Naive Bayesian 推导

Soul Walker

2020 年 7 月 22 日

1 背景知识

如果 X 和 Y 相互独立，那么有

$$P(X,Y)=P(X)P(Y) \tag{1}$$

条件概率公式

$$P(Y|X)=\frac{P(X,Y)}{P(X)} \tag{2}$$

$$P(X|Y)=\frac{P(X,Y)}{P(Y)} \tag{3}$$

$$P(Y|X)=\frac{P(X|Y)P(Y)}{P(X)} \tag{4}$$

全概率公式

$$P(X)=\sum_k P(X|Y=Y_k)P(Y_k)\text{其中}\sum_k P(Y_k)=1 \tag{5}$$

贝叶斯公式

$$P(Y_k|X)=\frac{P(X|Y_k)P(Y_k)}{\sum_k P(X|Y=Y_k)P(Y_k)} \tag{6}$$

2 朴素贝叶斯模型

输入：训练数据集

$$T=\{\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N)\}=\{\mathbf{x}_i\}_{i=1}^N$$

其中, $\mathbf{x}_i\in\chi\in\mathbb{R}^N$ 为实例的特征向量, $\mathbf{x}_i=(\mathbf{x}_i^{(1)},\mathbf{x}_i^{(2)},...,\mathbf{x}_i^{(n)})^T$, 对于每一个样本 $\mathbf{x}_i$, $i=1,2,\cdots,N$, 第 i 个样本的第 j 个特征，即 $x_i^{(j)}\in\{a_{j1},a_{j2},\cdots,a_{jS_j}\}$， $y\in\{c_1,c_2,\cdots,c_k\}$， $k=1,2,\cdots,K$，实例特征向量 $\mathbf{x}$，$\mathbf{X}$ 为训练样本集，形状为 (N,p) ，$\mathbf{Y}$ 是训练标签集，形状为 $(N,1)$。

输出：实例 $\mathbf{x}$ 所属的类别 y。

那我们可以首先根据输入数据学习到贝叶斯先验概率

$$P(\mathbf{Y}=c_k)=\frac{\sum_{i=1}^N I(y_i=c_k)}{N},k=1,2,\cdots,K \tag{7}$$

接着我们可以学习出条件概率分布

$$P(\mathbf{X}^{(j)}=a_{jl}|\mathbf{Y}=c_k)=\frac{\sum_{i=1}^{N}I(\mathbf{x}_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N}I(y_i=c_k)} \tag{8}$$

然后对于给定的实例 $\mathbf{x}=(\mathbf{x}^{(1)},\mathbf{x}^{(2)},\cdots,\mathbf{x}^{(n)})^T$，我们可以计算

$$f(c_k)=P(\mathbf{Y}+c_k)\prod_{j=1}^{n}P(\mathbf{X}^{(j)}=\mathbf{x}^{(j)}|\mathbf{Y}=c_k),k=1,2,\cdots,K \tag{9}$$

根据贝叶斯公式

$$P(\mathbf{Y}=c_k|\mathbf{X}=\mathbf{x})=\frac{P(\mathbf{X}=\mathbf{x}|\mathbf{Y}=c_k)P(\mathbf{Y}=c_k)}{P(\mathbf{X}=\mathbf{x})}=\frac{\prod_{j=1}^{n}P(\mathbf{X}^{(j)}=\mathbf{x}^{(j)}|\mathbf{Y}=c_k)P(\mathbf{Y}=c_k)}{P(\mathbf{X}=\mathbf{x})} \tag{10}$$

我们可以确定实例 $\mathbf{x}$ 的分类

$$y=\arg\max_{c_k}P(Y=c_k)\prod_{j=1}^{n}P(\mathbf{X}^{(j)}=\mathbf{x}^{(j)}|\mathbf{Y}=c_k)=\arg\max_{c_k}f(c_k) \tag{11}$$

也就是后验概率最大化

3 参数的贝叶斯估计

因为如果在训练集里面，某一类别的实例数为 0，将导致条件概率的分母为 0。这里我们只考虑贝叶斯估计。为解决这个问题，引入拉普拉斯平滑，增加一个正数 $\lambda>0$（这里就是这个算法需要注意的地方）。

假设已知我们经过极大似然估计的结果

先验概率

$$P(\mathbf{Y}=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)}{N} \tag{12}$$

后验概率

$$P(\mathbf{X}^{(j)}=a_{jl}|\mathbf{Y}=c_k)=\frac{\sum_{i=1}^{N}I(\mathbf{x}_i^{(j)}=a_{jl},y_i=c_k)}{\sum_{i=1}^{N}I(y_i=c_k)} \tag{13}$$

加入拉普拉斯平滑之后

先验概率

$$P(\mathbf{Y}=c_k)=\frac{\sum_{i=1}^{N}I(y_i=c_k)+\lambda}{N+K\lambda} \tag{14}$$

后验概率

$$P(\mathbf{X}^{(j)}=a_{jl}|\mathbf{Y}=c_k)=\frac{\sum_{i=1}^{N}I(\mathbf{x}_i^{(j)}=a_{jl},y_i=c_k)+\lambda}{\sum_{i=1}^{N}I(y_i=c_k)+S_j\lambda} \tag{15}$$