6_朴素贝叶斯

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。

6_1_朴素贝叶斯法的学习与分类

训练数据集

$$T = \left\{ \left(\mathbf{x}_1, y_1 \right), \left(\mathbf{x}_2, y_2 \right), \cdots, \left(\mathbf{x}_N, y_N \right) \right\}$$

由$P\left(X, Y\right)$独立同分布产生。其中，$\mathbf{x}_i \in \mathcal{X} \subseteq \mathbb{R}^n, y_i \in \mathcal{Y} = \left\{ c_1, c_2, \cdots, c_K \right\}, i = 1, 2, \cdots, N$，$\mathbf{x}_i$为第$i$个特征向量（实例），$y_i$为$\mathbf{x}_i$的类标记，$X$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。$P\left(X, Y\right)$是$X$和$Y$的联合概率分布。

条件独立性假设

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

即，用于分类的特征在类确定的条件下都是条件独立的。

由

$$P(X=\mathbf{x}, Y=c_k) = P(X=\mathbf{x}|Y=c_k)\,P(Y=c_k)$$
$$P(X=\mathbf{x}, Y=c_k) = P(Y=c_k|X=\mathbf{x})\,P(X=\mathbf{x})$$

得

$$\begin{aligned} P(X=\mathbf{x}|Y=c_k)\,P(Y=c_k) &= P(Y=c_k|X=\mathbf{x})\,P(X=\mathbf{x}) \\ P(Y=c_k|X=\mathbf{x}) &= \frac{P(X=\mathbf{x}|Y=c_k)\,P(Y=c_k)}{P(X=\mathbf{x})} \\ &= \frac{P(X=\mathbf{x}|Y=c_k)\,P(Y=c_k)}{\sum_Y P(X=\mathbf{x}, Y=c_k)} \\ &= \frac{P(X=\mathbf{x}|Y=c_k)\,P(Y=c_k)}{\sum_Y P(X=\mathbf{x}|Y=c_k)\,P(Y=c_k)} \\ &= \frac{P(Y=c_k)\prod_{j=1}^n P\left(X^{(j)}=x^{(j)}|Y=c_k\right)}{\sum_Y P(Y=c_k)\prod_{j=1}^n P\left(X^{(j)}=x^{(j)}|Y=c_k\right)} \end{aligned}$$

朴素贝叶斯分类器可表示为

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

6_2_朴素贝叶斯法的参数估计

6_2_1_极大似然估计

朴素贝叶斯模型参数的极大似然估计

1. 先验概率$P(Y=c_k)$的极大似然估计

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

2. 设第j个特征$x^{(j)}$可能取值的集合为$\left\{a_{j1}, a_{j2}, \cdots, a_{jS_j}\right\}$，条件概率$P\left(X^{(j)}=a_{jl}|Y=c_k\right)$的极大似然估计

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

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是指示函数。

朴素贝叶斯算法：

输入：线性可分训练数据集$T=\{(\mathbf{x}_1,y_1),(\mathbf{x}_2,y_2),\cdots,(\mathbf{x}_N,y_N)\}$，其中$\mathbf{x}_i=\left(x_i^{(1)},x_i^{(2)},\cdots,x_i^{(n)}\right)^T$，$x_i^{(j)}$是第$i$个样本的第$j$个特征，$x_i^{(j)}\in\left\{a_{j1},a_{j2},\cdots,a_{jS_j}\right\}$，$a_{jl}$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$\mathbf{x}$；

输出：实例$\mathbf{x}$的分类

1. 计算先验概率及条件概率

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

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

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

3. 确定实例x的类别

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

6_2_1_贝叶斯估计

朴素贝叶斯模型参数的贝叶斯估计

1. 条件概率的贝叶斯估计

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

   式中$\lambda\geq 0$。当$\lambda=0$时，是极大似然估计；当$\lambda=1$时，称为拉普拉斯平滑。

2. 先验概率的贝叶斯估计

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