一、朴素贝叶斯

二. 朴素贝叶斯的优缺点

三. 朴素贝叶斯实战

一、朴素贝叶斯

朴素贝叶斯中的朴素一词的来源就是**假设各特征之间相互独立**。这一假设使得朴素贝叶斯算法变得简单，但有时会牺牲一定的分类准确率。

首先给出贝叶斯公式：

$$P(B|A)=\frac{P(A|B)P(B)}{P(A)}$$

换成分类任务的表达式：

$$\text{p(类别|特征)}=\frac{p(\text{特征}|\text{类别})\text{p(类别)}}{\text{p(特征)}}$$

这是朴素贝叶斯法分类的基本公式．于是，朴素贝叶斯分类器可表示为

$$y=f(x)=\arg\max_{c_k}\frac{P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)}{\sum_k P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k)} \tag{4.6}$$

注意到，在式 (4.6) 中分母对所有 c_k 都是相同的，所以，

$$y=\arg\max_{c_k} P(Y=c_k)\prod_j P(X^{(j)}=x^{(j)}\mid Y=c_k) \tag{4.7}$$

因为它假定所有的特征在数据集中的作用是同样重要和独立的。正如我们所知，这个假设在现实世界中是很不真实的，因此，说朴素贝叶斯真的很“朴素”。

朴素贝叶斯模型(Naive Bayesian Model)的朴素(Naive)的含

义是"很简单很天真"地假设样本特征彼此独立. 这个假设现实中基本上不存在，但特征相关性很小的实际情况还是很多的，所以这个模型仍然能够工作得很好。

二. 朴素贝叶斯的优缺点

优点：

（1） 算法逻辑简单,易于实现（算法思路很简单，只要使用贝叶斯公式转化即可！）

（2）分类过程中时空开销小（假设特征相互独立，只会涉及到二维存储）

缺点：

朴素贝叶斯假设属性之间相互独立，这种假设在实际过程中往往是不成立的。在属性之间相关性越大，分类误差也就越大。

三. 朴素贝叶斯实战

sklearn中有3种不同类型的朴素贝叶斯：

高斯分布型：用于classification问题，假定属性/特征服从正态分布的。

多项式型：用于离散值模型里。比如文本分类问题里面我们提到过，我们不光看词语是否在文本中出现，也得看出现次数。如果总词数为n，出现词数为m的话，有点像掷骰子n次出现m次这个词的场景。

伯努利型：最后得到的特征只有0(没出现)和1(出现过)。

朴素贝叶斯

优点：在数据较少的情况下仍然有效，可以处理多类别问题

缺点：对于输入数据的准备方式较为敏感

适用数据类型：标称型数据

核心：贝叶斯定理

”朴素“：属性条件独立性假设，所有属性相互独立

$$P(c \mid \boldsymbol{x}) = \frac{P(\boldsymbol{x}, c)}{P(\boldsymbol{x})} . \tag{7.7}$$

基于贝叶斯定理, $P(c \mid \boldsymbol{x})$可写为

$$P(c \mid \boldsymbol{x}) = \frac{P(c)\, P(\boldsymbol{x} \mid c)}{P(\boldsymbol{x})} , \tag{7.8}$$

其中, $P(c)$ 是类“先验”(prior)概率; $P(\boldsymbol{x} \mid c)$ 是样本 $\boldsymbol{x}$ 相对于类标记 c 的类条件概率(class-conditional probability), 或称为“似然”(likelihood); $P(\boldsymbol{x})$ 是用于归一化的“证据”(evidence)因子. 对给定样本 $\boldsymbol{x}$, 证据因子 $P(\boldsymbol{x})$ 与类标记无关, 因此估计 $P(c \mid \boldsymbol{x})$ 的问题就转化为如何基于训练数据 D 来估计先验 $P(c)$ 和似然 $P(\boldsymbol{x} \mid c)$.

基于属性条件独立性假设, 式(7.8)可重写为

$$P(c \mid \boldsymbol{x}) = \frac{P(c)\, P(\boldsymbol{x} \mid c)}{P(\boldsymbol{x})} = \frac{P(c)}{P(\boldsymbol{x})} \prod_{i=1}^{d} P(x_i \mid c) , \tag{7.14}$$

参数估计：极大似然估计和贝叶斯估计（多使用极大似然估计）

估计类条件概率的一种常用策略是先假定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. 具体地, 记关于类别 c 的类条件概率为 $P(\boldsymbol{x} \mid c)$, 假设 $P(\boldsymbol{x} \mid c)$ 具有确定的形式并且被参数向量 $\boldsymbol{\theta}_c$ 唯一确定, 则我们的任务就是利用训练集 D 估计参数 $\boldsymbol{\theta}_c$. 为明确起见, 我们将 $P(\boldsymbol{x} \mid c)$ 记为 $P(\boldsymbol{x} \mid \boldsymbol{\theta}_c)$.

ps:朴素贝叶斯实现分类代码参考

https://blog.csdn.net/chuhang_zhqr/article/details/50755369

https://www.cnblogs.com/zy230530/p/6847243.html