实验一 贝叶斯分类和k-近邻学习

1. 实验目的
2. 理解机器学习基本原理和实现过程；
3. 初步学会数据预处理、数据划分方法；
4. 能用python 实现朴素贝叶斯分类和k-近邻学习基本算法
5. 学会机器学习效果的评价方法
6. 实验设备

1． PC机或云计算平台

2. 操作系统不限

3. Python 3.x

4. 使用numpy, pandas等基本算法包

1. 实验原理
2. 数据准备

数据集：复旦新闻语料库（fudancorpus）

预处理：（1）对语料库进行分词、去停用词形成字典；

（2）构建（one-hot）文本向量

（3）划分训练集和测试集

2. 学习模型构建

朴素贝叶斯分类思想及数学原理

k-近邻学习基本算法

1. 实验过程记录

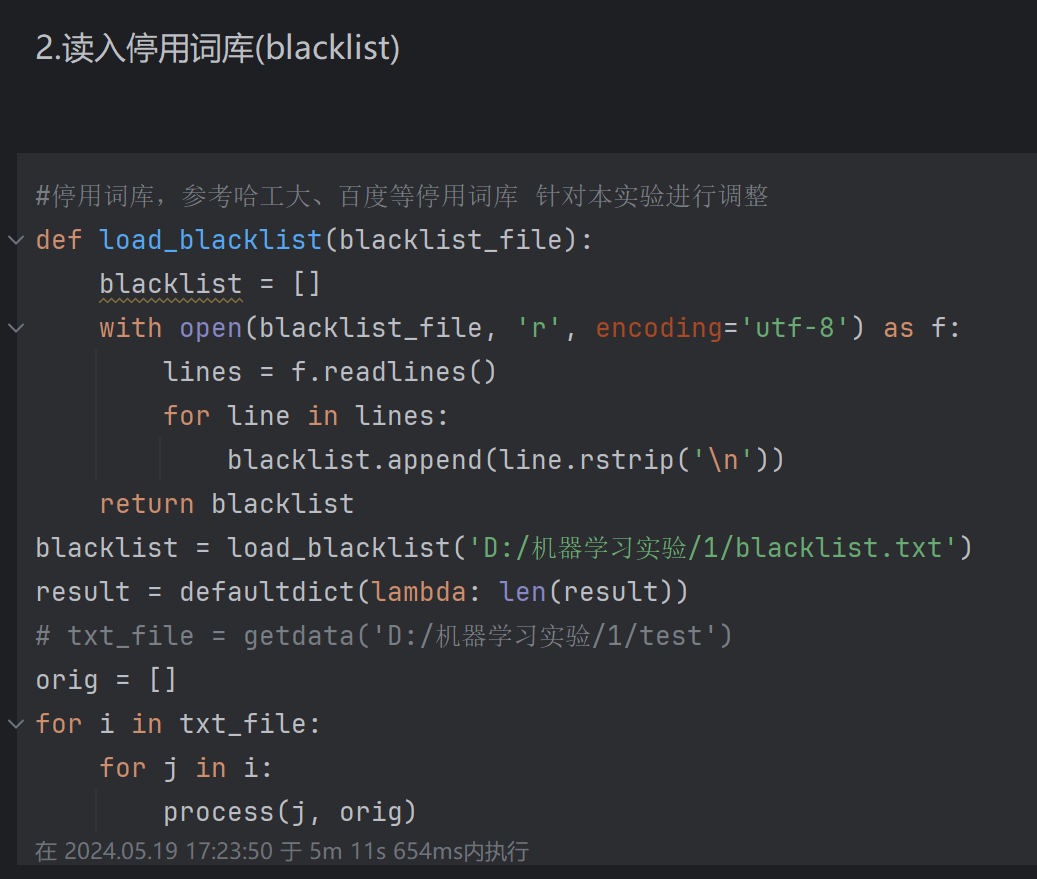
