# 朴素贝叶斯

2021-09-13



# 1. 朴素贝叶斯的学习方法

## 1.1. 概念定义

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$\mathcal{Y}=\{c1, c2, \cdots, c_K\}$。输入为**特征向量**$\mathbf{x} \in \mathcal{X}$，输出为类标记$y \in \mathcal{Y}$。$X$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$T$由$P(X, Y)$独立同分布产生：

$$T=\{(\mathbf{x_1}, y_1),(\mathbf{x_2}, y_2), \cdots,(\mathbf{x_M}, y_M)\}$$

$$\text{注意：}\mathbf{x}\text{为}n \times 1\text{向量，}M\text{为训练集大小}$$

## 1.2. 学习方法

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。

1. 先验概率分布

$$P(Y=c_k), \quad k=1,2, \cdots, K \tag{21}$$

2. 条件概率分布

$$P(X=\mathbf{x} \mid Y=c_k)=P\left(X^{(1)}=x^{(1)}, \cdots, X^{(n)}=x^{(n)} \mid Y=c_k\right), \quad k=1,2, \cdots, K \tag{22}$$

于是，学习到联合概率分布$P(X, Y)$：

$$P(X=\mathbf{x}, Y=c_k)=P(Y=c_k) \cdot P(X=\mathbf{x} \mid Y=c_k) \tag{23}$$

## 1.3. 朴素的含义

问题：条件概率分布$P(X=\mathbf{x} \mid Y=c_k)=P\left(X^{(1)}=x^{(1)}, \cdots, X^{(n)}=x^{(n)} \mid Y=c_k\right)$的参数是指数级的，其估计实际不可实行。

解决方法：对上述条件概率做**条件独立性假设**，即在给定$Y=c_k$的条件下，随机变量$\mathbf{x}$的各分量独立。具体如下所示：

$$\begin{aligned} P(X=\mathbf{x} \mid Y=c_k) &=P\left(X^{(1)}=x^{(1)}, \cdots, X^{(n)}=x^{(n)} \mid Y=c_k\right) \\ &=\prod_{j=1}^n P\left(X^{(j)}=x^{(j)} \mid Y=c_k\right) \end{aligned} \tag{24}$$

朴素贝叶斯实际上学习到生成数据的机制，所以属于生成模型。

## 1.4. 如何做分类

朴素贝叶斯算法分类时，对给定的输入$\mathbf{x}$，通过学习到的模型计算**后验概率分布**$P\left(X=\mathbf{x} \mid Y=c_k\right)$，将后验概率最大的类作为$\mathbf{x}$的预测类别。根据贝叶斯定理：

$$\begin{aligned}
P\left(Y=c_k \mid X=\mathbf{x}\right) &= \frac{P\left(Y=c_k, X=\mathbf{x}\right)}{P(\mathbf{x})} \\
&= \frac{P\left(X=\mathbf{x} \mid Y=c_k\right) P\left(Y=c_k\right)}{\sum_k P(\mathbf{x}, y)} \\
&= \frac{P\left(X=\mathbf{x} \mid Y=c_k\right) P\left(Y=c_k\right)}{\sum_k P\left(X=\mathbf{x} \mid Y=c_k\right) P\left(Y=c_k\right)} \\
&\approx \frac{P\left(Y=c_k\right) \prod_{j=1}^n P\left(X^{(j)}=x^{(j)} \mid Y=c_k\right)}{\sum_k P\left(Y=c_k\right) \prod_{j=1}^n P\left(X^{(j)}=x^{(j)} \mid Y=c_k\right)} \\
&\text{其中，} k=1,2,\cdots,K
\end{aligned} \tag{25}$$

后验概率的朴素解释：现在判断一封电子邮件是否为垃圾邮件，不看内容随机猜，50%的胜率，但是，如果能看到邮件内容，就是知道了特征$\mathbf{x}$，再去判断是否为垃圾邮件，就是所谓的后验概率。

## 1.5. 贝叶斯分类器

朴素贝叶斯分类器可以表示为：

$$y=f(x)=\arg \max_{c_k} \frac{P\left(Y=c_k\right) \prod_{j=1}^n P\left(X^{(j)}=x^{(j)} \mid Y=c_k\right)}{\sum_k P\left(Y=c_k\right) \prod_{j=1}^n P\left(X^{(j)}=x^{(j)} \mid Y=c_k\right)} \tag{26}$$

由于分母对所有$c_k$都是相同的，因此，上式可以简化为：

$$\begin{aligned}
y=f(x) &= \arg \max_{c_k} P\left(Y=c_k \mid X=\mathbf{x}\right) \\
&= \arg \max_{c_k} P\left(Y=c_k\right) \prod_{j=1}^n P\left(X^{(j)}=x^{(j)} \mid Y=c_k\right)
\end{aligned} \tag{27}$$

# 2. 后验概率最大化的含义

朴素贝叶斯法将实例预测为后验概率最大的类别，这等价于期望风险最小化。

假设选择0-1损失函数：

$$L(Y, f(X))=\begin{cases}1, & Y \neq f(X) \\ 0, & Y=f(X)\end{cases} \text{其中，} f(X) \text{是分类决策函数} \tag{28}$$

那么期望风险代表的就是损失的平均值，期望是对联合分布$P(X, Y)$取的，期望风险函数为：

$$\begin{aligned}
& R_{\exp}(f)=E[L(Y, f(X))] \\
&= \int_{\chi \times y} L(y, f(x)) P(x, y) d x d y \\
&= \int_{\chi \times y} L(y, f(x)) P(y \mid x) P(x) d x d y \\
&= \int_\chi \int_Y L(y, f(x)) P(y \mid x) d y P(x) d x
\end{aligned} \tag{29}$$

