朴素贝叶斯

cloud

2016.12.15

1 朴素贝叶斯法

1.1 基本方法

朴素贝叶斯法是基于贝叶斯定理和条件独立性假设的分类方法。简单来说对于训练数据集，假设特征之间相互独立，在这里我们定义 X 是输入空间上的随机向量，Y 是定义在输出空间上的随机变量，x 是输入的特征向量，y 是输出的类标记，c_k 表示所有可能的输出 y，其中 $k = 1, 2, ...K$，那我们求的目标就是使得后验概率最大化的那个 c_k：

$$arg \max_{c_k} P(Y = c_k | X = x)$$

为了求解概率，概率公式要进一步写成：

$$\begin{aligned} P(Y = c_k | X = x) &= \frac{P(X = x, Y = c_k)}{P(X = x)} \\ &= \frac{P(X = x | Y = c_k) P(Y = c_k)}{P(X = x)} \end{aligned} \tag{1}$$

问题就变成需要求解 (1) 式中的这 3 个概率。因为朴素贝叶斯作了条件独立性假设，所以 $P(X = x | Y = c_k)$ 可以写成如下所示，x_j 表示样本 x 的第 j 个特征。

$$\begin{aligned} P(X = x | Y = c_k) &= P(X_1 = x_1, ...X_n = x_n | Y = c_k) \\ &= \prod_{j=1}^{n} P(X_j = x_j | Y = c_k) \end{aligned} \tag{2}$$

下面来求 $P(X = x)$。在这里 $P(X = x)$ 可以写成样本 x 为每一个类的联合概率的加和。

$$\begin{aligned} P(X = x) &= \sum_{k=1}^{K} P(X = x, Y = c_k) \\ &= \sum_{k=1}^{K} P(X = x | Y = c_k) P(Y = c_k) \end{aligned} \tag{3}$$

将 (2)(3) 代入 (1) 可得

$$P(Y=c_k|X=x)=\frac{P(Y=c_k)\prod_{j=1}^{n}P(X_j=x_j|Y=c_k)}{\sum_{k=1}^{K}P(Y=c_k)\prod_{j=1}^{n}P(X_j=x_j|Y=c_k)}$$

又因为上式中分母对于所有的 c_k 都是相同的，所以当比较哪个种类条件概率最大时只需比较分子即可。所以我们求的概率最大的类别 y 可以写成：

$$y=arg\max_{c_k}P(Y=c_k)\prod_{j=1}^{n}P(X_j=x_j|Y=c_k) \tag{4}$$

所以问题最终转化为求解参数 $P(Y=c_k)$ 和所有的 $P(X_j=x_j|Y=c_k)$。

1.2 参数估计（极大似然估计）

这里的参数估计可以理解为去求解 $P(Y=c_k)$ 和 $P(X_j=x_j|Y=c_k)$，那么怎么用极大似然估计去求解呢？对我们来说先验概率 $P(Y=c_k)$ 很自然的理解为用 c_k 种类的数目除以样本总数 N，如下所示, 这里的 $y^{(i)}$ 是第 i 个样本的类别。

$$P(Y=c_k)=\frac{\sum_{i=1}^{N}I(y^{(i)}=c_k)}{N} \tag{5}$$

那这个公式从数学上是怎么用极大似然估计法推导出来的呢？下面给出证明过程：

首先令参数 $P(Y=c_k)=\theta_k$，在这里 N 指的是样本总数，N_k 指的是类别为 c_k 的数目。对于参数 θ 的极大似然函数可以写成如下形式：

$$L(\theta)=\prod_{i=1}^{N}P(Y=y^{(i)})=\prod_{k=1}^{K}\theta_k^{N_k}$$

对似然函数取对数得：

$$l(\theta)=ln(L(\theta))=\sum_{k=1}^{K}N_k ln\theta_k$$

要求该函数的极大值，我们注意到有约束条件 $\sum_{k=1}^{K}\theta_k=1$，我们利用拉格朗日乘子法，即将目标函数写成：

$$l(\theta,\lambda)=\sum_{k=1}^{K}N_k ln\theta_k+\lambda(\sum_{k=1}^{K}\theta_k-1)$$

对所有的 θ_k 和 λ 求偏导为 0：

$$\frac{\partial l}{\partial\theta_k}=\frac{N_k}{\theta_k}+\lambda=0 \quad k=1,2,...K$$

$$\frac{\partial l}{\partial \lambda} = \sum_{k=1}^{K} \theta_k - 1 = 0$$

联立方程式可得 $\lambda = -N$，然后代入第一个式子得到：

$$\theta_k = \frac{N_k}{N} = \frac{\sum_{i=1}^{N} I(y^{(i)} = c_k)}{N}$$

下面进一步求解 $P(X_j = x_j|Y = c_k)$，这里就不做极大似然估计的推导了，直接给出公式。我们设样本第 j 个特征 x_j 可能取值的集合为 $\{a_{j1}, a_{j2}, ..., a_{jS_j}\}$。