描述程序设计思想、主要代码等



1. 导入语料库并进行处理，通过jieba库分词，利用黑名单去停用词



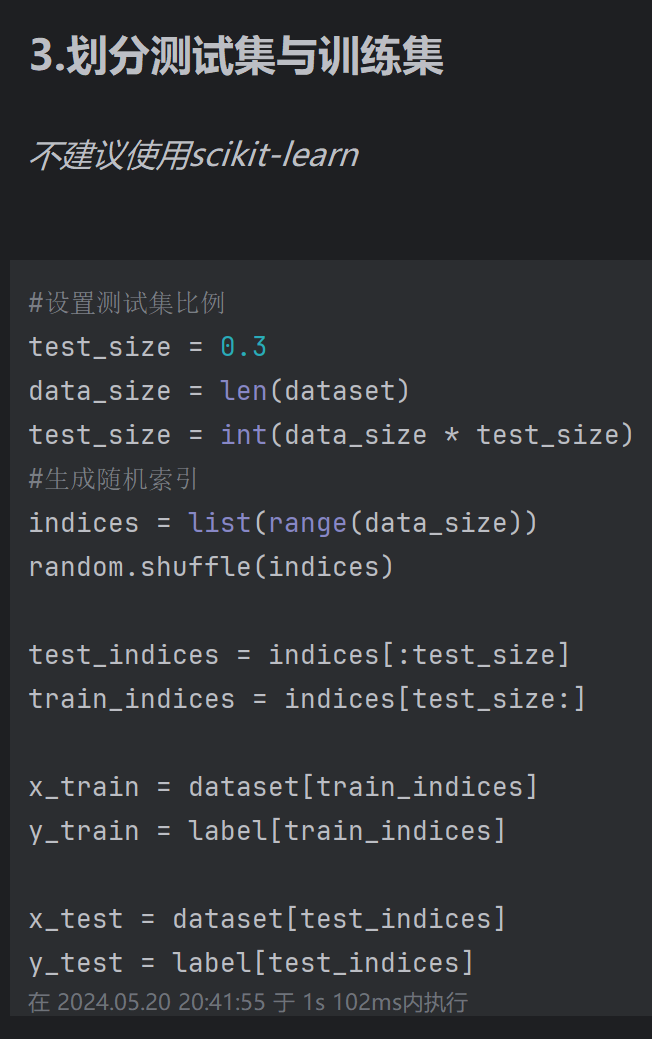




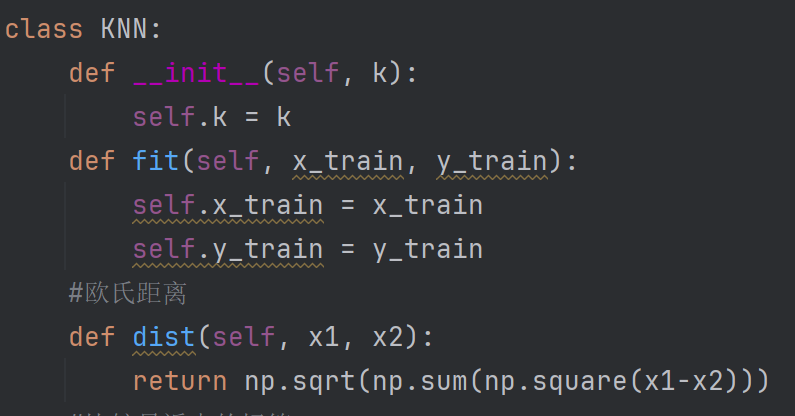
1. 构建onehot文本向量，这种向量便于计算欧氏距离（knn算法）。



