第六章 朴素贝叶斯

一、朴素贝叶斯

1. 朴素贝叶斯基本思想
   例如说根据邮件内容猜测是否为垃圾邮件，邮件X中有若干单词W，我们可以假设

$$P(X=x|Y=y) \approx \prod_{i=1}^{n} P(W=w_i|Y=y)$$

   假设我们想要用X来分类，其中Y是类标号
   使用贝叶斯法则来估计

$$P(Y=y|X=x) = \frac{P(X=x|Y=y)P(Y=y)}{P(X=x)}$$

   这里因为x已经给定，所以这个样本x出现的总体概率已经该固定，我们目前要做的是在所有的y中找到可以使得后验概率最大的y
   最后得到最终决策

$$\hat{y} = arg\,max_y\, P(Y=y)\prod_{i=1}^{n} P(X=x_i|Y=y_i)$$

2. **朴素贝叶斯的假设**
   - 特征独立：数据的特征是在给定类标号下相互条件独立的
   - 特征同等重要：所有特征都被假定对类别标签的预测具有同等的贡献作用。

二、数值属性

假设服从正态分布，或者叫高斯分布

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

对于连续型数据，我们可以将这个拟合为一个高斯分布，对于每一个数值类型属性，计算每个类别平均值和标准差，可以得到每个属性的概率分布
随后可以将属性值带入概率函数得到对应概率

三、零概率

假设测试的某个特征在样本里从来没出现过，那么它的概率就是0，放在连乘里面得到结果必然是0，这样的结果显然是不合适的，这个零概率对其他的概率具有否决权
解决方法——拉普拉斯平滑
该数据集规模足够大，因此添加每一类中的一行数据都不会对估计的概率值产生影响。这

将解决概率值变为零的问题。

$$P(c) = \frac{|D_c| + 1}{|D| + N}$$

$$P(x_i|c) = \frac{|D_{c,x_i}| + 1}{|D_c| + N_i}$$

其中N是类别的数量
这将导致从这些类别中去除所有零值，同时也不会影响这些类别的总体相对频率

四、总结

1. **朴素贝叶斯的假设**
   - **特征独立性**：在已知类别标签的情况下，数据的各个特征是相互独立的。
   - **特征同样重要**：所有特征都被假定对类别标签的预测具有同等的贡献作用。
   - 连续特征呈正态分布：如果一个特征是连续型的，那么就假定它在每个类别中都呈正态分布。
   - 离散特征具有多项式分布的特征：如果一个特征是离散的，那么就假定它在每个类别中都具有多项式分布。
     我们大量使用朴素贝叶斯算法，尽管我们知道它存在缺陷，但它为我们提供了计算效率极高的算法，而且在实际应用中表现得非常出色。
2. 朴素贝叶斯的优点和缺点

优点	缺点
实现简单，并且计算有效率	假设特征之间独立，这与真实世界不一定一致
在存在分类情况下表现良好	在训练数据中没有出现相关情况时，它可能会难以准确估算概率（即零频问题）。
在大量特征情况下表现有效率	对于数字特征数据，假设来自正态分布