4.4 朴素贝叶斯分类器（Naïve Bayes）

朴素贝叶斯模型是针对分类问题求解时的一种简化的贝叶斯网络模型。当我们面对的样本数据维度比较高时，如果直接用第二章学习的贝叶斯决策分类器，就要计算在所有特征上的联合概率分布，需要用大量样本计算复杂的条件概率表。这在很多实际问题上是不现实的。所谓朴素贝叶斯（Naïve Bayes）方法采取了一个简化的策略，就是假设各个特征的取值只依赖于类别标签，而特征之间是互相独立的，即

$$p(x_l x_k|\omega_i) = \mathrm{p}(x_l|\omega_i)\mathrm{p}(x_k|\omega_i), \qquad l,k = 1,\cdots,d, k \neq l$$

（4-28）

在该假设下，需要求解的联合概率可以分解为：

$$p(x_1, x_2, \cdots, x_d, \omega_i) = p(x_1|\omega_i)p(x_2|\omega_i)\cdots p(x_d|\omega_i)p(\omega_i)$$

（4-29）

大大简化了联合概率分布的计算。

图 4-8 给出了朴素贝叶斯对应的贝叶斯网络图模型。

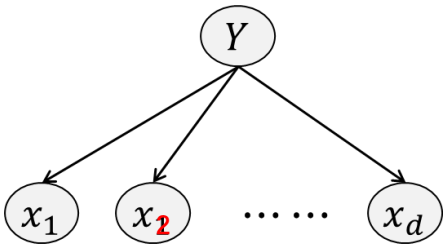

图 4-8　朴素贝叶斯模型的图模型表示

对于朴素贝叶斯模型来说，模型参数的估计变得容易得多。对于各类别的先验概率，可以通过统计训练样本中第 i 类样本占总训练样本的比率来进行估计：

$$p(Y = \omega_i) = \frac{\sum_{j=1}^{N} I(y_j = \omega_i)}{N}$$

（4-30）

其中$I(\cdot)$是指示函数，当括号中条件值满足时取值为 1，不满足时取值为 0。

对于各个特征的条件概率，可以通过第 i 类样本在该特征上的取值进行估计。例如，对于离散取值的特征，对第 k 个特征x_k，若其有S_k种可能的取值，即$\{v_1, v_2, \cdots, v_{S_k}\}$，则参数的极大似然估计为：

$$p(x_k = v_l|Y = \omega_i) = \frac{\sum_{j=1}^{N} I(x_k^{(j)} = v_l, y_j = \omega_i)}{\sum_{j=1}^{N} I(y_j = \omega_i)}$$

（4-31）

其中$x_k^{(j)}$表示第 j 个样本的第 k 个特征的取值。

当训练样本量较少，或者某些特征取值概率较低时，可能会出现 $\sum_{j=1}^{N} I(x_k^{(j)} = v_l, y_j = \omega_i) = 0$ 的情况。这时如果将$\hat{p}(x_k = v_l|Y = \omega_i)$直接设置为 0 可能并不太合理。为了避免因为训练样本的局限使得某些概率取值过低，影响未来判断，通常会在估计概率分布时加入伪计数（pseudo count）来对概率值进行平滑矫正，也被称为拉普拉斯平滑（Laplacian smoothing）。例如用如下方法加入平滑项：

$$p(Y = \omega_i) = \frac{\sum_{j=1}^{N} I(y_j = \omega_i) + 1}{N + C} \tag{4-32}$$

$$p(x_k = v_l|Y = \omega_i) = \frac{\sum_{j=1}^{N} I(x_k^{(j)} = v_l,\ y_j = \omega_i) + 1}{\sum_{j=1}^{N} I(y_j = \omega_i) + S_k} \tag{4-33}$$

其中 C 为类别数，S_k为第k维特征的可能取值数。

对于连续取值的变量特征，我们也可以采用正态分布、均匀分布等模型进行建模和估计分布参数。

朴素贝叶斯方法虽然简单，但在实际中有很多应用。下面举一个简单的例子来说明朴素贝叶斯算法流程。

假设有表 4-2 中的一组数据，反映的是一组客户的年龄、性别、收入信息已经是否买车的记录。我们要用朴素贝叶斯模型建立客户特征与是否买车之间的关系，用于对新客户是否可能会买车进行预测。

