否会买计算机。给定如表 5-2 所示的 15 个训练实例，其中每个实例由属性 age，income，student，credit rating 来表征，样本集合的类别属性为 buy computer，该属性有两个不同的取值，即{yes，no}，因此就有两个不同的类别($m=2$)。设 C_1 对应 yes 类别，C_2 对应 no 类别。

表 5-2　15 个训练实例

age	income	student	credit rating	buy computer
≤30	high	no	fair	no
≤30	high	no	excellent	no
31～40	high	no	fair	yes
>40	medium	no	fair	yes
>40	low	yes	fair	yes
>40	low	yes	excellent	no
31～40	low	yes	excellent	yes
≤30	medium	no	fair	no
≤30	low	yes	fair	yes
>40	medium	yes	fair	yes

续表

age	income	student	credit rating	buy computer
≤30	medium	yes	excellent	yes
31～40	medium	no	excellent	yes
31～40	high	yes	fair	yes
>40	medium	no	excellent	no

现有一测试实例 x：(age≤30，income=medium，student=yes，credit rating=fair)，问：这一实例是否会买计算机？我们的任务是要判断给定的测试实例是属于 C_1 还是 C_2。

根据公式有

$$V(x)=\underset{c\in\{\text{yes},\text{no}\}}{\operatorname{argmax}} P(c)P(\text{age}\leqslant 30\,|\,c)P(\text{medium}\,|\,c)P(\text{yes}\,|\,c)P(\text{fair}\,|\,c)$$

为计算 $V(x)$，计算每个类的先验概率 $P(C_i)$，即

$$P(C_i)：P(\text{buy computer}=\text{'yes'})=9/14=0.643$$
$$P(\text{buy computer}=\text{'no'})=5/14=0.357$$

为计算 $P(X\,|\,C_i)$，$i=1$、2，计算下面的条件概率：

$$P(\text{age}=\text{'}\leqslant 30\text{'}\,|\,\text{buy computer}=\text{'yes'})=2/9=0.222$$
$$P(\text{age}=\text{'}\leqslant 30\text{'}\,|\,\text{buy computer}=\text{'no'})=3/5=0.6$$
$$P(\text{income}=\text{'medium'}\,|\,\text{buy computer}=\text{'yes'})=4/9=0.444$$
$$P(\text{income}=\text{'medium'}\,|\,\text{buy computer}=\text{'no'})=2/5=0.4$$
$$P(\text{student}=\text{'yes'}\,|\,\text{buy computer}=\text{'yes'})=6/9=0.667$$
$$P(\text{student}=\text{'yes'}\,|\,\text{buy computer}=\text{'no'})=1/5=0.2$$
$$P(\text{credit rating}=\text{'fair'}\,|\,\text{buy computer}=\text{'yes'})=6/9=0.667$$
$$P(\text{credit rating}=\text{'fair'}\,|\,\text{buy computer}=\text{'no'})=2/5=0.4$$
$$X=(\text{age}\leqslant 30,\text{income}=\text{medium},\text{student}=\text{yes},\text{credit rating}=\text{fair})$$
$$P(X\,|\,C_i)：P(X\,|\,\text{buy computer}=\text{'yes'})=0.222\times 0.444\times 0.667\times 0.667=0.044$$
$$P(X\,|\,\text{buy computer}=\text{'no'})=0.6\times 0.4\times 0.2\times 0.4=0.019$$
$$P(X\,|\,C_i)\cdot P(C_i)：P(X\,|\,\text{buy computer}=\text{'yes'})\cdot P(\text{buy computer}=\text{'yes'})=0.028$$
$$P(X\,|\,\text{buy computer}=\text{'no'})\cdot P(\text{buy computer}=\text{'no'})=0.007$$

因此，对于样本 X，朴素贝叶斯分类预测 buy computer='yes'。

5.3.2 贝叶斯信念网络应用

使用如图 5-3 所示的 BBN 来诊断一个人是否患有心脏病。下面阐释在不同的情况下如何做出诊断。

情况一：没有先验信息。