**机器学习上机报告**

**第二章**

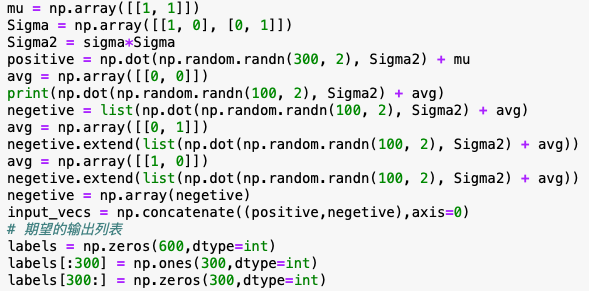
**实验内容：**

C2-1 构建一个二分类问题：逻辑与(AND)的一般问题。

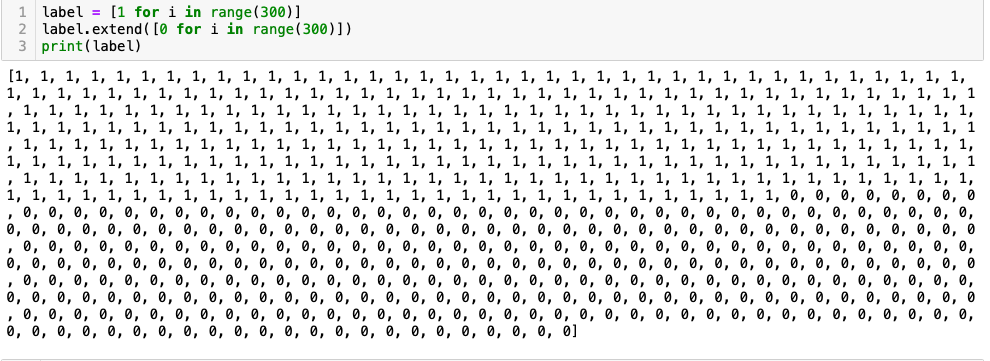
1. 生成数据：正样本样本服从高斯分布，均值为[1,1]，协方差矩阵为单位矩阵；负样本服从三个高斯分布的混合高斯分布，这三个高斯分布的均值分别为[0,0], [0,1], [1,0]，每个高斯分布的协方差矩阵均为(sigma\*单位矩阵)。
2. 学习：设sigma=0.01，请依上面的分布生成正负样本各300个，运用perceptron learning algorithm从数据中学习出一个perceptron，实现对正负样本的二分类。
3. 实验与讨论：请通过编程实验，讨论如下问题：a. 学习算法的收敛性与哪些因素存在怎样的关系？b. 讨论当sigma取值不断变大（如取值从0.01-1）情况下，学习算法的收敛性问题，由此可以得出怎样的结论？