表 4-2 一组假设的客户数据

编号	年龄	性别	收入	是否购买
1	<30	男	中	否
2	≥30	女	中	否
3	≥30	女	中	否
4	≥30	女	低	否
5	<30	男	高	否
6	≥30	女	低	否
7	<30	女	低	否
8	<30	女	高	是
9	≥30	男	中	是
10	<30	男	高	否
11	≥30	女	中	否
12	<30	男	低	否

13	≥30	女	中	否
14	≥30	男	低	是
15	≥30	男	中	是
16	≥30	女	低	否

首先通过带有拉普拉斯平滑的最大似然估计来估计类先验概率：

$$p(\text{购买}=\text{是})=\frac{4+1}{16+2}=\frac{5}{18};\ p(\text{购买}=\text{否})=\frac{13}{18}$$

然后分别估计类别条件下的各个特征的条件概率：

$$p(\text{年龄}\geq 30|\text{购买}=\text{是})=\frac{3+1}{4+2}=\frac{4}{6};\ p(\text{年龄}<30|\text{购买}=\text{是})=\frac{1+1}{4+2}=\frac{2}{6}$$

$$p(\text{年龄}\geq 30|\text{购买}=\text{否})=\frac{7+1}{12+2}=\frac{8}{14};\ p(\text{年龄}<30|\text{购买}=\text{否})=\frac{5+1}{12+2}=\frac{6}{14}$$

$$p(\text{性别}=\text{男}|\text{购买}=\text{是})=\frac{3+1}{4+2}=\frac{4}{6};\ p(\text{性别}=\text{女}|\text{购买}=\text{是})=\frac{1+1}{4+2}=\frac{2}{6}$$

$$p(\text{性别}=\text{男}|\text{购买}=\text{否})=\frac{4+1}{12+2}=\frac{5}{14};\ p(\text{性别}=\text{女}|\text{购买}=\text{否})=\frac{8+1}{12+2}=\frac{9}{14}$$

$$p(\text{收入}=\text{高}|\text{购买}=\text{是})=\frac{1+1}{4+3}=\frac{2}{7};\ p(\text{收入}=\text{中}|\text{购买}=\text{是})=\frac{2+1}{4+3}=\frac{3}{7}$$

$$p(\text{收入}=\text{低}|\text{购买}=\text{是})=\frac{1+1}{4+3}=\frac{2}{7};\ p(\text{收入}=\text{高}|\text{购买}=\text{否})=\frac{2+1}{12+3}=\frac{3}{15};$$

$$p(\text{收入}=\text{中}|\text{购买}=\text{否})=\frac{5+1}{12+3}=\frac{6}{15};\quad p(\text{收入}=\text{低}|\text{购买}=\text{否})=\frac{5+1}{12+3}=\frac{6}{15}$$

对于新来的一个测试样本，若其特征取值为{年龄 = 38 岁，性别 = 男，收入 = 高}，我们通过该朴素贝叶斯模型来预测其是否购买：

$$p(\text{购买}=\text{是},\text{年龄}=38\text{岁，性别}=\text{男，收入}=\text{高})$$

$$=p(\text{购买}=\text{是})p(\text{年龄}\geq 30|\text{购买}=\text{是})p(\text{性别}=\text{男}|\text{购买}=\text{是})p(\text{收入}=\text{高}|\text{购买}=\text{是})$$

$$=\frac{5}{18}*\frac{4}{6}*\frac{4}{6}*\frac{2}{7}=0.035$$

$$p(\text{购买}=\text{否},\text{年龄}=38\text{岁，性别}=\text{男，收入}=\text{中})$$

$$=p(\text{购买}=\text{否})p(\text{年龄}\geq 30|\text{购买}=\text{否})p(\text{性别}=\text{男}|\text{购买}=\text{否})p(\text{收入}=\text{高}|\text{购买}=\text{否})$$

$$=\frac{13}{18}*\frac{8}{14}*\frac{5}{14}*\frac{3}{15}=0.029$$

由此可知：

$$p(\text{购买} = \text{是}, \text{年龄} = 38\ \text{岁}，\ \text{性别} = \text{男}，\ \text{收入} = \text{高})$$
$$> p(\text{购买} = \text{否}, \text{年龄} = 38\ \text{岁}，\ \text{性别} = \text{男}，\ \text{收入} = \text{高})$$

因此预测该样本将会购买汽车。