**实验3：朴素贝叶斯**

1. 使用email文件夹下spam文件夹(垃圾邮件)和ham文件夹（正常邮件）的共50个邮件，采用交叉验证的方式，随机选取10个文件作为测试数据，其他作为训练数据，采用朴素贝叶斯进行垃圾邮件的分类。计算多次迭代后的平均错误率。（本题要求不使用sklearn库）

2.采用sklearn中的MultinomialNB朴素贝叶斯算法对sklearn的datasets中自带的digits数据集进行分类，测试分类效果。

3.对sklearn自带的breast\_cancer数据集，该数据集包含569个病例的数据样本，每个样本具有30个特征值，样本共分为2类:恶性和良性。采用朴素贝叶斯分类器、交叉验证法进行测试，计算训练集准确率（score）、测试集准确率（score）。

**实验4：支持向量机**

4. 使用支持向量机完成sklearn自带的digits数据集进行手写数字的识别。尝试采用不同核函数、调整不同gamma，C参数值等，采用交叉验证法，分别计算训练集错误率和测试集错误率，并进行调参分析。注:可使用sklearn库的SVM实现. 提示：调参可采用网格搜索法GridSearch实现。

5. 在实验3中我们采用了朴素贝叶斯进行了垃圾邮件分类：使用email文件夹下spam文件夹(垃圾邮件)和ham文件夹（正常邮件）的共50个邮件，采用交叉验证的方式，随机选取10个文件作为测试数据，其他作为训练数据，采用朴素贝叶斯进行垃圾邮件的分类。计算多次迭代后的平均错误率。

这里试采用SVM进行垃圾邮件分类，并对两种方法的性能（训练集错误率、测试集错误率、运行时间等）进行比较。注:可以使用sklearn库或python实现

6.\*选做题:

数据集：http://ai.stanford.edu/~amaas/data/sentiment/

为互联网电影评论数据，包含50000个关于电影的正面/负面评论，请采用逻辑回归模型或SVM或其他机器学习模型来拟合数据，并通过交叉验证法评估模型效果。

提示：可以参考网友做法，详见https://www.jianshu.com/p/c9165bfa7619?utm\_campaign=maleskine&utm\_content=note&utm\_medium=seo\_notes&utm\_source=recommendation

大致思路：读取数据、清洗文本、提取词汇、移除停用词、词带模型、数据集划分、逻辑回归模型拟合与评估。

需要使用python自然语言工具包nltk等。