**实验步骤：**

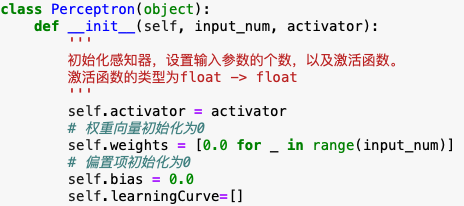
1. **利用numpy生成对应数据**

****

1. **生成label**

****

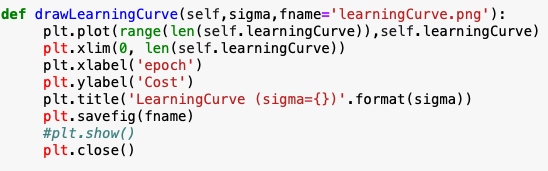
1. **初始化preception**

****

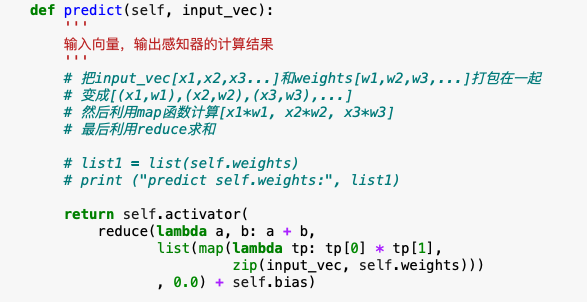
1. **训练preception以及更新权重**

****

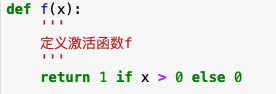
1. **画图函数**

****

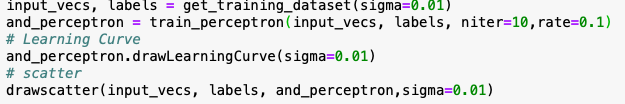
1. **predict函数**

****

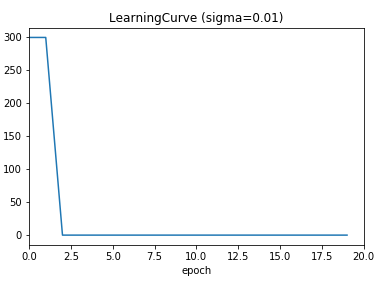
1. **定义0，1激活函数**

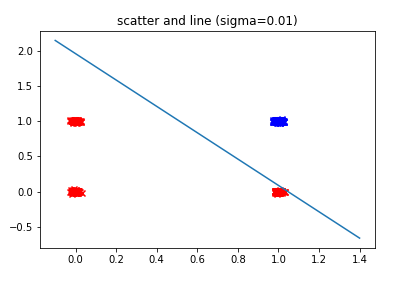
****

1. **训练感知器模型迭代10次**

****

**实验结果**

****

****

**实验与讨论：**

实验与讨论：请通过编程实验，讨论如下问题：a. 学习算法的收敛性与哪些因素存在怎样的关系？b. 讨论当sigma取值不断变大（如取值从0.01-1）情况下，学习算法的收敛性问题，由此可以得出怎样的结论

1. 收敛和数据的sigma大小和学习率以及迭代次数有关
2. sigma越大收敛越差

**第三章**

**实验内容：**

C3-1

(i) Generate 500 data (x,y), where y=x+n, n is Gaussian distributed with the mean of zero and standard deviation of delt. Please use linear regression learning algorithm for estimating y from input x;

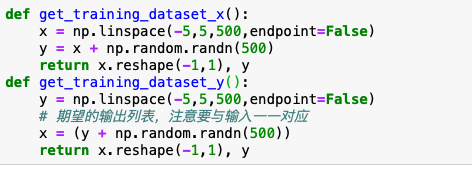
(ii) Do the same for x=y+n but still for estimating y from input x;

(iii) Make a comparison on the regression curves with x as the input and y as the output of the regressor obtained from (i) and (ii) respectively .

