第 4 章　朴素贝叶斯法

朴素贝叶斯（naïve Bayes）法是基于贝叶斯定理与特征条件独立假设的分类方法[①]. 对于给定的训练数据集，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最大的输出 y. 朴素贝叶斯法实现简单，学习与预测的效率都很高，是一种常用的方法.

本章叙述朴素贝叶斯法，包括朴素贝叶斯法的学习与分类、朴素贝叶斯法的参数估计算法.

4.1　朴素贝叶斯法的学习与分类

4.1.1　基本方法

设输入空间 $\mathcal{X} \subseteq \mathbf{R}^n$ 为 n 维向量的集合，输出空间为类标记集合 $\mathcal{Y} = \{c_1, c_2, \cdots, c_K\}$. 输入为特征向量 $x \in \mathcal{X}$，输出为类标记（class label）$y \in \mathcal{Y}$. X 是定义在输入空间 $\mathcal{X}$ 上的随机向量，Y 是定义在输出空间 $\mathcal{Y}$ 上的随机变量. $P(X,Y)$ 是 X 和 Y 的联合概率分布. 训练数据集

$$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$$

由 $P(X,Y)$ 独立同分布产生.

朴素贝叶斯法通过训练数据集学习联合概率分布 $P(X,Y)$. 具体地，学习以下先验概率分布及条件概率分布. 先验概率分布

$$P(Y = c_k), \quad k = 1, 2, \cdots, K \tag{4.1}$$

条件概率分布

$$P(X = x \mid Y = c_k) = P(X^{(1)} = x^{(1)}, \cdots, X^{(n)} = x^{(n)} \mid Y = c_k), \quad k = 1, 2, \cdots, K \tag{4.2}$$

于是学习到联合概率分布 $P(X,Y)$.

条件概率分布 $P(X = x \mid Y = c_k)$ 有指数级数量的参数，其估计实际是不可行的. 事实上，假设 $x^{(j)}$ 可取值有 S_j 个，$j = 1, 2, \cdots, n$，Y 可取值有 K 个，那么参数个数为 $K\prod_{j=1}^{n} S_j$.

① 注意：朴素贝叶斯法与贝叶斯估计（Bayesian estimation）是不同的概念.

朴素贝叶斯法对条件概率分布作了条件独立性的假设. 由于这是一个较强的假设，朴素贝叶斯法也由此得名. 具体地，条件独立性假设是

$$\begin{aligned}P(X=x\mid Y=c_k)&=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\mid Y=c_k)\\&=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}\mid Y=c_k)\end{aligned} \tag{4.3}$$

朴素贝叶斯法实际上学习到生成数据的机制，所以属于生成模型. 条件独立假设等于是说用于分类的特征在类确定的条件下都是条件独立的. 这一假设使朴素贝叶斯法变得简单，但有时会牺牲一定的分类准确率.

朴素贝叶斯法分类时，对给定的输入 x，通过学习到的模型计算后验概率分布 $P(Y=c_k\mid X=x)$，将后验概率最大的类作为 x 的类输出. 后验概率计算根据贝叶斯定理进行：

$$P(Y=c_k\mid X=x)=\frac{P(X=x\mid Y=c_k)P(Y=c_k)}{\sum_k P(X=x\mid Y=c_k)P(Y=c_k)} \tag{4.4}$$

将式 (4.3) 代入式 (4.4) 有

$$P(Y=c_k\mid X=x)=\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)},\quad k=1,2,\cdots,K \tag{4.5}$$

这是朴素贝叶斯法分类的基本公式. 于是，朴素贝叶斯分类器可表示为

$$y=f(x)=\arg\max_{c_k}\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)} \tag{4.6}$$

注意到，在式 (4.6) 中分母对所有 c_k 都是相同的，所以，

$$y=\arg\max_{c_k}P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k) \tag{4.7}$$

4.1.2 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. 这等价于期望风险最小化. 假设选择 0-1 损失函数：

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

式中 $f(X)$ 是分类决策函数. 这时，期望风险函数为

$$R_{\exp}(f)=E[L(Y,f(X))]$$

期望是对联合分布 $P(X,Y)$ 取的．由此取条件期望

$$R_{\exp}(f)=E_X\sum_{k=1}^{K}[L(c_k,f(X))]P(c_k\mid X)$$

为了使期望风险最小化，只需对 $X=x$ 逐个极小化，由此得到：

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

这样一来，根据期望风险最小化准则就得到了后验概率最大化准则：

$$f(x)=\arg\max_{c_k}P(c_k\mid X=x)$$

即朴素贝叶斯法所采用的原理．

4.2 朴素贝叶斯法的参数估计

4.2.1 极大似然估计

在朴素贝叶斯法中，学习意味着估计 $P(Y=c_k)$ 和 $P(X^{(j)}=x^{(j)}\mid Y=c_k)$．可以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

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)}{N},\quad k=1,2,\cdots,K \tag{4.8}$$

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 $P(X^{(j)}=a_{jl}\mid Y=c_k)$ 的极大似然估计是

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

$$j=1,2,\cdots,n;\quad l=1,2,\cdots,S_j;\quad k=1,2,\cdots,K \tag{4.9}$$

式中，$x_i^{(j)}$ 是第 i 个样本的第 j 个特征；a_{jl} 是第 j 个特征可能取的第 l 个值；I 为指示函数．

4.2.2 学习与分类算法

下面给出朴素贝叶斯法的学习与分类算法.

