作业五实验报告

赵虎201600301325

## 一．题目要求

使用5-6个模型（可以是knn svm 贝叶斯等等）用投票的方式做一个集成学习对MNIST数据集进行分类。

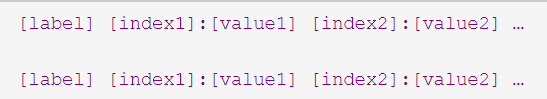
## 二．题目分析

本实验中我才用的模型有线性SVM，非线性（多项式）SVM，KNN算法，朴素贝叶斯估计和神经网络5个模型，根据五个模型的分类结果进行投票，票数多的为最终分类，若票数均一样，则任取一个模型的分类结果。

## 三．题目求解

1.加载数据

libsvm使用的训练数据和检验数据文件格式如下：



每一行第一个元素为标签值，之后为特征点的序号和值。

这里用python程序加载数据：

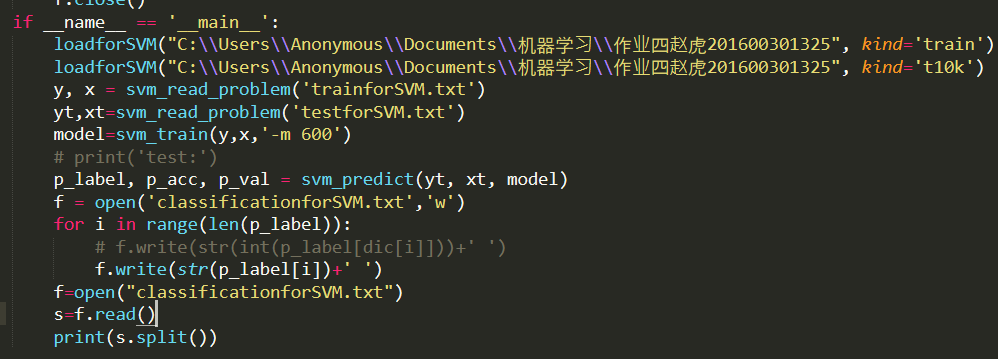


对读入的数据遍历特征值和标签值，将标签按照0-9的顺序排列存储到txt文件中，并将源数据序号和新排列数据的序号对应关系存到一个字典中，方便最终投票时数据绝对顺序不变。

由于每个模型的运行时间较长，所以这里五个模型采用四个程序独立运行，将每个程序的分类结果写入txt中，最后在Ensemble\_Learning类中加载所有结果并投票集成，这样可以极大的加快测试和运行程序的速度。

2.非线性SVM模型

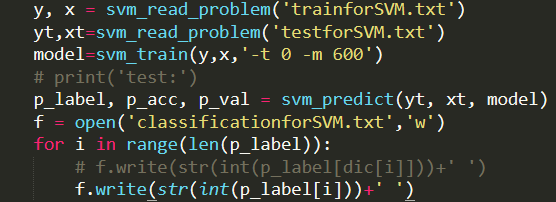
非线性SVM模型调用libSVM的函数库：



由于生成的txt文件较大，在运算时libSVM默认的cache内存大小为40M，这里扩充为600M，最终将生成的结果读入到txt中以便最后的投票。

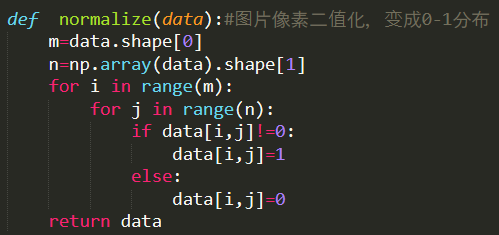
libSVM库中train函数默认采用的是非线性RBF核：exp(-γ\*||u-v||2)，所以这里为默认值。