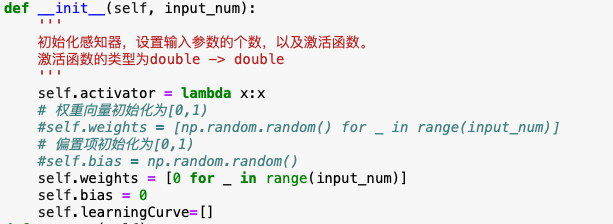
**实验步骤：**

1. **生成随机500个x和500个满足条件的n，以此来生成y**

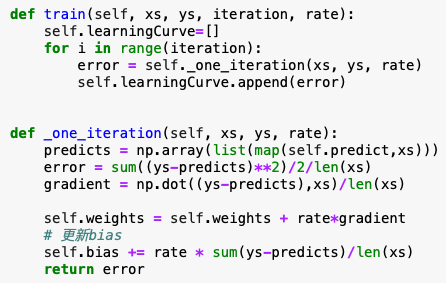
**生成随机500个y和500个满足条件的n，以此来生成y**

****

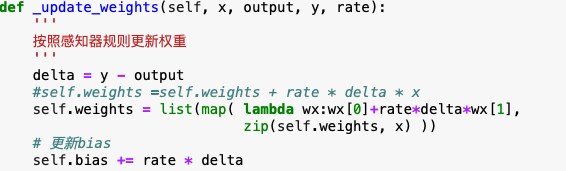
1. **初始化线性回归模型**

****

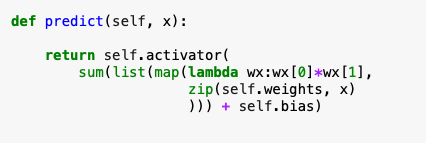
1. **模型训练函数**

****

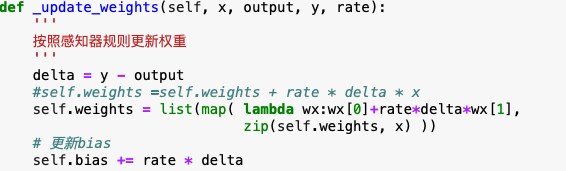
1. **训练过程中的权重更新函数**

****

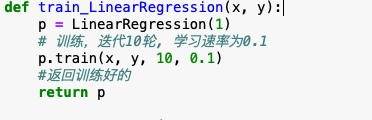
1. **预测predict函数**

****

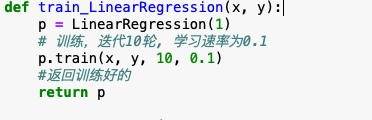
1. **画图函数**

****

1. **创建新型回归拟合器，输入参数个数为1**

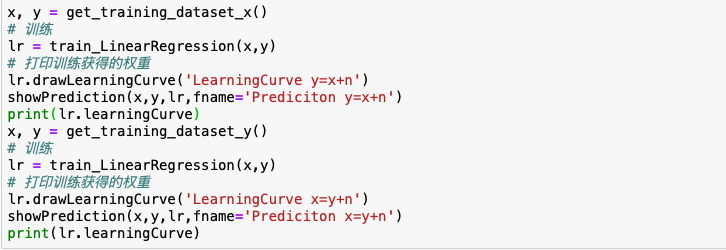
****

1. **定义分类直线画图函数**

