人工智能基础

编程作业 2

PB13011058
王悦

实验目的：

本次实验考虑机器学习中传统的监督学习问题与非监督学习，基于两个经典应用：手写数字识别和图片去噪，并结合课上介绍的相应学习算法，在数据集上分别进行实验，以加强对相关算法原理及应用的理解。

Part1.手写数字分类

数据集介绍：

USPS 手写数字识别数据集,我们将对其中的 3 和 8 两个数字进行分类，每张图片表示为一个 16x16 像素的黑白图片，对于每一个像素，用 1 个 8bit 数字(0-255 之间)表示其灰度值。一个 16x16 的图片，总共有 256 个像素，因此对于每张图片，可以用一个 256 个元素的向量表示。而在标记信息中，0 表示当前样本为数字 3，1 表示当前样本为数字 8。

训练与测试

在监督学习中，训练数据带有标号，在训练的过程中需要从训练数据 traindata 和其对应的标号 trainlabel 中学习相应的分类模型。

在测试过程中，用学习到的模型对测试集中的数据 testdata 作预测，并将预测结果与测试数据的真实标签 testlabel 进行比较，从而度量分类模型的性能。

$$\text{Accuracy} = \frac{\sum_{i \in test\ set} I(\text{predict}_i = testlabel_i)}{\#of\ test\ size}$$

实验要求：

1.实现一个朴素贝叶斯分类器

提交一个 Matlab 函数 nbayesclassifier，函数形式为

```
function [ypred, accuracy] = nbayesclassifier (traindata ,trainlabel , testdata, testlabel, threshold )
```

其中 threshold 为用于判断类别的后验概率的阈值，即如果 P(digit=8|x)>threshold 则判别为数字 8。要求函数返回对测试数据的预测 ypred，以及通过与真实标号比较计算得到的分类正确率 accuracy。ypred 与 trainlabel 和 testlabel 形式相同。

2.实现一个最小二乘分类器(引入规范化项后)

1).对引入了 L2 规范化项之后的最小二乘分类问题进行推导。即求解以下优化问题：

$$min_w (X\boldsymbol{w} - y)^2 + \lambda \|\boldsymbol{w}\|^2$$

2).基于 1 中的结果，实现并提交一个 Matlab 函数 lsclassifier

```
function [ypred, accuracy] = lsclassifier (traindata, trainlabel,
```

`testdata, testlabel, lambda)`

3.实现一个支持向量机分类器

提交一个 Matlab 函数 softsvm

`function [ypred,accuracy] = softsvm (traindata, trainlabel, testdata, testlabel , sigma, C)`

其中 C 为 softmarginSVM 的控制参数，sigma 为控制核函数的参数，当 sigma=0 时，使用线性核函数,其他情况则使用 RBF 核函数

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{sigma^2}}$$

4.在不同数据集上使用交叉验证选择各个算法的参数

实现交叉验证（代码需要提交），在各个数据集上：

- 使用 5-fold 交叉验证为每个算法挑选适当的参数(NaïveBayes 中的 threshold，最小二乘法中的 Lambda，SVM 中的 sigma 和 C)；
- 对每一个算法：
  - 返回一个矩阵，表示每一个参数（参数组合）在每一个 fold 上的正确率（若有 10 个参数，则返回 10x5 的矩阵）；
  - 挑选在 5 个 fold 中平均正确率最高的参数（参数组合）

