# 机器学习实验报告

## 线性模型与SVM

学院：航天学院

姓名： 石瑞河

学号：1180400510

**实验1：编程实现线性回归**

数据集1：

文件名：LR-Data1.csv，30个训练数据

数据格式：第1列为输入x，第2列为输出y

数据生成：

真实模型参数：

数据集2：

文件名：LR-Data2.csv，30个训练数据

数据格式：第1列为输入x，第2列为输出y

数据生成：

真实模型参数：

要求：

1. 编程实现伪逆矩阵线性回归；
2. 显示估计的模型参数和；
3. 显示训练数据点和回归的直线；

**实验2：编程实现线性分类**

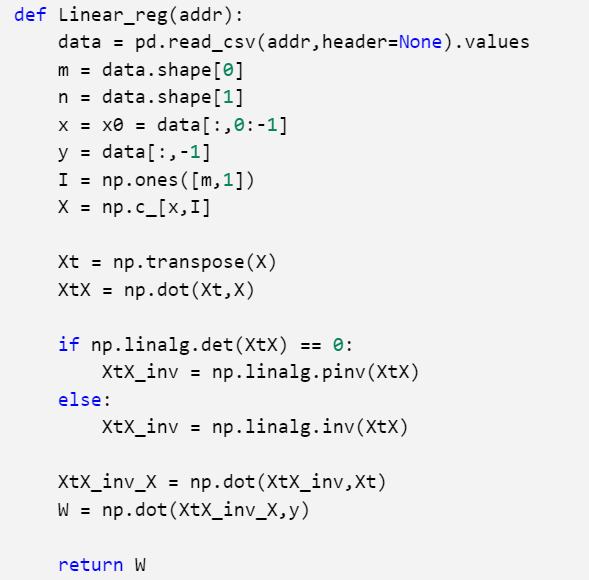
分类数据：

第1类数据：

第2类数据：

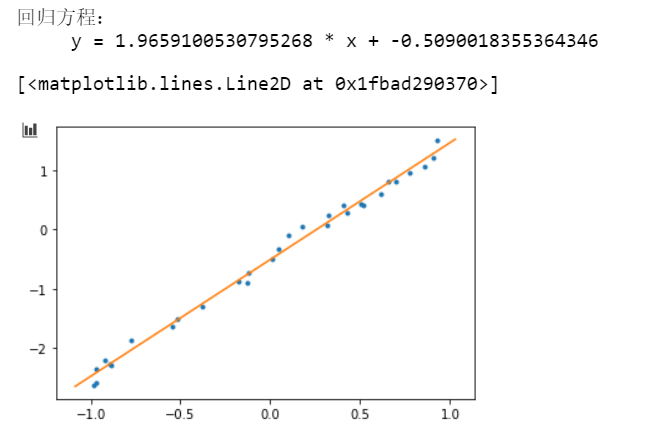
要求：

1. 编程实现伪逆矩阵线性分类；
2. 显示估计的模型参数和；
3. 显示训练数据和分类边界；

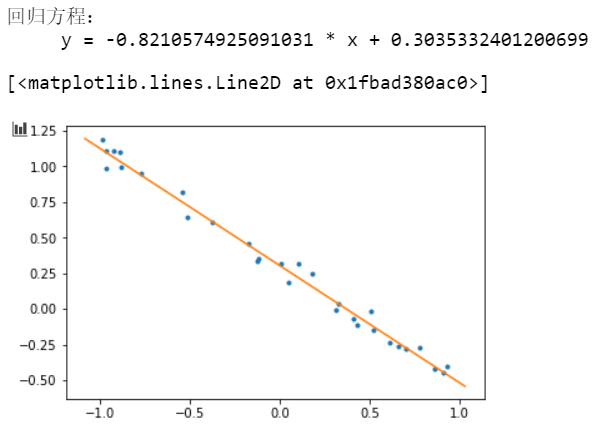
线性回归代码

实验1：

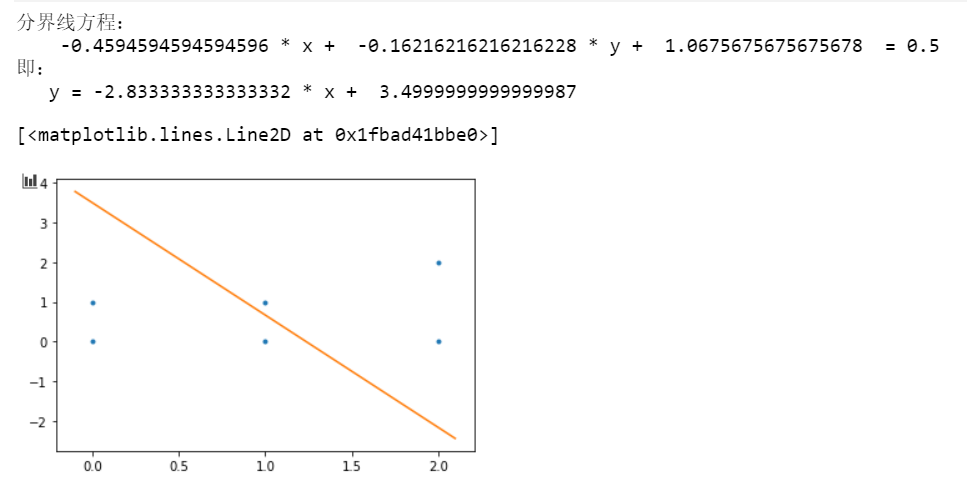
数据集1



数据集2



实验2：



其中(1,1) (2,2) (2,0)为一类，(0,0) (1,0) (0,1)为一类

对于这个简单样本，分类结果完全正确

**实验3：编程实现线性支持向量机分类**

训练数据：TrainSamples.csv，TrainLables.csv，包含30000个训练数据

测试数据：TestSamples.csv，TestLabels.csv，包含10000个测试数据；

数据内容：手写数字图像，图像大小，10个类别(0~9)

数据格式：

1. TrainSamples.csv和TestSamples.csv中，每行一个训练数据或测试数据的图像，784维属性；
2. TrainLabels.csv和TestLabels.csv中包含训练数据和测试数据的类别标记，每行对应一个训练数据的类别；

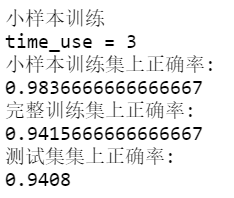