$$P(X_j = a_{jl}|Y = c_k) = \frac{\sum_{i=1}^{N} I(x_j^{(i)} = a_{jl}, y^{(i)} = c_k)}{\sum_{i=1}^{N} I(y^{(i)} = c_k)} \tag{6}$$

$$j = 1, 2, ..., N; \quad l = 1, 2, ..., S_j; \quad k = 1, 2, ..., K$$

最后，将 (5)(6) 代入 (4) 即可得到样本 x 的类别。

1.3 参数估计（贝叶斯估计）

用极大似然估计可能会出现所要估计的概率值为 0 的情况，这样会影响后验概率的计算结果造成偏差。解决这一问题的方法是采用贝叶斯估计，这里不作证明，直接给出公式：

$$P(X_j = a_{jl}|Y = c_k) = \frac{\sum_{i=1}^{N} I(x_j^{(i)} = a_{jl}, y^{(i)} = c_k) + \lambda}{\sum_{i=1}^{N} I(y^{(i)} = c_k) + S_j\lambda}$$

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

上面的两个式子中 $\lambda \geqslant 0$，当 $\lambda = 0$ 就是极大似然估计，当取 $\lambda = 1$ 时称为拉普拉斯平滑，这样就可以满足概率值都大于 0。

1.4 损失函数

朴素贝叶斯将实例分到后验概率最大的类中，这等价于期望风险最小化。假设选择 $0-1$ 损失函数，训练集中某个样本 x，其预测结果是 $f(x)$，真实结果是 y，损失函数如下：

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

$f(x)$ 是分类决策函数，这时期望风险函数为 $E[L(y,f(x))]$，又因为期望是对联合概率 $P(x,y)$ 取的，且 y 的可能取值有 k 种，所以条件期望为 $E[\sum_{k=1}^{K} L(y,f(x))P(Y=c_k|X=x)]$。使期望风险最小化就只需对所有样本逐个进行最小化，由此得到：

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

由此可知，期望风险最小化准则得到了后验概率最大化准则，即朴素贝叶斯采用的原理。

2 参数估计方法

在机器学习中最常用的参数估计方法有极大似然估计，极大后验估计和贝叶斯估计。我们以本文的朴素贝叶斯公式为例，注意贝叶斯估计和朴素贝叶斯法是两个东西，不要理解混淆了。下面来解释一下它们的原理，贝叶斯公式这里统一简写成 $P(\theta|X)=\frac{P(X|\theta)P(\theta)}{P(X)}$，$\mathcal{X}$ 表示特征向量的输入空间，$\mathcal{Y}$ 表示类别标记的输出空间。

2.1 极大似然估计

极大似然估计法属于频率学派，他们认为参数是客观存在的，使得样本分布概率最大的参数就是客观存在的值。所以极大似然估计公式如下：

首先求 $L(\theta)$:

$$L(\theta)=\prod_{x\in\mathcal{X}} P(x|\theta)$$

然后取对数求极大：

$$
\begin{aligned}
\theta_{MLE} &= arg \max_{\theta} log \prod_{x\in\mathcal{X}} P(x|\theta) \\
&= arg \max_{\theta} \sum_{x\in\mathcal{X}} logP(x|\theta)
\end{aligned}
$$

2.2 极大后验估计

极大似然估计是在对被估计量没有任何先验知识的前提下求得的，如果已知被估计参数满足某种先验分布，则需要用到极大后验估计。极大后验公式如下：

$$
\begin{aligned}
\theta_{MAP} &= arg\max_{\theta} \frac{P(X|\theta)P(\theta)}{P(X)} \\
&= arg\max_{\theta} P(X|\theta)P(\theta) \\
&= arg\max_{\theta}\{logP(X|\theta) + logP(\theta)\} \\
&= arg\max_{\theta}\{\sum_{x\in\mathcal{X}} logP(x|\theta) + logP(\theta)\}
\end{aligned}
$$

2.3 贝叶斯估计

贝叶斯估计属于贝叶斯学派，他们认为参数是随机的，和一般随机变量没有本质区别。正是因为参数不能固定，当给定一个输入 x 后，就不能用一个确定的 y 表示输出结果，必须用一个概率的方式表达出来，所以贝叶斯估计的输出是一个期望值。

在这里贝叶斯估计其实要解决的不是如何去估计参数，而是如何估计新的测量数据出现的概率，但其过程并不需要要计算参数 θ，而是通过对 θ 的积分得出，可以写成如下形式。在这里 x 是新的样本。

$$
P(x|\theta) = \int_{\theta\sim N(\mu,\sigma^2)} P(x|\theta)P(\theta|X)d\theta
$$

写到这里，贝叶斯估计还是感觉理解不够透彻，希望有相关见解的同学请留下你的留言。

3 参考文献

1. 李航博士的统计学习方法

2.http://blog.csdn.net/yangliuy/article/details/8296481

3.http://blog.csdn.net/andyelvis/article/details/42423185

4.http://www.cnblogs.com/xueliangliu/archive/2012/08/02/2962161.html