1. 划分测试集与训练集

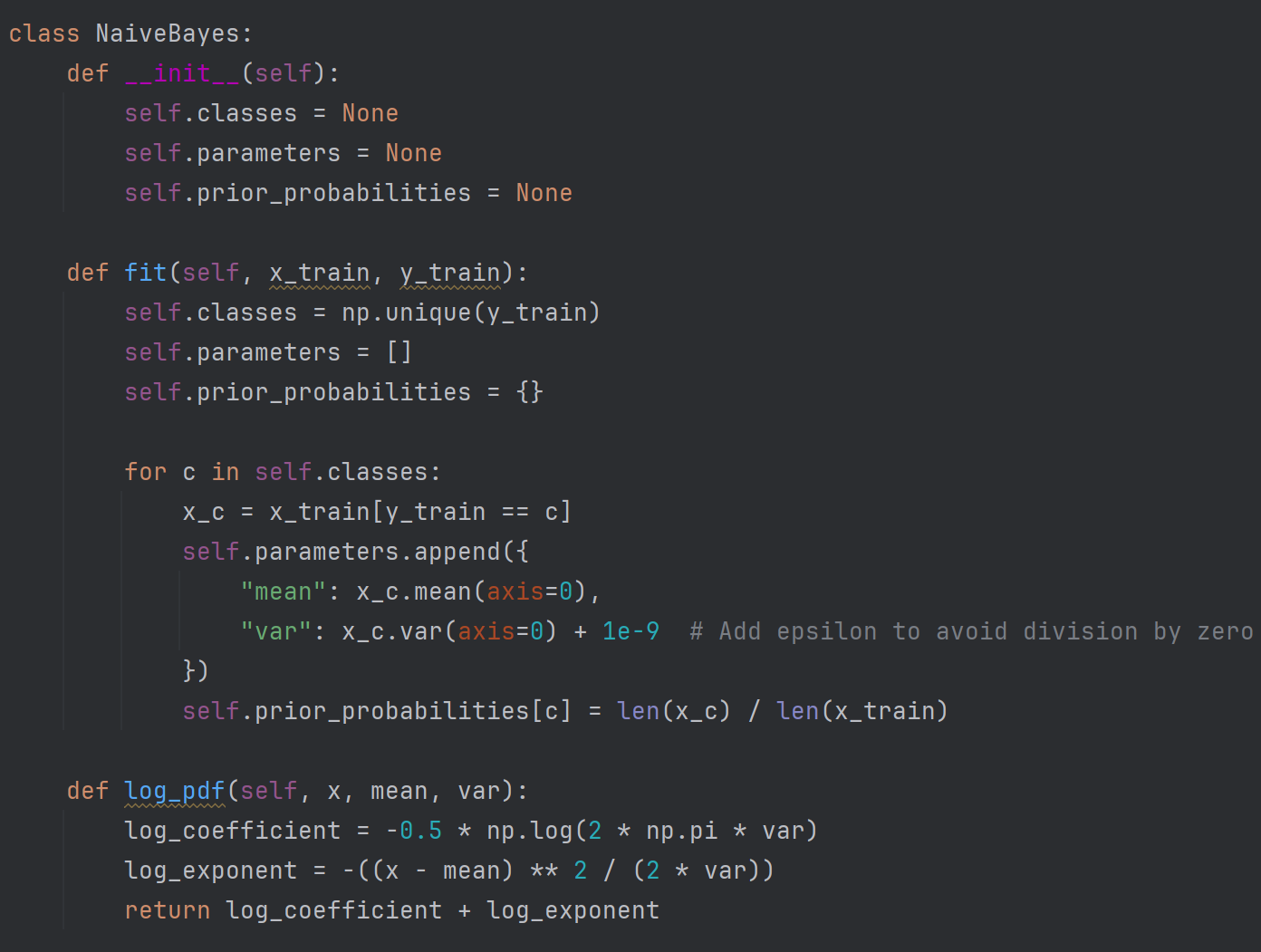


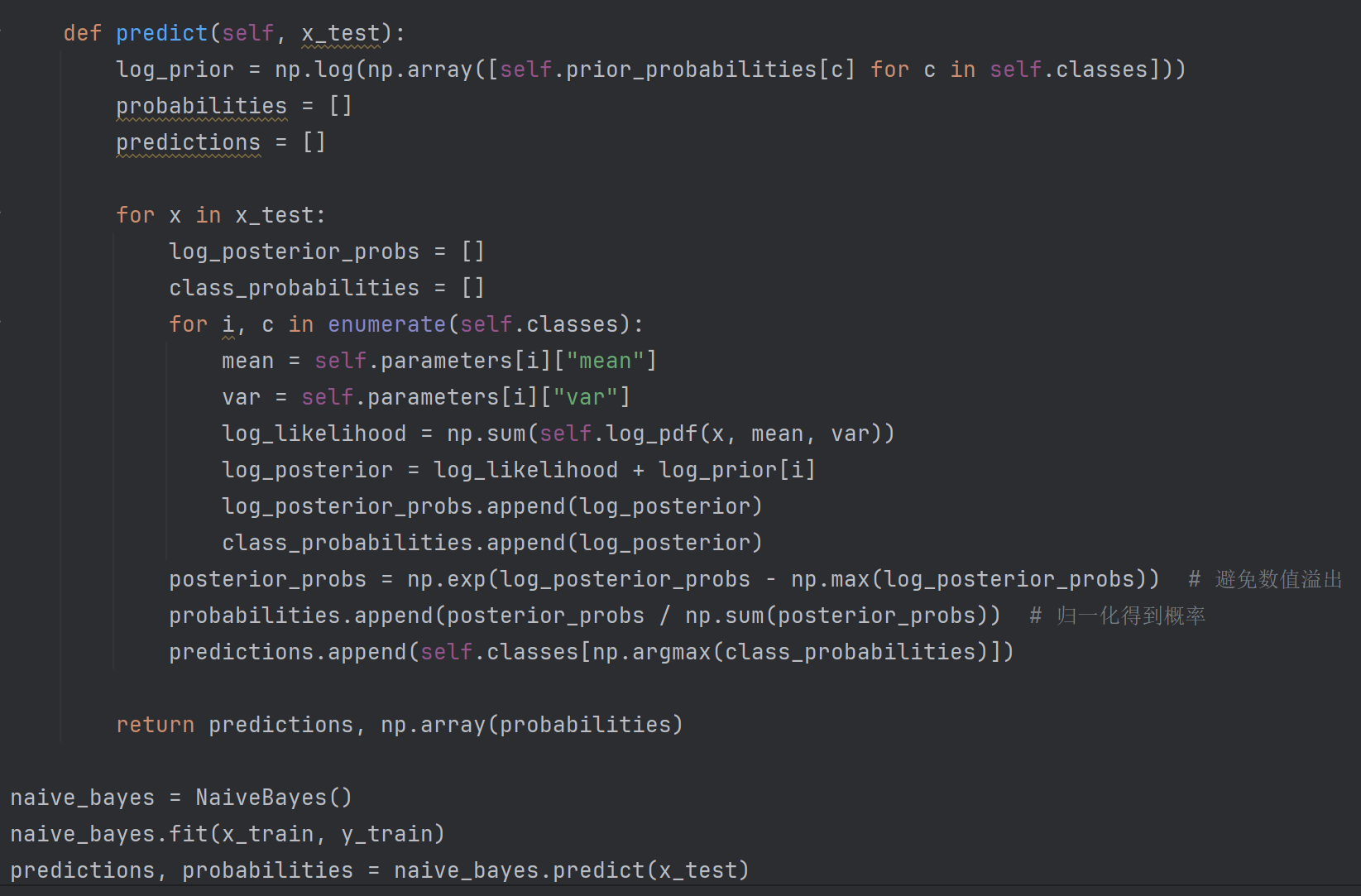
1. 使用knn算法进行训练。Knn算法通过存入训练集，对输入的测试集计算欧氏距离，筛选出k个距离最近的点，通过判断周围点标签的数量预测输入测试点的标签。





