实验十一：朴素贝叶斯分类

本次实验旨在让同学们了解并掌握朴素贝叶斯算法的原理和应用。

内容

1. 朴素贝叶斯

1. 朴素贝叶斯法是典型的生成学习方法。生成方法由训练数据学习联合概率分布$P(X,Y)$，然后求得后验概率分布$P(Y|X)$。具体来说，利用训练数据学习$P(X|Y)$和$P(Y)$的估计，得到联合概率分布：

$$P(X,Y) = P(Y)P(X|Y)$$

概率估计方法可以是极大似然估计或贝叶斯估计。

2. 朴素贝叶斯法的基本假设是条件独立性，

$$\begin{aligned} P(X=x|Y=c_k) &= P\left(X^{(1)}=x^{(1)},\cdots,X^{(n)}=x^{(n)}|Y=c_k\right) \\ &= \prod_{j=1}^{n} P\left(X^{(j)}=x^{(j)}|Y=c_k\right) \end{aligned}$$

这是一个较强的假设。由于这一假设，模型包含的条件概率的数量大为减少，朴素贝叶斯法的学习与预测大为简化。因而朴素贝叶斯法高效，且易于实现。其缺点是分类的性能不一定很高。

3. 朴素贝叶斯法利用贝叶斯定理与学到的联合概率模型进行分类预测。它的思想基础是这样的：对于给出的待分类项，求解在此项出现的条件下各个类别出现的概率，哪个最大，就认为此待分类项属于哪个类别。

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

将输入x分到后验概率最大的类y。

$$y = \arg\max_{c_k} P\left(Y=c_k\right) \prod_{j=1}^{n} P\left(X_j=x^{(j)}|Y=c_k\right)$$

后验概率最大等价于0-1损失函数时的期望风险最小化。

2. 朴素贝叶斯高斯模型

如果特征是**连续型数据**，推荐使用高斯模型来实现，高斯模型即正态分布。当特征是连续变量的时候，运用多项式模型就会导致很多误差，此时即使做平滑，所得到的条件概率也难以描述真实情况。所以处理连续的特征变量，应该采用高斯模型。

概率密度函数：

$$P\left(X_j = x^{(j)}|Y = c_k\right) = \frac{1}{\sqrt{2\pi\sigma^2}}exp(-\frac{(x^{(i)} - \mu)^2}{2\sigma^2})$$

数学期望(mean)：μ

方差：$\sigma^2 = \frac{\sum(X-\mu)^2}{N}$

可以使用`scipy.stats.norm`中的`pdf()`实现。`scipy.stats.norm`模块是`scipy`库中用于正态分布的模块，它提供了统计数据和一些基本操作的计算，例如概率密度函数（PDF）、累积分布函数（CDF）和反函数。官方文档