令$H(x)=\int_Y L(y, f(x)) P(y \mid x) d y$，$H(x)$中损失函数大于等于0，条件概率$P(y \mid x)$大于0，因此$H(x)$也大于0。同时$P(x)$也大于0，且当$X=x$时，$P(x)$（先验概率）为常数，因此期望风险最小化可转换为条件期望最小化，即

$$\arg \min_{y \in \mathcal{Y}} R_{\exp}(f)=\arg \min_{y \in \mathcal{Y}} H(x) \tag{30}$$

为了使期望风险最小化，只需对数据集$T$中每个$\mathbf{x}$逐个最小化即可，由此可得到：

$$\begin{aligned}
f(x) &= \arg \min_{y \in \mathcal{Y}} R_{\exp}(f) \\
&= \arg \min_{y \in \mathcal{Y}} H(\mathbf{x})
\end{aligned}$$

$$\begin{aligned}
&= \arg\min_{y\in\mathcal{Y}} \sum_{k=1}^{K} L(c_k, y) P(c_k \mid X=\mathbf{x}) \\
&= \arg\min_{y\in\mathcal{Y}} \sum_{k=1}^{K} P(y \neq c_k \mid X=\mathbf{x}) \\
&= \arg\min_{y\in\mathcal{Y}} (1 - P(y = c_k \mid X=\mathbf{x})) \\
&= \arg\max_{y\in\mathcal{Y}} P(y = c_k \mid X=\mathbf{x})
\end{aligned} \tag{31}$$

其中，$y$为模型预测的输出类别，$c_k$为真是类别

公式解释：$y$必然属于$\mathcal{Y} = \{c1, c2, \cdots, c_K\}$中的一个，假设为$c_k$。那么剩下的$c_1, c_2, c_{k-1}, c_{k+1}, \cdots, c_K$的概率和必然为$1 - P(c_k)$

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，即朴素贝叶斯采用的原理：

$$f(x) = \arg\max_{c_k} P(c_k \mid X=\mathbf{x}) \tag{32}$$

# 3. 朴素贝叶斯的参数估计

## 3.1. 极大似然估计

在朴素贝叶斯方法中，学习意味着估计：

1. $P(Y = c_k)$
2. $P\left(X^{(j)} = x^{(j)} \mid Y = c_k\right)$

因此，可以使用极大似然法估计相应的概率。

### 3.1.1. $P(Y = c_k)$估计

先验概率$P(Y = c_k)$的极大似然估计是：

$$P(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, \quad k = 1, 2, \cdots, K \tag{33}$$

### 3.1.2. $P\left(X^{(j)} = x^{(j)} \mid Y = c_k\right)$估计

设第$j$个特征$x^{(j)}$可能取值的集合是$\{a_{j1}, a_{j2}, \cdots, a_{jS_j}\}$，条件概率$P(X^{(j)} = a_{jl} \mid Y = c_k)$的极大似然估计是：

$$\begin{aligned}
&P\left(X^{(j)} = a_{jl} \mid Y = c_k\right) = \frac{\sum_{i=1}^{N} I\left(x_i^{(j)} = a_{jl}, y_i = c_k\right)}{\sum_{i=1}^{N} I(y_i = c_k)} \\
&j = 1, 2, \cdots, n; \quad l = 1, 2, \cdots, S_j; \quad k = 1, 2, \cdots, K
\end{aligned} \tag{34}$$

其中，$x_i^{(j)}$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。

### 3.1.3. 学习与分类算法

算法1.1(朴素贝叶斯算法)

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

输出：$\mathbf{x}$的分类

(1)计算先验概率及条件概率

$$\begin{aligned}
&P(Y = c_k) = \frac{\sum_{i=1}^{N} I(y_i = c_k)}{N}, \quad k = 1, 2, \cdots, K \\
&P\left(X^{(j)} = a_{jl} \mid Y = c_k\right) = \frac{\sum_{i=1}^{N} I\left(x_i^{(j)} = a_{jl}, y_i = c_k\right)}{\sum_{i=1}^{N} I(y_i = c_k)}
\end{aligned} \tag{35}$$

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

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

$$P\left(Y=c_k\right)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\mid Y=c_k\right),\quad k=1,2,\cdots,K \tag{36}$$

(3)预测实例$\mathbf{x}$的类别

$$y=\arg\max_{c_k}P\left(Y=c_k\right)\prod_{j=1}^{n}P\left(X^{(j)}=x^{(j)}\mid Y=c_k\right) \tag{37}$$

## 3.2. 贝叶斯估计

使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，使分类产生误差。解决这一问题的方法是使用贝叶斯估计。具体地，条件概率的贝叶斯估计是：

$$P_\lambda\left(X^{(j)}=a_{jl}\mid Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(x_i^{(j)}=a_{jl},y_i=c_k\right)+\lambda}{\sum_{i=1}^{N}I\left(y_i=c_k\right)+S_j\lambda}\text{其中，}\lambda\geq 0 \tag{38}$$

等价于在随机变量各个取值的频数上加上一个正数$\lambda>0$。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。显然，对于任何$l=1,2,\cdots,S_j$，$k=1,2,\cdots,K$，有：

$$\begin{aligned}&P_\lambda\left(X^{(j)}=a_{jl}\mid Y=c_k\right)>0\\&\sum_{l=1}^{S_j}P\left(X^{(j)}=a_{jl}\mid Y=c_k\right)=1\end{aligned} \tag{39}$$

说明公式(18)确实为一种概率分布。

同样，先验概率的贝叶斯估计是：

$$P_\lambda\left(Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(y_i=c_k\right)+\lambda}{N+K\lambda} \tag{40}$$