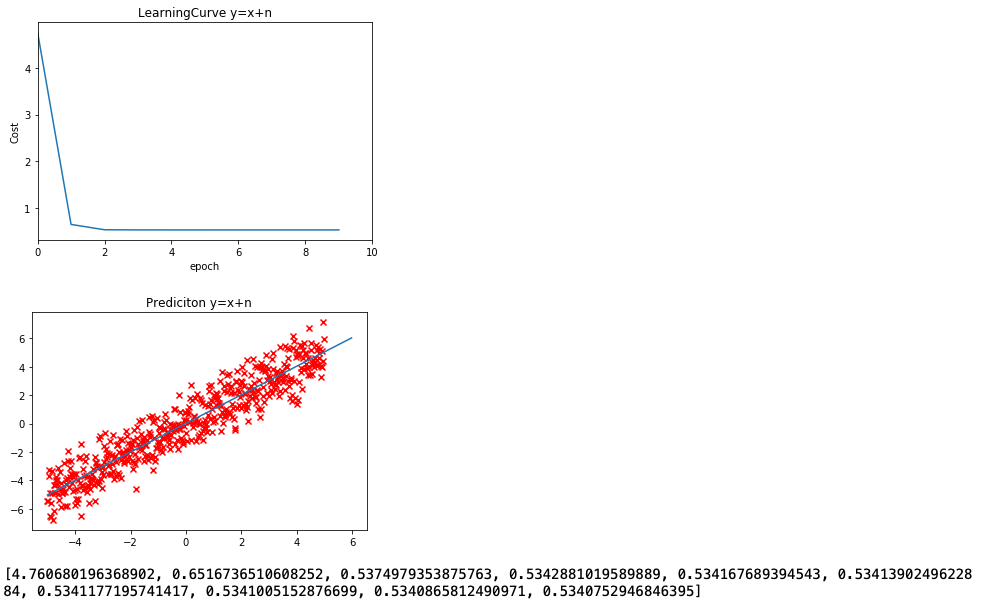
****

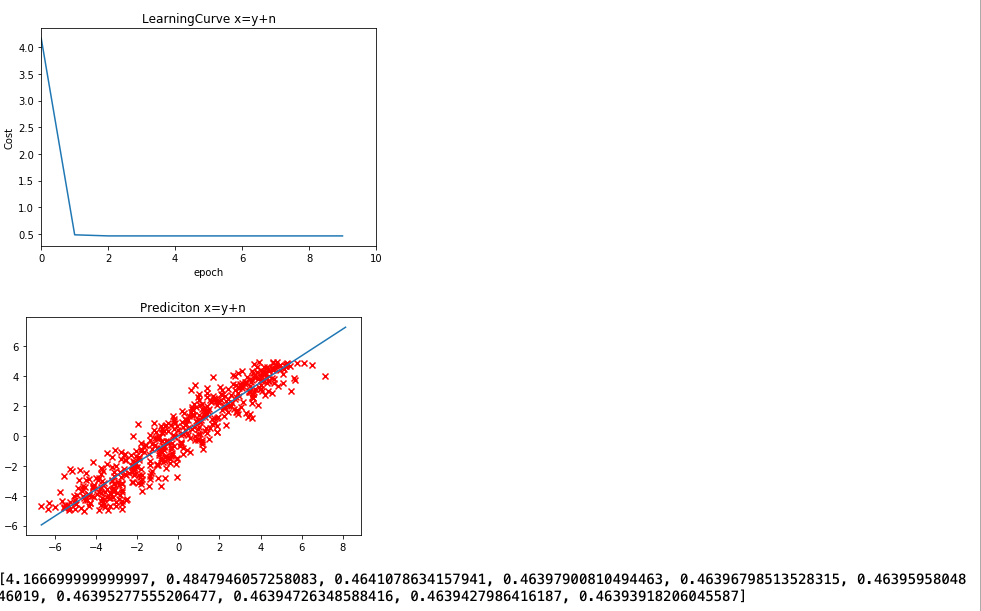
1. **主函数**

**训练x=y+n和y=x+n打印对应权重**

****

**实验结果：**

****

****

**第四章**

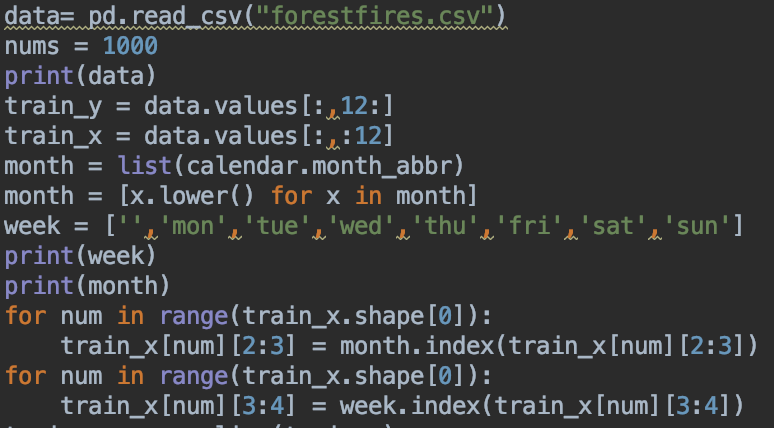
**实验内容：**

(regression problem) train a neural network, to predict the burned area of the forest, using the data from <http://archive.ics.uci.edu/ml/datasets/Forest+Fires>

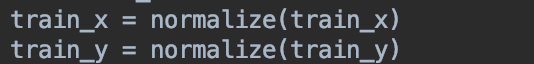
**实验步骤：**

1. **分析数据格式：**

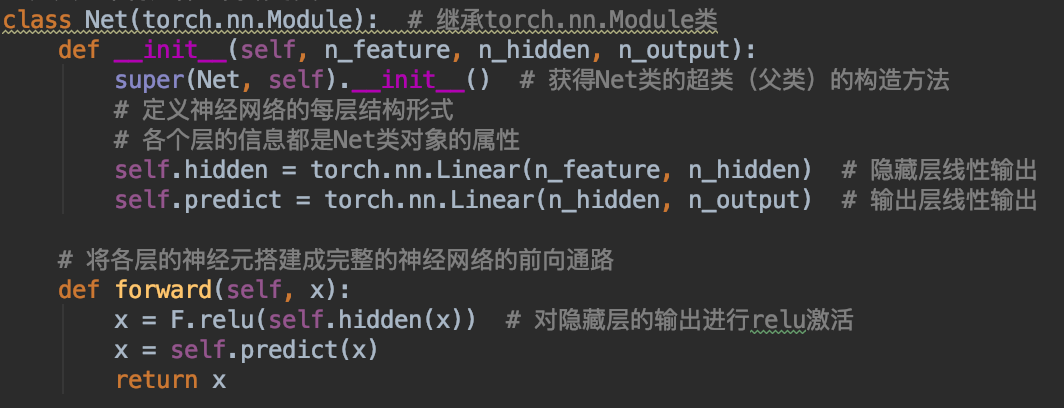
X,Y,month,day,FFMC,DMC,DC,ISI,temp,RH,wind,rain,area,其中month和day为字符串需要把他们转为float类型，并且把最后一项area设为train\_y