算法 4.1（朴素贝叶斯算法（naïve Bayes algorithm））

输入：训练数据 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i=(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)})^{\mathrm{T}}$，$x_i^{(j)}$ 是第 i 个样本的第 j 个特征，$x_i^{(j)}\in\{a_{j1},a_{j2},\cdots,a_{jS_j}\}$，$a_{jl}$ 是第 j 个特征可能取的第 l 个值，$j=1,2,\cdots,n$，$l=1,2,\cdots,S_j$，$y_i\in\{c_1,c_2,\cdots,c_K\}$；实例 x；

输出：实例 x 的分类.

（1）计算先验概率及条件概率

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

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

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

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

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

（3）确定实例 x 的类

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

例 4.1 试由表 4.1 的训练数据学习一个朴素贝叶斯分类器并确定 $x=(2,S)^{\mathrm{T}}$ 的类标记 y. 表中 $X^{(1)}$，$X^{(2)}$ 为特征，取值的集合分别为 $A_1=\{1,2,3\}$，$A_2=\{S,M,L\}$，Y 为类标记，$Y\in C=\{1,-1\}$.

表 4.1 训练数据

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
$X^{(1)}$	1	1	1	1	1	2	2	2	2	2	3	3	3	3	3
$X^{(2)}$	S	M	M	S	S	S	M	M	L	L	L	M	M	L	L
Y	−1	−1	1	1	−1	−1	−1	1	1	1	1	1	1	1	−1

解 根据算法 4.1，由表 4.1，容易计算下列概率：

$$P(Y=1)=\frac{9}{15},\quad P(Y=-1)=\frac{6}{15}$$

$$P(X^{(1)}=1\mid Y=1)=\frac{2}{9},\quad P(X^{(1)}=2\mid Y=1)=\frac{3}{9},\quad P(X^{(1)}=3\mid Y=1)=\frac{4}{9}$$

$$P(X^{(2)}=S\mid Y=1)=\frac{1}{9},\quad P(X^{(2)}=M\mid Y=1)=\frac{4}{9},\quad P(X^{(2)}=L\mid Y=1)=\frac{4}{9}$$

$$P(X^{(1)}=1\mid Y=-1)=\frac{3}{6},\quad P(X^{(1)}=2\mid Y=-1)=\frac{2}{6},\quad P(X^{(1)}=3\mid Y=-1)=\frac{1}{6}$$

$$P(X^{(2)}=S\mid Y=-1)=\frac{3}{6},\quad P(X^{(2)}=M\mid Y=-1)=\frac{2}{6},\quad P(X^{(2)}=L\mid Y=-1)=\frac{1}{6}$$

对于给定的 $x=(2,S)^{\mathrm{T}}$ 计算：

$$P(Y=1)P(X^{(1)}=2\mid Y=1)P(X^{(2)}=S\mid Y=1)=\frac{9}{15}\cdot\frac{3}{9}\cdot\frac{1}{9}=\frac{1}{45}$$

$$P(Y=-1)P(X^{(1)}=2\mid Y=-1)P(X^{(2)}=S\mid Y=-1)=\frac{6}{15}\cdot\frac{2}{6}\cdot\frac{3}{6}=\frac{1}{15}$$

因为 $P(Y=-1)P(X^{(1)}=2\mid Y=-1)P(X^{(2)}=S\mid Y=-1)$ 最大，所以 $y=-1$． ■

4.2.3 贝叶斯估计

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．具体地，条件概率的贝叶斯估计是

$$P_{\lambda}(X^{(j)}=a_{jl}\mid Y=c_k)=\frac{\sum_{i=1}^{N}I(x_i^{(j)}=a_{jl},y_i=c_k)+\lambda}{\sum_{i=1}^{N}I(y_i=c_k)+S_j\lambda}\tag{4.10}$$

式中 $\lambda\geqslant 0$．等价于在随机变量各个取值的频数上赋予一个正数 $\lambda>0$．当 $\lambda=0$ 时就是极大似然估计．常取 $\lambda=1$，这时称为拉普拉斯平滑（Laplace smoothing）．显然，对任何 $l=1,2,\cdots,S_j$， $k=1,2,\cdots,K$，有

$$P_{\lambda}(X^{(j)}=a_{jl}\mid Y=c_k)>0$$

$$\sum_{l=1}^{S_j}P(X^{(j)}=a_{jl}\mid Y=c_k)=1$$

表明式 (4.10) 确为一种概率分布．同样，先验概率的贝叶斯估计是

$$P_{\lambda}(Y=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)+\lambda}{N+K\lambda}\tag{4.11}$$

例 4.2 问题同例 4.1，按照拉普拉斯平滑估计概率，即取 $\lambda=1$.