在实验报告中需要记录交叉验证的结果,即对于每个参数(参数组合)在 5 个 fold 上的平均正确率。

5.实验报告

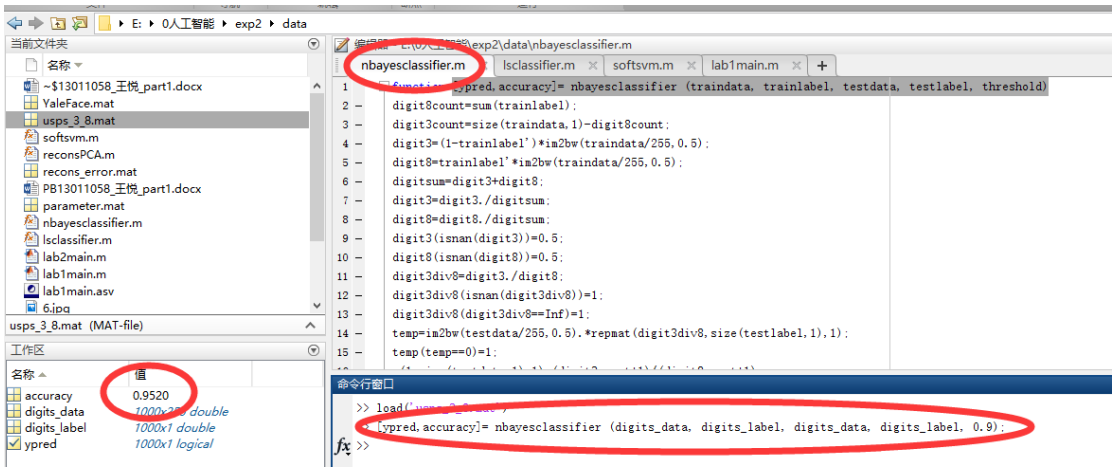
总结以上的实验结果，并对实验结果进行分析。

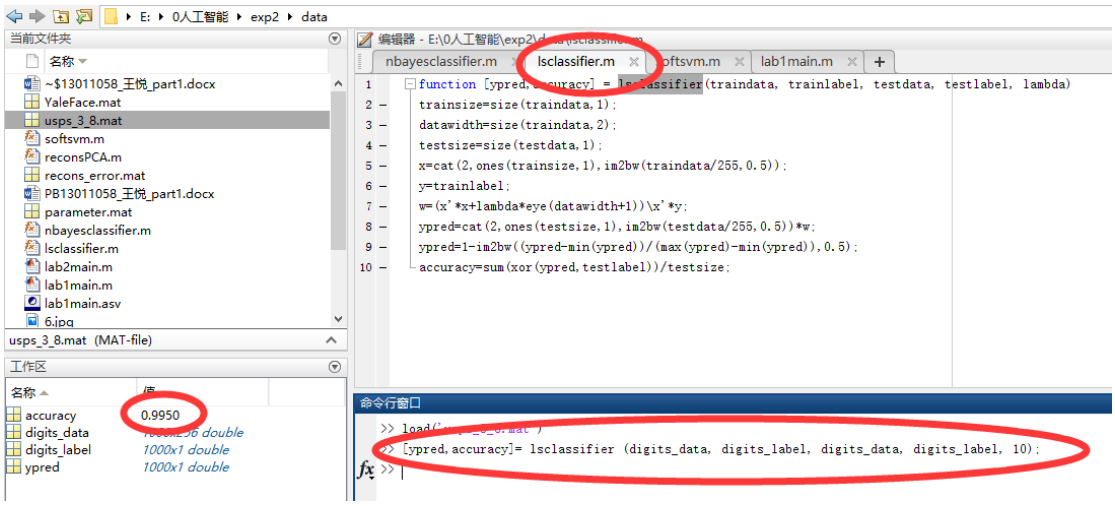
6.实验测试结果评价

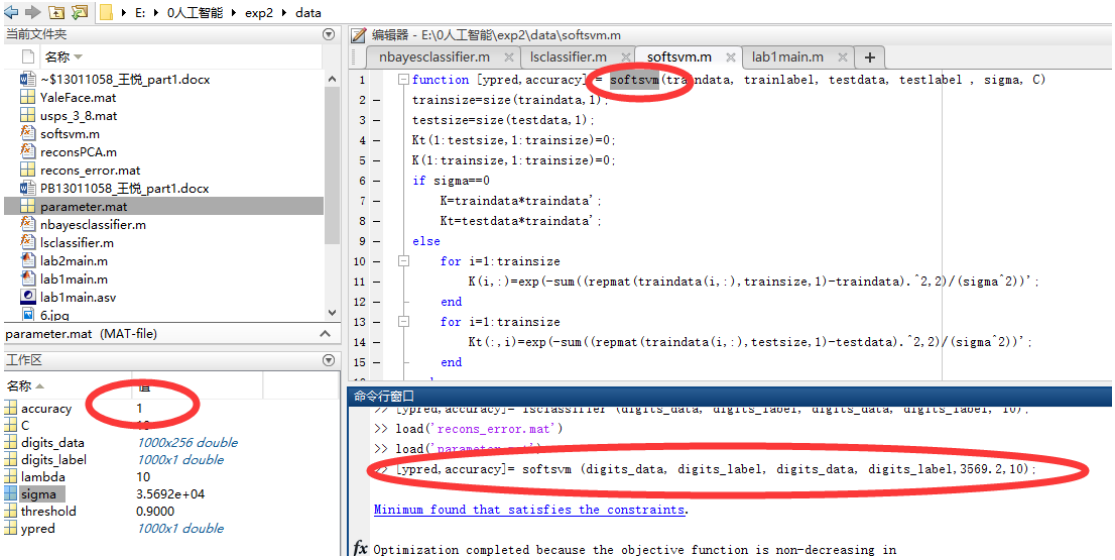
对于这部分，保存每个算法在相应数据集上对应的最佳参数并提交。如对于分类算法，需要保存 NaïveBayes 和 SVM 在相应数据集上使用 5-fold 交叉验证得到的参数(NaïveBayes 的 threshold，LeastSquares 的 lambda，SVM 的 sigma 和 C)，保存文件名统一为“数据集名”_parameters.mat。我们将会基于你们的算法代码以及最优参数，在保留下来的一部分数据上进行测试，并度量各个算法的性能。