3.线性SVM模型

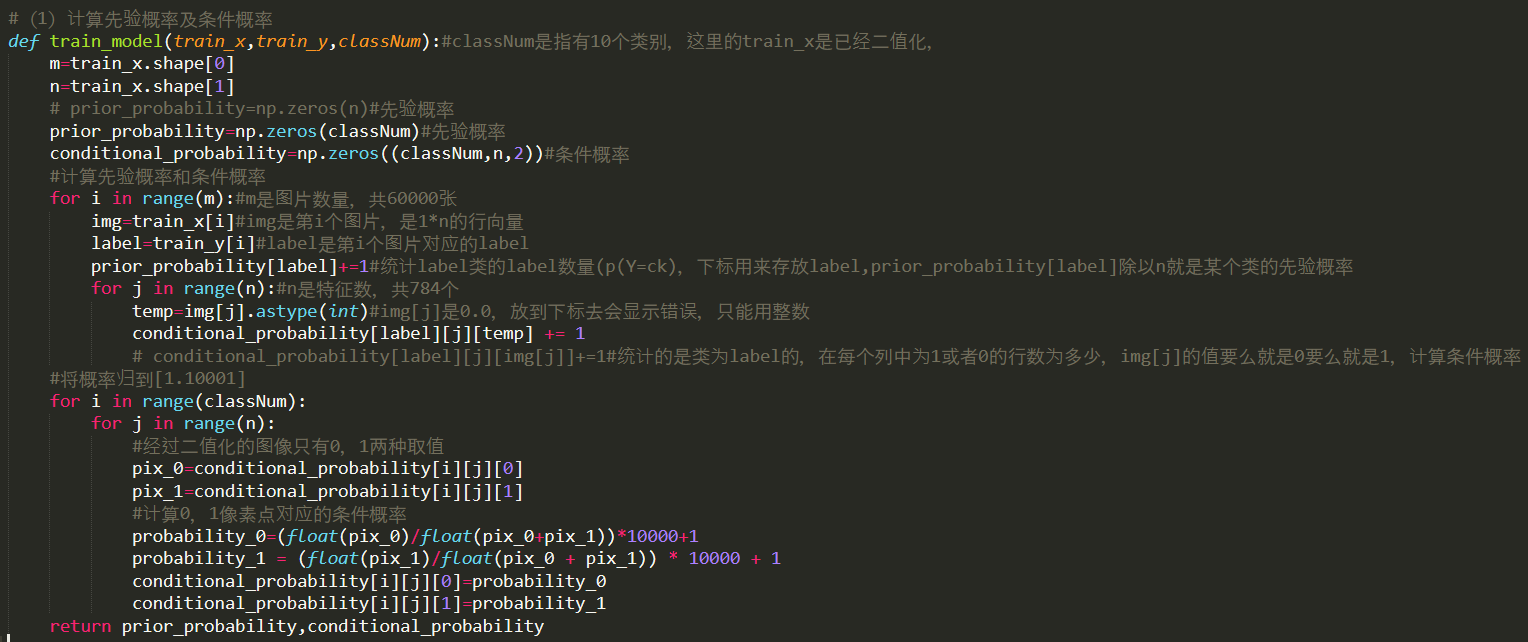


4.朴素贝叶斯估计模型

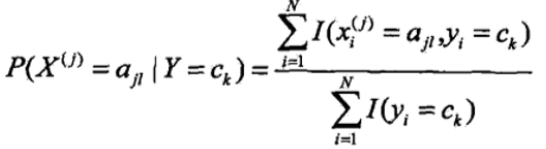
在模型中读入数据后，为了简化计算，对图片进行二值化处理：



根据贝叶斯估计的过程，首先计算先验概率和条件概率：

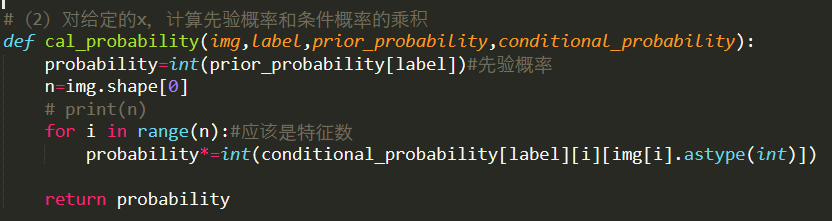


先验概率的计算方法就是每一类图片占全部图片的百分比，根据公式：

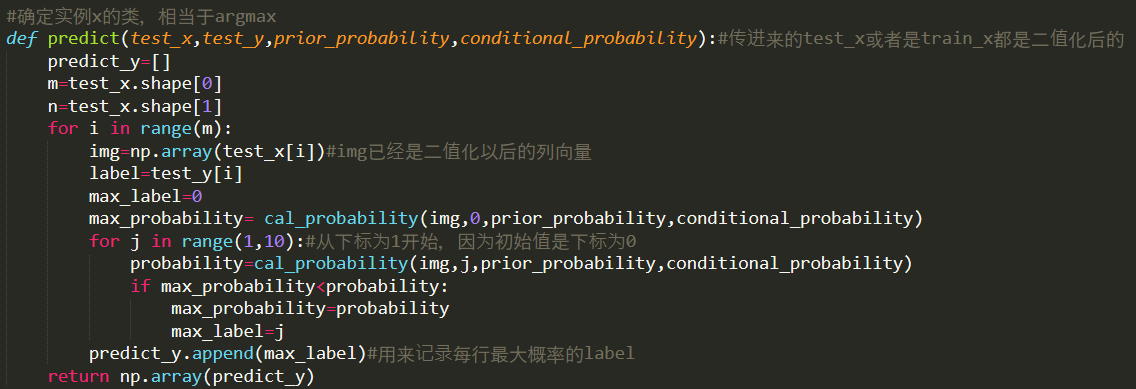


计算出条件概率。

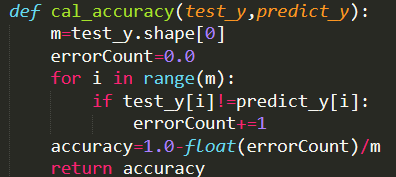
对于给定的类，计算先验概率和条件概率的乘积：



进行预测，概率最大的即为估计的分类：



最终，计算分类准确率，将分类标签存到classificationforBayes.txt中以便最后的投票。



5.KNN算法