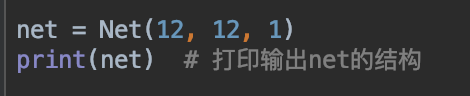
1. **对数据进行归一化**

****

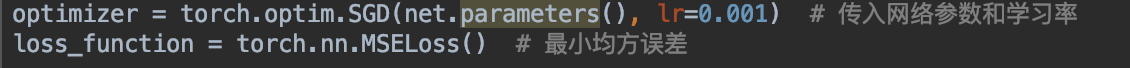
1. **利用pytorch初始化神经网络类**

****

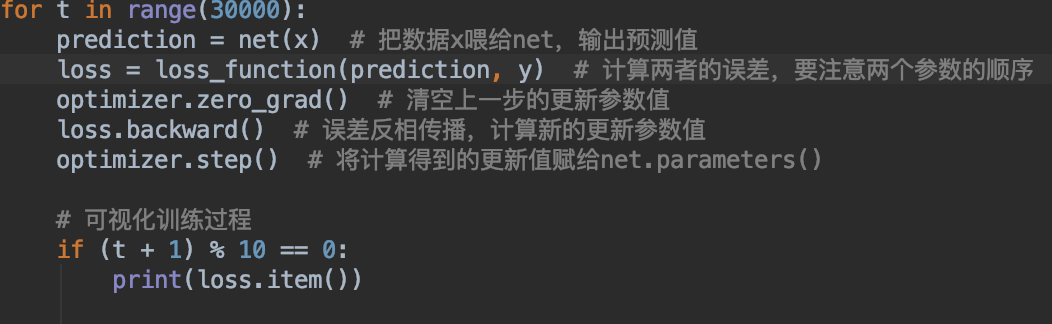
1. **初始化net类为12输入，12层，1输出**

****

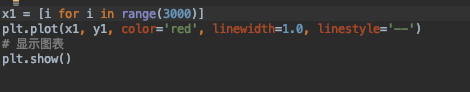
**5、给网络喂入参数**

****

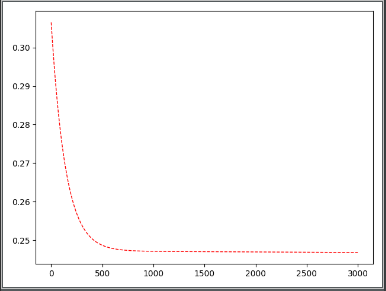
**6、训练3000次**

****

1. **可视化训练结果**

****

**实验结果：**

****

**第五章**

**实验内容：**

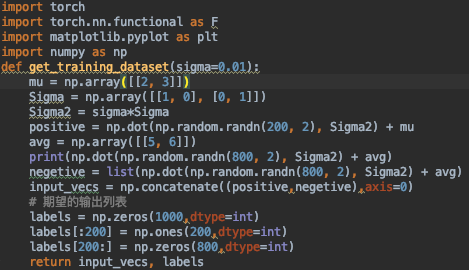
C5-1 构建一个类别不平衡的二分类问题。

1. 生成数据：正样本样本服从高斯分布，均值为[2,3]，协方差矩阵为单位矩阵；负样本服从高斯分布，均值为[5,6]，协方差矩阵为单位矩阵。
2. 学习：请依上面的分布生成正样本200个，负样本800个，将其划分为包含60%样本的训练集、20%样本的验证集和20%样本的测试集，通过分别构建两个不同的MLP模型实现对正负样本的二分类。其中第一个MLP模型含有一个隐层，第二个MLP模型含有两个隐层。

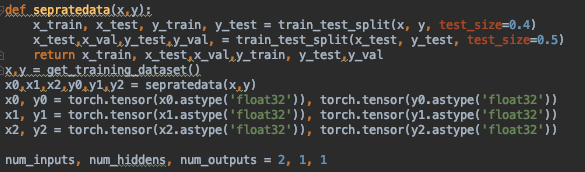
实验与讨论：请通过编程实验，讨论如下问题：a. 若要求MLP模型对于正例样本的预测有更高的查准率和查全率，请考虑在模型选择中采用哪种性能衡量指标；b.通过绘制学习曲线，分析模型的偏差与方差；c.通过绘制ROC曲线，比较两个MLP模型。

