第四章　朴素贝叶斯法

- 朴素贝叶斯法是基于贝叶斯定理与特征条件独立假设的分类方法。
- 朴素贝叶斯法与贝叶斯估计不同。

4.1　朴素贝叶斯法的学习和分类

4.1.1　基本方法

- 输入：$\mathcal{X} \subseteq R^n$为n维向量集合
- 输出：$\mathcal{Y} = \{c_1, c_2, ..., c_k\}$，输出为类标记，$y \in \mathcal{Y}$
- 数据集：$T = \{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$由$P(X, Y)$独立同分布产生。
- 训练过程：通过训练数据集，学习$P(X, Y)$（联合分布概率），也就是分别学习$\text{先验概率分布} P(Y = c_k), k = 1, 2, 3...k \text{和条件概率分布} P(X = x|y = c_k) = P(X^{(1)} = x^{(1)}, ..., X^{(n)} = x^{(n)}|Y = c_k\}$
    - 但是这种条件概率的计算方式由指数数量的参数，其估计实际是不可行的。

朴素贝叶斯法：

- 朴素贝叶斯法为了解决参数过多的问题，做出了条件独立性的假设：

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

即假设用于分类的特征在类确定的条件下是相互独立的。

- 后验概率的计算由贝叶斯定理得出：

$$P(Y = c_k|X = x) = \frac{P(X = x|Y = c_k)P(Y = c_k)}{\sum_k P(X = x|Y = c_k)P(Y = c_k)}$$

代入条件独立性假设可得：

$$P(Y = c_k|X = x) = \frac{P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)}|Y = c_k)}{\sum_k P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)}|Y = c_k)} \\ k = 1, 2, ..., K$$

进而得到朴素贝叶斯分类器L：

$$y = f(x) = argmax_{ck} \frac{P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)}|Y = c_k)}{\sum_k P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)}|Y = c_k)}$$

我们又知道，对于任意一个数据项或特征，上式分母都相同，所以：

$$y = argmax_{ck} P(Y = c_k)\prod_j P(X^{(j)} = x^{(j)}|Y = c_k)$$

4.1.2　后验概率最大化的含义

- 这一部分主要介绍选择$P(Y = c_k|X = x)$最大的c_k作为x的分类结果的合理性证明。
- 假设选择0-1损失函数：

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

我们可以得到期望风险函数：

$$\begin{aligned}R_{exp}(f)&=E[L(Y,f(X))]\\&=E_X\sum_{k=1}^{K}[L(c_k,f(X))]P(c_k|X)\end{aligned}$$

为了使期望风险最小化，只需对$X=x$逐个极小化，得到：

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

通过以上推导我们可以看出，当$P(y=c_k|X=x)$取最大时，我们的期望风险将降到最低。这也就是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

- 在计算先验概率和条件概率时我们可以应用极大似然估计，计算如下：

先验概率：

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

条件概率：

$$\begin{gathered}P(X^{(j)}=a_{jl}|Y=c_k)=\frac{\sum_{i=1}^{N}I(x_i^{(i)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N}I(y_i=c_k)}\\j=1,2,...,n;l=1,2,...,S_j;k=1,2,...,K\end{gathered}$$

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（计数函数）

4.2.2 学习与分类算法

1. 计算先验概率和条件概率
2. 对于给定的实例$x=(x^{(1)},x^{(2)},...,x^{(n)})^T$,计算：

$$P(Y=C_k)\prod_{j=1}^{n}P(X^{(j)}=x^{(j)}|Y=c_k),k=1,2,...,K$$

3. 确定实例x的类：

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

4.2.3 贝叶斯估计

- 在极大似然估计中，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。这一问题可通过贝叶斯估计解决。
- 贝叶斯估计实际上是在每个频数上都加了一个相同的正数λ，当$\lambda=0$时，贝叶斯估计就是极大似然估计。
- 通常我们将λ取1，此式成为拉普拉斯平滑。并且可知，此式的贝叶斯估计是一种概率分布。
- 条件概率的贝叶斯估计：

$$P_\lambda = (X^{(j)}=a_{jl}|Y=c_k) = \frac{\sum_{i=1}^{N} I(x_i^{(j)}=a_{jl}, y_i=c_k)+\lambda}{\sum_{i=1}^{N} I(y_i=c_k)+S_j\lambda}，\lambda \geq 0$$

- 先验概率的贝叶斯估计：

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