解 $A_1=\{1,2,3\}$，$A_2=\{S,M,L\}$，$C=\{1,-1\}$. 按照式 (4.10) 和式 (4.11) 计算下列概率：

$$P(Y=1)=\frac{10}{17},\quad P(Y=-1)=\frac{7}{17}$$

$$P(X^{(1)}=1\mid Y=1)=\frac{3}{12},\quad P(X^{(1)}=2\mid Y=1)=\frac{4}{12},\quad P(X^{(1)}=3\mid Y=1)=\frac{5}{12}$$

$$P(X^{(2)}=S\mid Y=1)=\frac{2}{12},\quad P(X^{(2)}=M\mid Y=1)=\frac{5}{12},\quad P(X^{(2)}=L\mid Y=1)=\frac{5}{12}$$

$$P(X^{(1)}=1\mid Y=-1)=\frac{4}{9},\quad P(X^{(1)}=2\mid Y=-1)=\frac{3}{9},\quad P(X^{(1)}=3\mid Y=-1)=\frac{2}{9}$$

$$P(X^{(2)}=S\mid Y=-1)=\frac{4}{9},\quad P(X^{(2)}=M\mid Y=-1)=\frac{3}{9},\quad P(X^{(2)}=L\mid Y=-1)=\frac{2}{9}$$

对于给定的 $x=(2,S)^{\mathrm{T}}$ 计算：

$$P(Y=1)P(X^{(1)}=2\mid Y=1)P(X^{(2)}=S\mid Y=1)=\frac{10}{17}\cdot\frac{4}{12}\cdot\frac{2}{12}=\frac{5}{153}=0.0327$$

$$P(Y=-1)P(X^{(1)}=2\mid Y=-1)P(X^{(2)}=S\mid Y=-1)=\frac{7}{17}\cdot\frac{3}{9}\cdot\frac{4}{9}=\frac{28}{459}=0.0610$$

由于 $P(Y=-1)P(X^{(1)}=2\mid Y=-1)P(X^{(2)}=S\mid Y=-1)$ 最大，所以 $y=-1$. ■

本章概要

1．朴素贝叶斯法是典型的生成学习方法．生成方法由训练数据学习联合概率分布 $P(X,Y)$，然后求得后验概率分布 $P(Y\mid X)$．具体来说，利用训练数据学习 $P(X\mid Y)$ 和 $P(Y)$ 的估计，得到联合概率分布：

$$P(X,Y)=P(Y)P(X\mid Y)$$

概率估计方法可以是极大似然估计或贝叶斯估计：

2．朴素贝叶斯法的基本假设是条件独立性，

$$\begin{aligned}P(X=x\mid Y=c_k)&=P(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}\mid Y=c_k)\\&=\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}\mid Y=c_k)\end{aligned}$$

这是一个较强的假设．由于这一假设，模型包含的条件概率的数量大为减少，朴素贝叶斯法的学习与预测大为简化．因而朴素贝叶斯法高效，且易于实现．其缺

点是分类的性能不一定很高.

3．朴素贝叶斯法利用贝叶斯定理与学到的联合概率模型进行分类预测.

$$P(Y \mid X)=\frac{P(X,Y)}{P(X)}=\frac{P(Y)P(X \mid Y)}{\sum_{Y} P(Y)P(X \mid Y)}$$

将输入 x 分到后验概率最大的类 y .

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

后验概率最大等价于 0-1 损失函数时的期望风险最小化.

继 续 阅 读

朴素贝叶斯法的介绍可见文献[1, 2]. 朴素贝叶斯法中假设输入变量都是条件独立的，如果假设它们之间存在概率依存关系，模型就变成了贝叶斯网络，参见文献[3].

习 题

4.1 用极大似然估计法推出朴素贝叶斯法中的概率估计公式 (4.8) 及公式 (4.9) .

4.2 用贝叶斯估计法推出朴素贝叶斯法中的概率估计公式 (4.10) 及公式 (4.11) .

参 考 文 献

[1] Mitchell TM. Chapter 1: Generative and discriminative classifiers: Naïve Bayes and logistic regression. In: Machine Learning. Draft, 2005. http://www.cs.cmu.edu/~tom/mlbook/NBayeslogReg.pdf

[2] Hastie T, Tibshirani R, Friedman J. The Elements of Statistical Learning. Data Mining, Inference, and Prediction. Springer-Verlag, 2001（中译本：统计学习基础——数据挖掘、推理与预测. 范明，柴玉梅，昝红英等译. 北京：电子工业出版社，2004)

[3] Bishop C. Pattern Recognition and Machine Learning, Springer, 2006