**实验步骤：**

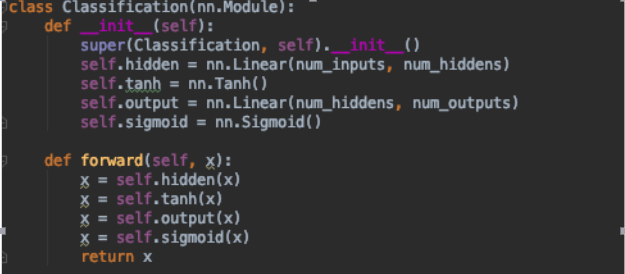
1. 生成数据：



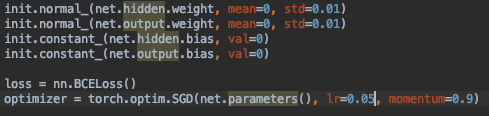
1. 划分数据集合为验证集，训练集，测试集，设定一层单隐层



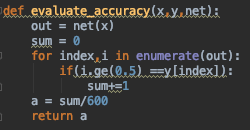
1. 初始化二分类



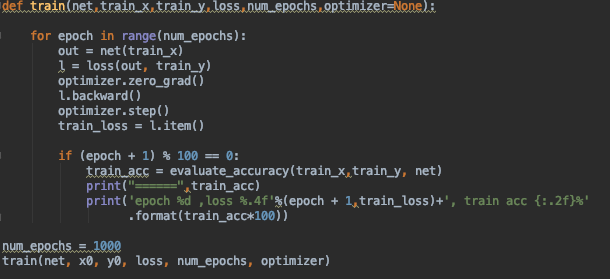
1. 初始化层内参数，损失函数和学习率等参数



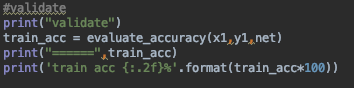
5、定义激活函数



1. 训练集训练

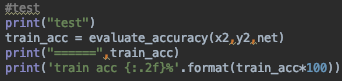


1. 利用验证集合调参，最后参数为lr=0.01



截屏2020-06-27 00.18.35

1. 利用验证集合验证

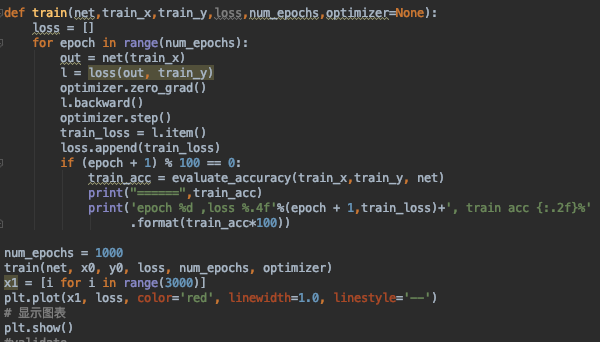


截屏2020-06-27 00.18.40

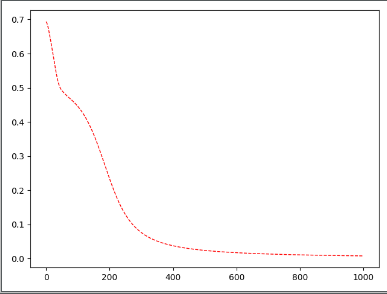
9、绘制roc曲线



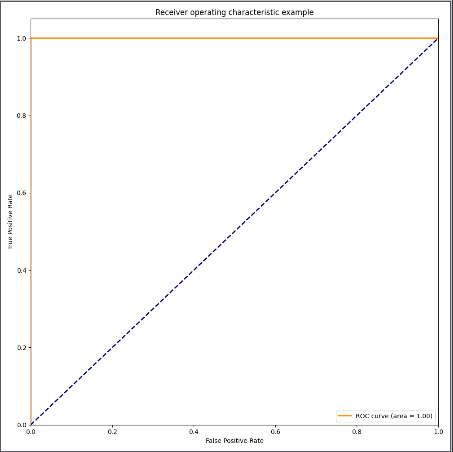
1. 绘制训练曲线



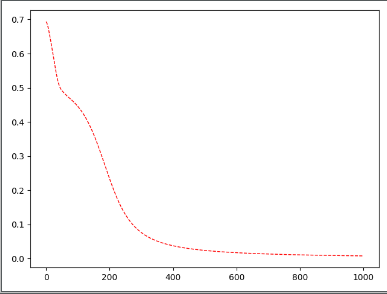
11、训练曲线：



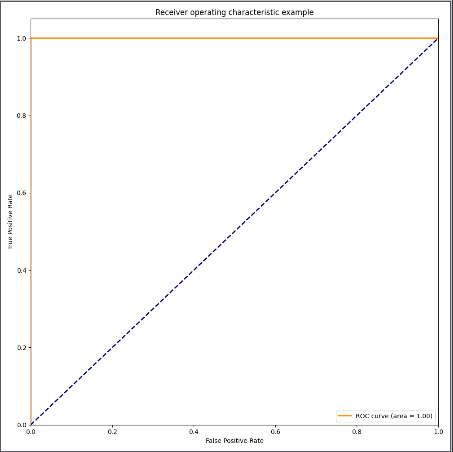
12roc曲线：



双隐层:



roc曲线：



实验讨论：

1. 利用交叉验证法
2. 偏差过高过拟合
3. 双隐层较好能够更快的到达训练手腕点

**第六章**

**实验内容：**

C6-1 Fisheriris data can be read by the following matlab command:

