朴素贝叶斯法原理

【参考资料】

李航 《统计学习方法》

周志华 《机器学习》

1. 基本方法

设输入空间$\mathcal{X} \subseteq \mathbf{R}^{n}$，输出空间为类标记集合$\mathcal{Y}=\{c_{1}, c_{2}, \ldots, c_{K}\}$。训练数据集
$T=\{(x_{1}, y_{1}),(x_{2}, y_{2}), \cdots,(x_{N}, y_{N})\}$由联合概率分布$P(X, Y)$独立同分布产生。

朴素贝叶斯法通过训练数据集来学习联合概率分布$P(X, Y)$。具体通过学习先验概率分布

$$P\left(Y=c_{k}\right), \quad k=1,2, \cdots, K$$

以及条件概率分布

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

来学习联合概率分布$P(X, Y)$。

问题的难点在于条件概率分布$P(X=x|Y=c_{k})$有指数量级的参数，估计参数较困难。假设某个具体的特征属性$x^{(j)}$的可能取值有S_{j}个，$j=1,2, \cdots, n$， 类别Y的可能取值有K个，那么参数的个数为$K \prod_{j=1}^{n} S_{j}$。

为了削减参数，朴素贝叶斯法采用**条件独立假设**，这是一个很强的假设。具体地，条件独立性假设是：

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

即各个特征属性相互独立，不存在依赖关系，并且它们对分类的贡献都相同。

在条件独立性假设下，可以用贝叶斯公式来计算后验概率

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

这是朴素贝叶斯分类器的基本形式，可以进一步表示为

$$y=f(x)=\arg \max_{c_{k}} \frac{P\left(Y=c_{k}\right) \prod_{j} P\left(X^{(j)}=x^{(j)}|Y=c_{k}\right)}{\sum_{k} P\left(Y=c_{k}\right) \prod_{j} P\left(X^{(j)}=x^{(j)}|Y=c_{k}\right)}$$

注意到上式中的分母对所有c_{k}都是相同的，所以

$$y=\arg\max_{c_k} P\left(Y=c_k\right)\prod_j P\left(X^{(j)}=x^{(j)}|Y=c_k\right)$$

2. 参数估计

朴素贝叶斯法中，采用极大似然法估计参数。那么先验概率$P\left(Y=c_k\right)$的极大似然估计是

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

即类别为c_k的样本数量占总样本数量的比例。

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

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

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

即在类别为c_k的样本中，特征$x^{(j)}$的取值为a_{jl}的样本所占的比例。

3. 拉普拉斯平滑

当特征$x^{(j)}$的某个可能取值a_{jl}没有出现在类别为c_k的样本中时，极大似然估计的结果会变为0，从而导致整个条件概率连乘的结果为0，这是不合理的。因此会采用贝叶斯估计的方法对概率进行修正。

修正后的条件概率为

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

其中，$\lambda\geqslant 0$。这相当于用特征$x^{(j)}$可能取值的数目对概率进行了修正。当$\lambda=0$时，就是极大似然估计。常取$\lambda=1$，这时又称为**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

$$P_\lambda\left(X^{(j)}=a_{jl}|Y=c_k\right)>0$$
$$\sum_{l=1}^{s_j} P\left(X^{(j)}=a_{jl}|Y=c_k\right)=1$$

表明修正后的结果仍然时一种概率分布。

同理，先验概率修正后为

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

即使用类别的可能取值数目进行修正。