实验记录：

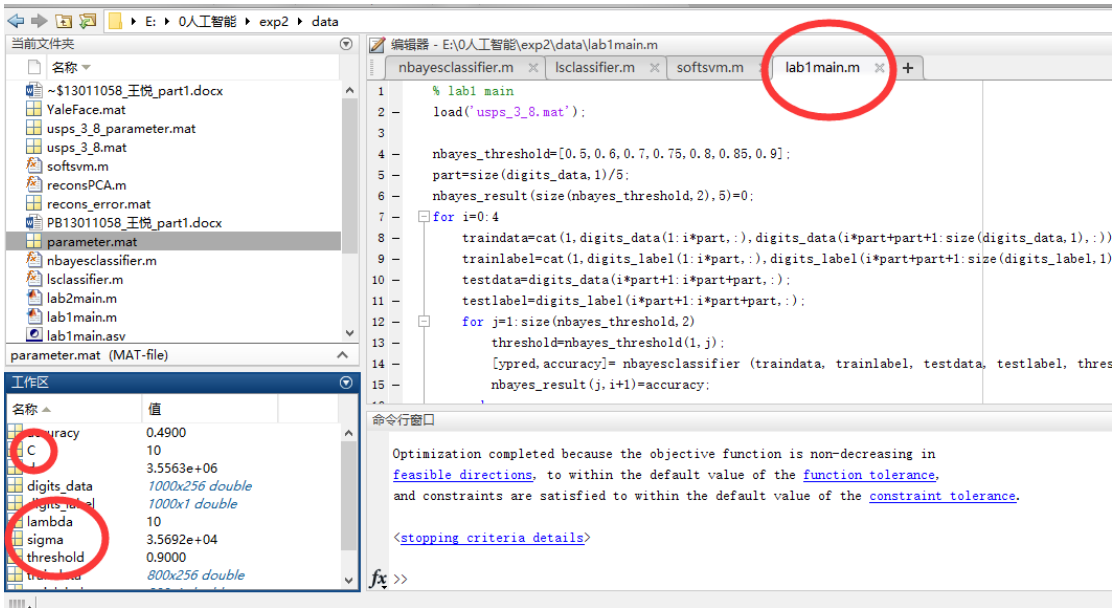
首先，编写三个子函数：`nbayesclassifier`、`lsclassifier`、`softsvm`，并独立测试通过。