‘load fisheriris’.

Its meaning can be found from the internet.

1. Use one-against-all strategy to decompose the 4-ary classification problem into binary classification problems. Use SVM to solve each binary classification problems and combine all the binary classifiers to solve the problem.
2. Use KNN to solve the fisheriris data classification problem.
3. Use MLP to solve each binary classification problems and combine all the binary classifiers to solve the problem.

Compare the generalization performance and the storage required for the classifiers obtained in (a), (b) and (c).

**实验步骤：**

**a)**

1. **利用matlib的fisheriris导入数据**

clc,clear,close all

load fisheriris

X = meas;

Y = species;

tabulate(Y);

**2、创建SVM模板（二分类模型），并对分类变量进行标准化处理**

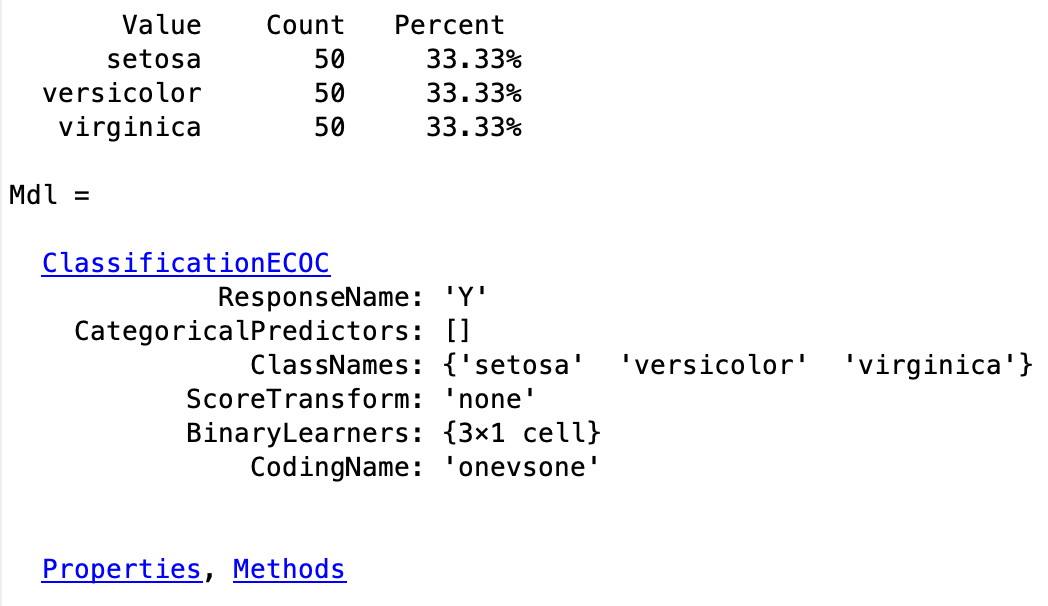
t = templateSVM( 'Standardize' , 1); % 对分类变量进行标准化处理