1. 使用朴素贝叶斯方法进行训练。朴素贝叶斯法通过计算条件概率判断测试集向量的标签。

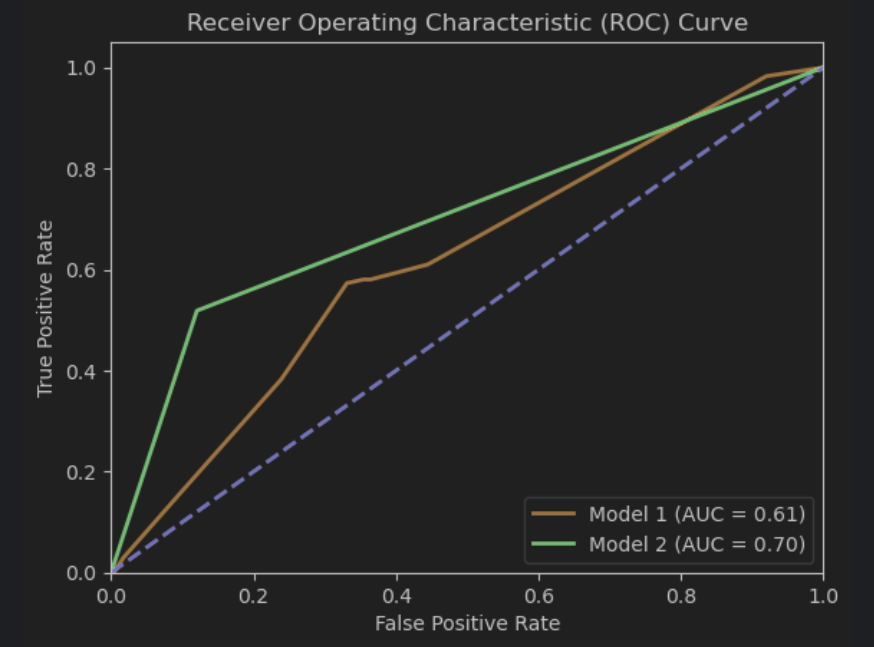
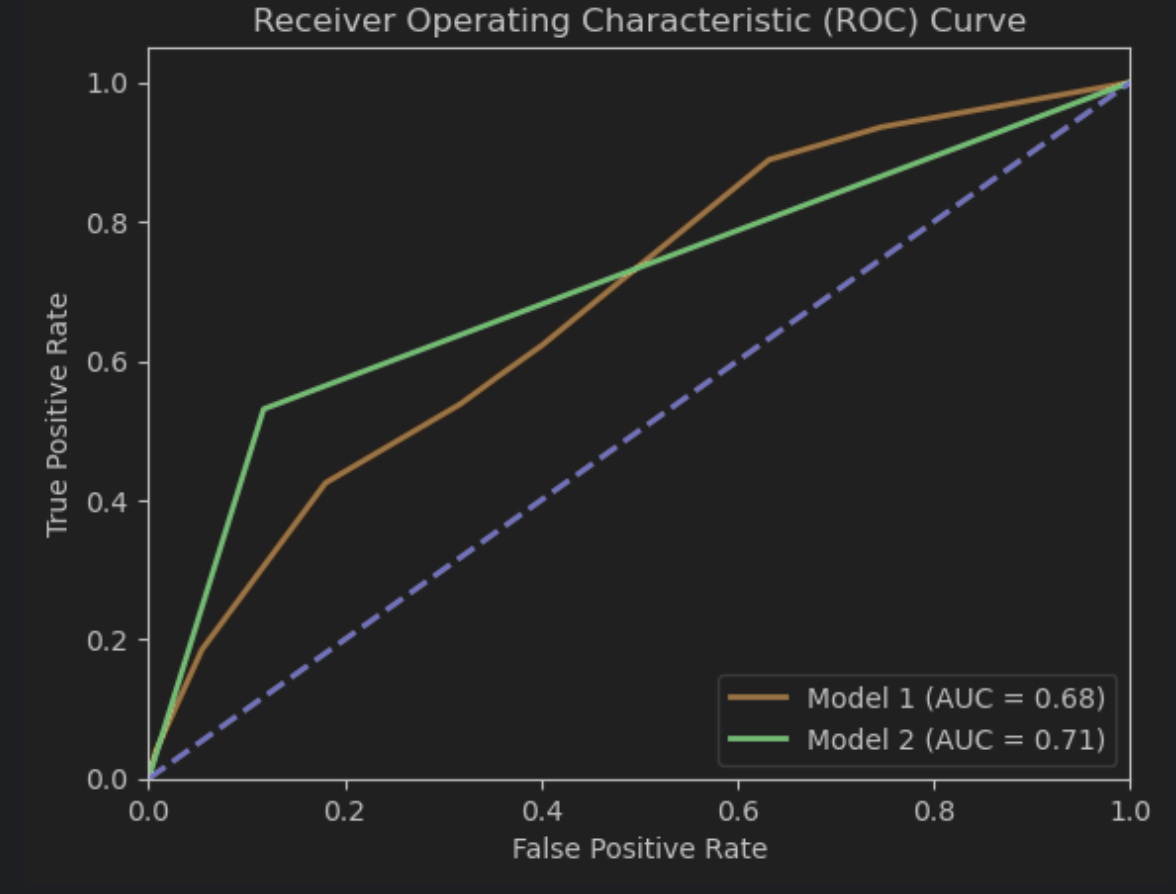
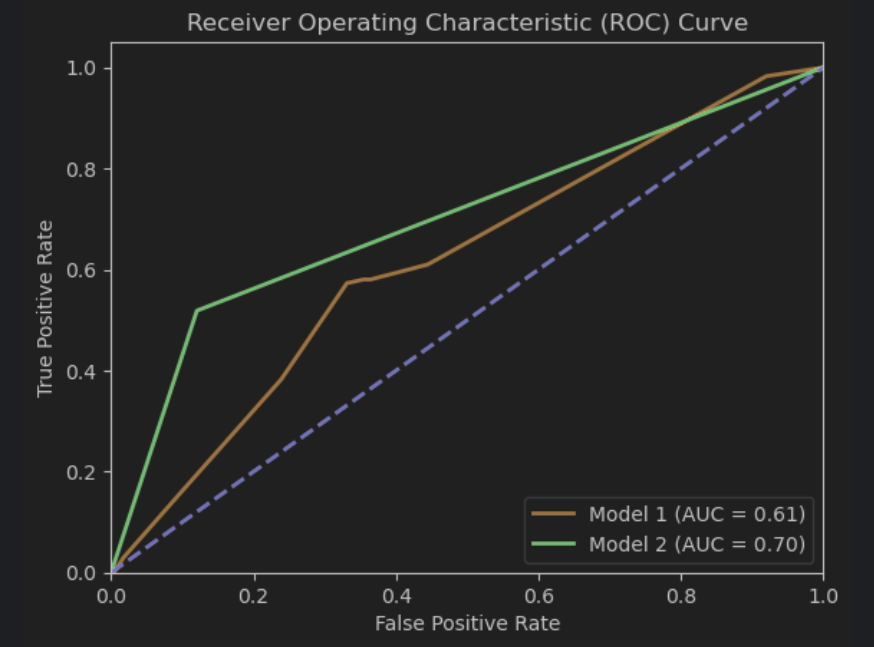




