Naive Bayes Classifier (朴素贝叶斯方法)

1. 朴素贝叶斯法的学习及分类

(1) 模型假设

输入：p 维特征向量 (Covariate) $x \in \mathbb{R}^p$;

输出：类别标记 (class label) $y \in \{c_1, \cdots, c_K\}$。当 $K = 2$ 时，对应二分类问题。

训练数据集：$T = \{(x_1, y_1), (x_2, y_2), \cdots, (x_N, y_N)\}$; 假设$(x_i, y_i)$ 由联合概率分布 $P(X, Y)$ 独立同分布产生。

目标: 朴素贝叶斯法通过训练数据学习联合概率分布 $P(X, Y)$, 进而利用贝叶斯定理, 求出后验概率最大的输出 y. 为得到 $P(X, Y)$，可以学习先验概率及条件概率分布.

- 先验概率分布：$P(Y = c_k)$ $(k = 1, \cdots, K)$.
- 条件概率分布：

$$P(X = x|Y = c_k) = P(X^{(1)} = x^{(1)}, \cdots, X^{(p)} = x^{(p)}|Y = c_k),$$

  其中 $X^{(j)}$ 和 $x^{(j)}$ 分别代表 X 和 x 的第 j 个分量。

(2) 后验概率最大化

给定输入 x，将后验概率最大的类作为 x 的类的输出。后验概率可由贝叶斯定理而得：

$$P\left(Y = c_k|X = x\right) = \frac{P\left(X = x|Y = c_k\right) P\left(Y = c_k\right)}{\sum_k P\left(X = x|Y = c_k\right) P\left(Y = c_k\right)} \tag{1.1}$$

(3) 条件独立性假设

条件概率分布 $P(X = x|Y = c_k)$ 的估计难度较大。假设 $x^{(j)}$ 为离散型，可能的取值有 S_j 个，Y 的可能取值有 K 个，那么待估的参数个数为 $K\prod_{j=1}^{p} S_j$.

因此，朴素贝叶斯法对条件概率做了条件独立性假设：

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

这一假设使得朴素贝叶斯法变得简单，但会损失一定的分类准确性。

后验概率代入条件独立性假设的结果得：

$$P\left(Y=c_k|X=x\right)=\frac{P\left(Y=c_k\right)\prod_j P\left(X^{(j)}=x^{(j)}|Y=c_k\right)}{\sum_k P\left(Y=c_k\right)\prod_j P\left(X^{(j)}=x^{(j)}|Y=c_k\right)},\quad k=1,2,\cdots,K \tag{1.2}$$

由于上式分母对于所有类别 c_k 都相同，所以有：

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

注：<u>后验概率最大化等价于期望风险最小化（此时选择 0-1 损失函数）。</u>

以下进一步说明：<u>后验概率最大化等价于期望风险最小化（此时选择 0-1 损失函数）。</u>

记 0-1 损失函数为 $L(Y,f(X))=1$ 如果 $Y\neq f(X)$; 否则 $L(Y,f(X))=0$. 此时期望函数为

$$R_{exp}(f)=E\{L(Y,f(X))\}=E_X\Big\{\sum_k L(c_k,f(X))P(c_k|X)\Big\}.$$

对期望风险极小化只需要对每个 $X=x$ 逐个极小化，由此可得：

$$\begin{aligned} f(x) &= \arg\min_y \sum_k L(c_k, y) P(c_k|X=x) \\ &= \arg\min_y \sum_k I(y \neq c_k) P(Y=c_k|X=x) \\ &= \arg\min_y \left\{ P(Y \neq y|X=x) \right\} \\ &= \arg\max_y P(Y=y|X=x). \end{aligned}$$

因此，

$$f(x) = \arg\max_y P(Y=y|X=x).$$

2 参数估计

2.1 极大似然估计

先验概率的极大似然估计为：

$$P(Y=c_k) = \frac{\sum_{i=1}^N I(y_i = c_k)}{N} \tag{2.1}$$

设第 j 个特征可能的取值集合为 $a_{j1}, \cdots, a_{jS_j}$, 则条件概率的极大似然估计为：

$$P(X^{(j)} = a_{jl}|Y=c_k) = \frac{\sum_i I(x_i^{(j)} = a_{jl}, y_i = c_k)}{\sum_i I(y_i = c_k)} \tag{2.2}$$

2.2 学习与分类算法

输入：训练数据 $\{(x_1, y_1), \cdots, (x_N, y_N)\}$

输出：实例 x 的分类。

(1) 计算先验概率 (2.1) 及条件概率 (2.2)

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

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

(3) 确定 x 的分类：

注：对于连续变量 X，可使用核方法（kernel methods）估计其概率密度函数，具体方法参见 Elements Chapter 6.6.

2.3 贝叶斯估计

极大似然估计可能会出现估计的概率值为 0 的情形，会影响到后验概率的计算。此时往往采用贝叶斯估计。此时条件概率的估计为：

$$P_{\lambda}\left(X^{(j)}=a_{jl}|Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(x_i^{(j)}=a_{jl},y_i=c_k\right)+\lambda}{\sum_{i=1}^{N}I\left(y_i=c_k\right)+S_j\lambda}\tag{2.3}$$

贝叶斯估计等价于随机变量在各个取值的频数上加 λ。一般取 $\lambda=1$。

先验概率的贝叶斯估计为：

$$P_{\lambda}\left(Y=c_k\right)=\frac{\sum_{i=1}^{N}I\left(y_i=c_k\right)+\lambda}{N+K\lambda}\tag{2.4}$$