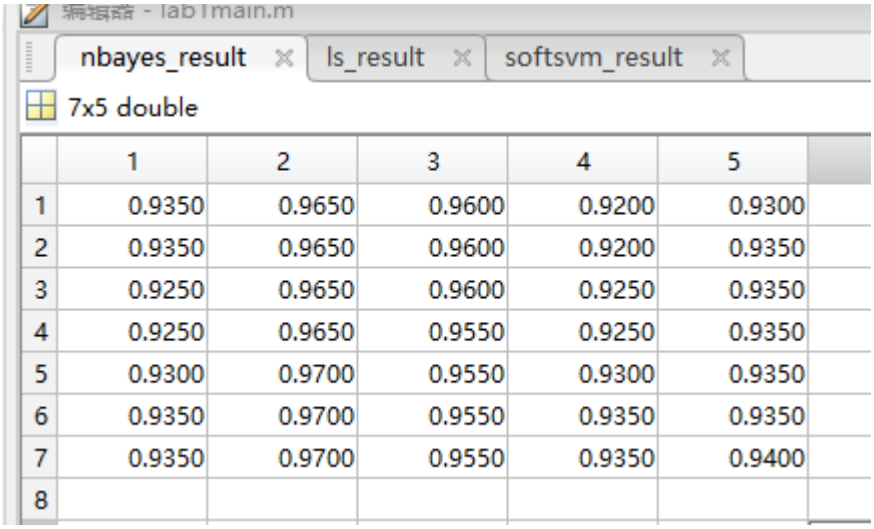


然后编写主函数，测试最佳参数。(测试时间较长，需要五六分钟，因为 svm 的测试比较耗时间)。



实验结果：

成功并且较为精简高效的实现了三种算法和主程序，对许多循环的处理采用了矩阵的方式去运算，提升性能。



nbayes_result | ls_result | softsvm_result

7x5 double

	1	2	3	4	5
1	0.9350	0.9650	0.9600	0.9200	0.9300
2	0.9350	0.9650	0.9600	0.9200	0.9350
3	0.9250	0.9650	0.9600	0.9250	0.9350
4	0.9250	0.9650	0.9550	0.9250	0.9350
5	0.9300	0.9700	0.9550	0.9300	0.9350
6	0.9350	0.9700	0.9550	0.9350	0.9350
7	0.9350	0.9700	0.9550	0.9350	0.9400
8					

编辑器 - lab1main.m

nbayes_result | ls_result | softsvm_result

11x5 double

	1	2	3	4	5
1	0.9600	0.9650	0.9700	0.9600	0.9750
2	0.9600	0.9650	0.9700	0.9600	0.9750
3	0.9600	0.9650	0.9700	0.9600	0.9800
4	0.9600	0.9650	0.9700	0.9650	0.9800
5	0.9600	0.9650	0.9750	0.9600	0.9750
6	0.9700	0.9700	0.9800	0.9600	0.9750
7	0.9700	0.9800	0.9800	0.9600	0.9700
8	0.9650	0.9800	0.9500	0.9650	0.9800
9	0.8900	0.9150	0.8950	0.9450	0.9350
10	0.8550	0.8300	0.8500	0.8750	0.8650
11	0.8150	0.7750	0.8300	0.8450	0.8400
12					

变量 | 所选内容 | 编辑

nbayes_result | ls_result | softsvm_result

20x5 double

	1	2	3	4	5
1	0.9050	0.9450	0.9350	0.9350	0.9200
2	0.9550	0.9600	0.9550	0.9700	0.9700
3	0.8000	0.6850	0.7600	0.7850	0.7850
4	0.6750	0.9800	0.5750	0.9150	0.9850
5	0.4550	0.6300	0.4400	0.9350	0.4900
6	0.7250	0.9500	0.5350	0.9350	0.9250
7	0.9050	0.9450	0.9350	0.9350	0.9200
8	0.9550	0.9600	0.9550	0.9700	0.9650
9	0.4550	0.4950	0.4400	0.9350	0.4900
10	0.4550	0.4950	0.4400	0.9350	0.4900
11	0.4550	0.6300	0.4400	0.9350	0.4900
12	0.7250	0.9500	0.5400	0.9350	0.9250
13	0.4550	0.4950	0.4400	0.9350	0.4900
14	0.4550	0.4950	0.4400	0.9350	0.4900
15	0.4550	0.4950	0.4400	0.9350	0.4900
16	0.4550	0.4950	0.4400	0.9350	0.4900
17	0.4550	0.4950	0.4400	0.9350	0.4900
18	0.4550	0.4950	0.4400	0.9350	0.4900
19	0.4550	0.4950	0.4400	0.9350	0.4900
20	0.4550	0.4950	0.4400	0.9350	0.4900
21					

三种算法都可以实现分类，从结果分析，svm 的效果最好，最小二乘分类器次之，最后是朴素贝叶斯分类器。但是从计算的时间上，最小二乘分类器最快，朴素贝叶斯次之，svm 最慢。个人分析认为 svm 慢的原因是其维度较高，并且朴素贝叶斯程序和最小二乘分类器对原始数据做了二值化处理，svm 没有处理，因此运算的时间消耗增加。

测试得到的最佳参数保存在了`usps_3_8_parameter.mat`中。

实验总结：

1) 概率模型分类器是一个条件概率模型。朴素贝叶斯分类器包括了朴素贝叶斯概率模型和相应的决策规则。朴素贝叶斯分类器的一个优势在于只需要根据少量的训练数据估计出必要的参数（变量的均值和方差）。由于变量独立假设，只需要估计各个变量的方法，而不需要确定整个协方差矩阵。
2) 最小二乘分类器，计算出一组参数，这组参数可以让计算出来的数据与观测数据最为接近，分为线性和非线性两种。
3) 支持向量机属于一般化线性分类器，能够同时最小化经验误差与最大化几何边缘区。支持向量机建构一个或多个高维（甚至是无限多维）的超平面来分类资料点，这个超平面即为分类边界。