机器学习常见知识汇总之贝叶斯分类器

田晓彬
xiaobin9652@163.com

贝叶斯决策论

首先我们来回忆下经典的贝叶斯公式。假设有N种可能的类别标记，$\boldsymbol{y}=\{c_1,c_2,....c_N\}$，每个样本$\boldsymbol{x}$属于第$c$类的后验概率为：

$$P(c|\boldsymbol{x})=\frac{P(\boldsymbol{x}|c)P(c)}{P(\boldsymbol{x})}$$

其中$P(c)$为类先验概率，表示类别c出现的概率；$P(\boldsymbol{x}|c)$为类条件概率，表示在类别c下$\boldsymbol{x}$出现的概率；$P(c|\boldsymbol{x})$为后验概率，表示数据$\boldsymbol{x}$属于类c的概率；$P(\boldsymbol{x})$是用于归一化的因子，给定数据$\boldsymbol{x}$,$P(\boldsymbol{x})$和类标记无关。有了后验概率，我们就可以对样本数据进行分类。后验概率越大，说明该样本属于这一类别的可能性越大，我们就越有把握地将这个数据归到这个类别下面。

若想在有限的数据中准确地估计出后验概率，大致有两种策略：判别式模型，基于$\boldsymbol{x}$直接建模$P(c|\boldsymbol{x})$来预测c；生成式模型，通过贝叶斯公式求得$P(c|\boldsymbol{x})$来预测c。

对于生成式模型，在使用贝叶斯公式时需要首先得到类先验概率$P(c)$和类条件概率$P(c|\boldsymbol{x})$。但在实际问题中，通常只能获得有限数目的样本数据，这样类先验概率$P(c)$和类条件概率$P(c|\boldsymbol{x})$都是未知的。

类先验概率$P(c)$表达了样本空间中各类样本所占的比例，当数据集包含充足的独立同分布的样本时，$P(c)$可以通过各类样本出现的频率来进行估计。

而类条件概率$P(c|\boldsymbol{x})$的估计是非常困难的，而将未知的类条件概率估计问题转化成估计参数问题可以很好的解决这个问题。极大似然估计就是一种常用的参数估计方法。

极大似然估计

估计类条件概率的一种常见策略就是先假设其具有某种确定的概率分布形式，再基于训练样本对概率分布的参数进行估计。概率模型的训练过程就是 基于样本数据来进行参数估计的过程，极大似然估计就是通过优化似然函数来确定参数值的方法，该方法根据数据采样来估计概率分布参数。简单来说，最

大似然估计就是利用已知的样本数据，来反推最有可能导致这种结果的参数值。

令D_c表示训练集D中第c类样本组成的集合，假设这些样本是独立同分布的，则参数$\boldsymbol{\theta_c}$对于数据集D_c的似然是

$$P(D_c|\boldsymbol{\theta}_c) = \prod_{\boldsymbol{x}\in D_c} P(\boldsymbol{x}|\boldsymbol{\theta_c}) \tag{1}$$

对$\boldsymbol{\theta_c}$进行极大似然估计，就是找到可以使$P(D_c|\boldsymbol{\theta}_c)$最大化的参数值$\hat{\boldsymbol{\theta}}_c$。直观上来看，最大似然估计是试图在所有$\boldsymbol{\theta}_c$的取值中，找到一个可以使数据出现的可能性最大的值。

式子(1)中的连乘容易造成计算结果下溢，所以对式子(1)取对数转化成对数似然

$$\begin{aligned} LL(\boldsymbol{\theta}_c) &= \log P(D_c|\boldsymbol{\theta}_c) \\ &= \sum_{\boldsymbol{x}\in D_c} \log P(\boldsymbol{x}|\boldsymbol{\theta_c}) \end{aligned} \tag{2}$$

通过对此对数似然求一阶导，可以得到参数$\boldsymbol{\theta}_c$的极大似然估计$\hat{\boldsymbol{\theta}}_c$为

$$\hat{\boldsymbol{\theta}}_c = \arg\max_{\boldsymbol{\theta}_c} LL(\boldsymbol{\theta}_c)$$