要求：

1. 编程实现支持向量机分类器；
2. 以TrainSamples.cvs和TrainLabels.csv为训练数据，训练SVM分类器；
3. 分类测试数据TestSamples.csv，与TestLabels.csv对比，显示分类的正确率；
4. 设置不同的SVM学习参数，包括核函数，核函数的参数，以及平衡参数C，显示不同训练参数学习得到分类器的分类正确率；

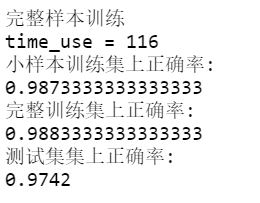
为了训练速度，从训练集中截取了一组3000个数据的小样本，称之为小样本训练集

无参数默认svm （rbf，C=1，gamma=auto=1/n\_features， 样本特征数的倒数）

小样本训练结果

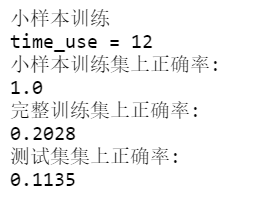


完整样本训练结果



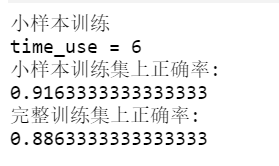
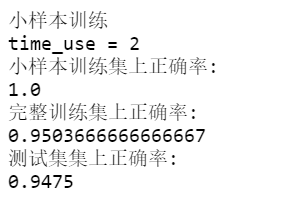
调节gamma；

gamma = 1时 （gamma = 20，0.1，0.0001时同结果）



调节C

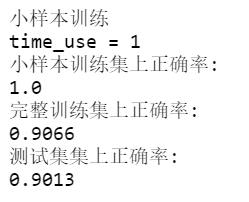
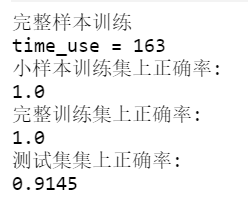
C=0.1时 C=10时

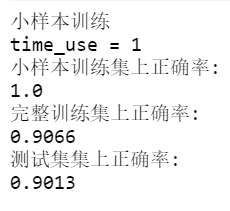
似乎默认选择的参数比较优秀

线性核 kernel = ‘linear’

C=1

C=0.1 （C=10时同结果）



rbf核函数默认参数下训练结果较为优秀，即便只使用一部分训练数据，也能训练出不错的模型，加大训练数据量时，模型性能能够再提升以部分，模型训练时间却会多花几十倍。Rbf核函数的gamma值不易设定，认为设定出的训练模型寻年效果都极差。C值能够影响模型的准确度，但不宜过大或过小，过小时分类错误多，过大时泛化能力差。

线性核计算快，效果可行，但没有rbf的拟合效果好。

附件清单：

实验1、2：linear.ipynb

实验3： svm.ipynb