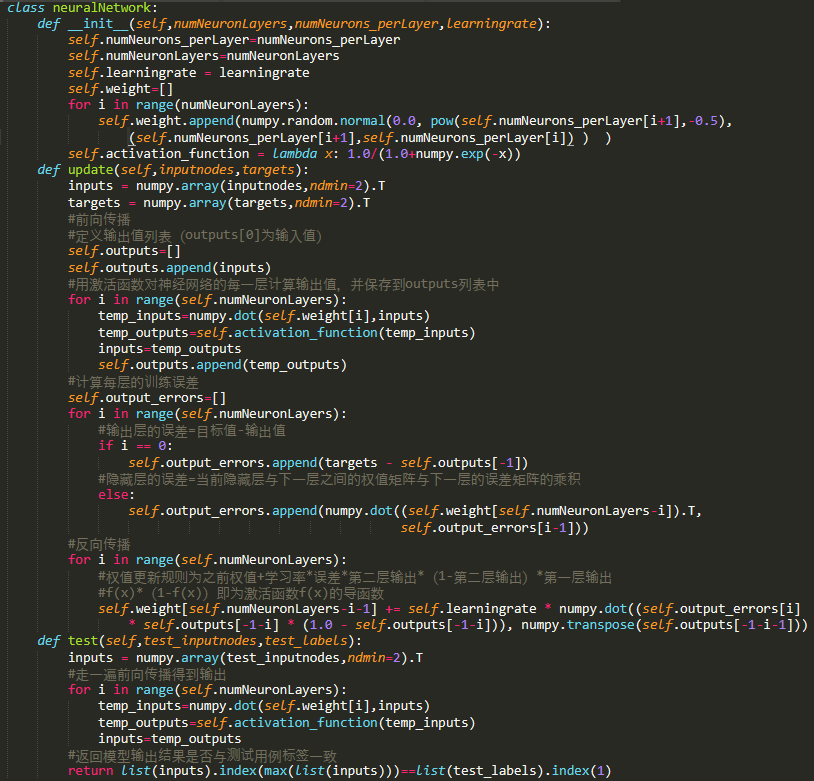
KNN方法的原理这里不再赘述，K这里取30：



同样的，将分类结果存到classificationforKNN.txt中以便最后的投票。

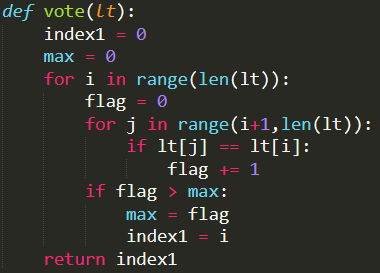
6.神经网络

这里采用上一个实验中的程序，神经网络采用三层结构，隐藏层节点为30，迭代次数为5，学习率为0.1。

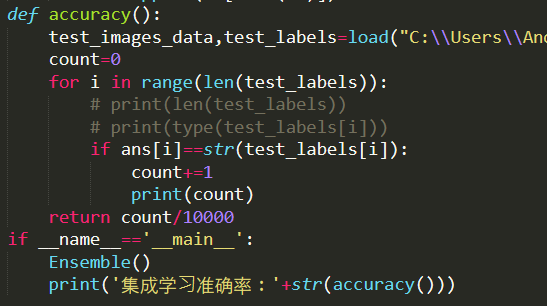


7.集成学习

将之前五个模型的分类结果加载进来，对测试集每一个数据的分类结果进行投票，票数多的为最终分类，若票数均一样，则任取一个模型的分类结果。







根据预期，集成学习的正确率应高于任意一个弱分类器的正确率，线性SVM正确率94.41％，非线性SVM正确率94.46％，贝叶斯估计正确率84.41％，KNN正确率82.26％，神经网络正确率94.76％，最终集成学习的正确率为95.49％。