通过极大似然估计求得的解只是一个估计值，只有在样本数量足够多且每个样本都满足独立同分布时，这个值才会接近真实值。

朴素贝叶斯算法

根据贝叶斯决策论可以发现，类条件概率$p(\boldsymbol{c}|\boldsymbol{x})$的估计是非常困难的。为了解决这个问题，朴素贝叶斯分类器采用了“属性条件独立假设”，对于已知类别，假设数据的所有属性互相独立。即每个属性独立地对分类结果产生影响。

基于属性条件独立假设，可以将贝叶斯公式重写为：

$$P(c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|c)P(c)}{P(\boldsymbol{x})} = \frac{P(c)}{P(\boldsymbol{x})}\prod_{i=1}^{d} P(x_i|c) \tag{1}$$

其中d为属性数目，x_i为$\boldsymbol{x}$在第i个属性上的取值。

$P(\boldsymbol{x})$与类别无关，所以根据式子(1)可以得出

$$f(\boldsymbol{x}) = \arg\max_{c\in \boldsymbol{y}} P(c)\prod_{i=1}^{d} P(x_i|c) \tag{2}$$

这就是朴素贝叶斯模型的表达式。

首先，基于数据集D来估计先验概率$P(c)$。令D_c表示数据集D中第c类样本组成的集合，则

$$P(c) = \frac{|D_c|}{|D|} \tag{3}$$

对于离散属性来说，

$$P(x_i|c) = \frac{|D_{c,x_i}|}{|D_c|} \tag{4}$$

其中D_{c,x_i}表示D_c中在第i个属性上取值为x_i的样本组成的集合。

对于连续属性，可以考虑使用概率密度函数，假定$P(x_i|c)\ N(\mu_{c,i}, \sigma_{c,i}^2)$,

$$P(x_i|c) = \frac{1}{\sqrt{2\pi}\sigma_{c,i}} \exp\left(-\frac{(x_i - \mu_{c,i})^2}{2\sigma_{c,i}^2}\right) \tag{5}$$

其中$\mu_{c,i}$和$\sigma_{c,i}^2$分别是第c类样本在第i个属性上取值的均值和方差。

EM算法

在实际的应用中，通常会遇到包含隐变量的数据集。当遇到这种数据集时，若想对参数θ做极大似然估计，则应该最大化

$$LL(\theta|\boldsymbol{X}, \boldsymbol{Z}) \tag{1}$$

其中$\boldsymbol{Z}$表示隐变量，θ表示模型参数。

观察式子(1)可以发现，由于隐变量$\boldsymbol{Z}$的存在，我们无法直接对其进行极大似然估计。期望最大化（EM）算法是常用的估计参数和隐变量的算法。它是一种迭代式的方法，基本思想是：若参数θ一致，则可以根据数据样本推断出最优隐变量$\boldsymbol{Z}$的值；反之，若$\boldsymbol{Z}$的值已知，则可以对参数θ做极大似然估计。

假定初始参数值为θ^0，可迭代执行以下步骤直至收敛：基于θ^t推测隐变量$\boldsymbol{Z}$的期望值，记为$\boldsymbol{Z^t}$；基于数据$\boldsymbol{X}$和隐变量$\boldsymbol{Z}$对参数θ做极大似然估计，记为θ^{t+1}。

进一步，若我们希望基于θ计算隐变量$\boldsymbol{Z}$的概率分布$p(\boldsymbol{Z}|\boldsymbol{X}, \theta^t)$，则EM算法可以描述为：

E步：以当前参数θ^t推断隐变量分布$p(\boldsymbol{Z}|\boldsymbol{X}, \theta^t)$，并计算对数似然$LL(\theta|\boldsymbol{X}, \boldsymbol{Z})$关于$\boldsymbol{Z}$的期望

$$Q(\theta|\theta^t) = E_{\boldsymbol{Z}|\boldsymbol{X},\theta^t} LL(\theta|\boldsymbol{X}, \boldsymbol{Z})$$

M步：根据$\boldsymbol{X}$和隐变量$\boldsymbol{Z}$计算参数最大化期望似然,即

$$\theta^{t+1} = \arg\max Q(\theta|\theta^t)$$

简要来说，EM算法使用两个步骤交替优化计算。第一步利用当前估计的参数值来计算对数似然的期望值；第二步寻找能使第一步产生的似然期望最大的参数值。交替优化直至收敛到局部最优。