1. 结果与分析

至少做三组实验对比两种方法的P、R、F1（以上三个指标为表格）及ROC曲线，对实验效果进行评价

|  |  |  |  |
| --- | --- | --- | --- |
|  | Precision | Recall | F1 |
| NaiveBayes | 0.7021 | 0.3608 | 0.3694 |
| KNN | 0.1693 | 0.3220 | 0.1576 |
| NaiveBayes | 0.6189 | 0.3914 | 0.4084 |
| KNN | 0.1571 | 0.3203 | 0.1717 |
| NaiveBayes | 0.7149 | 0.3936 | 0.4205 |
| KNN | 0.4119 | 0.2114 | 0.1644 |



通过实验结果可以得出，朴素贝叶斯算法的准确度更高，更加拟合。并且在运行时间方面朴素贝叶斯算法快于KNN

1. 自评

本次实验学到什么？遇到什么困难，又是如何解决的？自己在什么地方得到了提升，又在什么地方需要加强？

本次实验我学会了通过使用python库对语料进行处理编码的方法以及实现实验提到的两种算法的注意事项。这是我第一次对语料库进行处理，开始花费了较多的时间，经过查阅资料等方式掌握了处理的方法。两种算法在实现过程中也有很多的细节，通过查阅资料等方法，修改了之前实现方法中的问题。补充了概率计算的内容用于绘制曲线我对这两种算法有了更深刻的认识。