**3、基于SVM二分类模型进行训练并生成多分类模型**

**Mdl = fitcecoc(X,Y,'Learners',t,...**

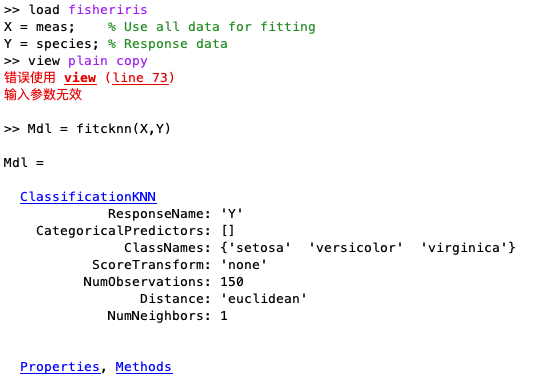
**'ClassNames',{'setosa','versicolor','virginica'})**

**实验结果:**

****

**B)KNN**

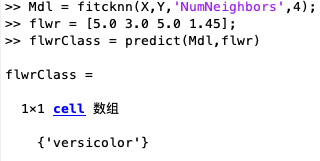
**导入数据，创建画布,画出数据点，创建knn**

****

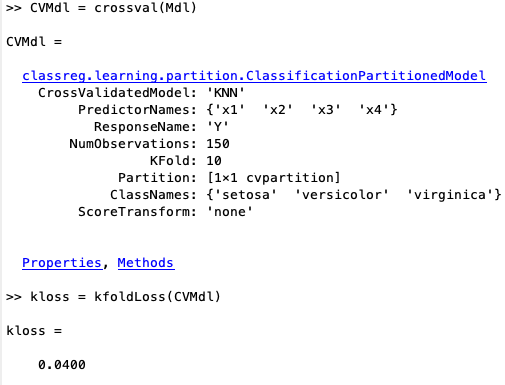
**调整最近邻数量：**

Mdl.NumNeighbors = 4;

**进行预测：**

****

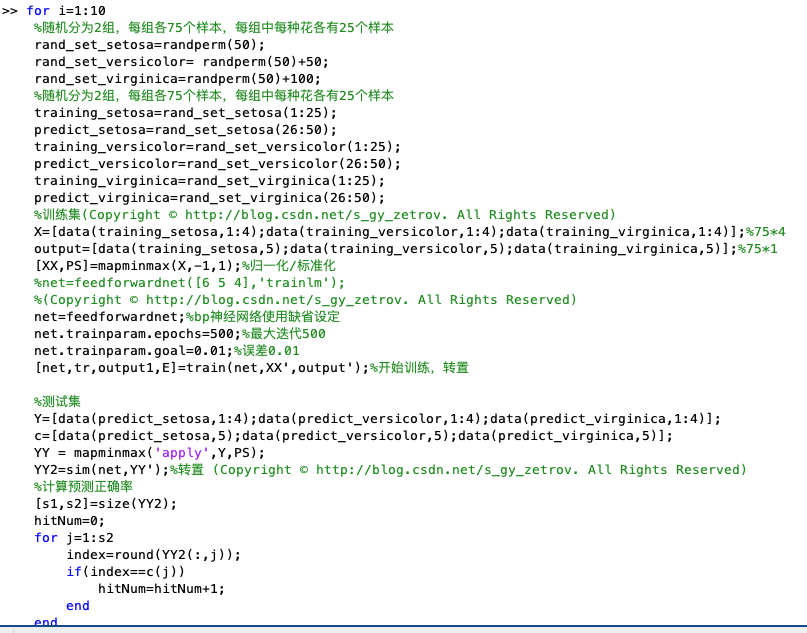
**利用crossval(Mdl)评估效果：**

****

**C）MLP**

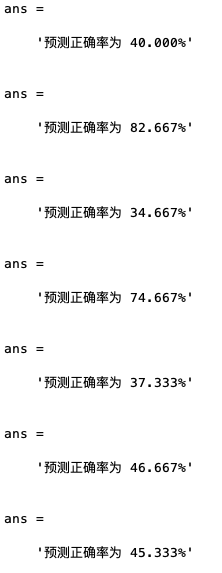
**导入数据**

**创建MLP模型：**

****

**截屏2020-06-27 02.02.23**

**结果最大值为：82%**

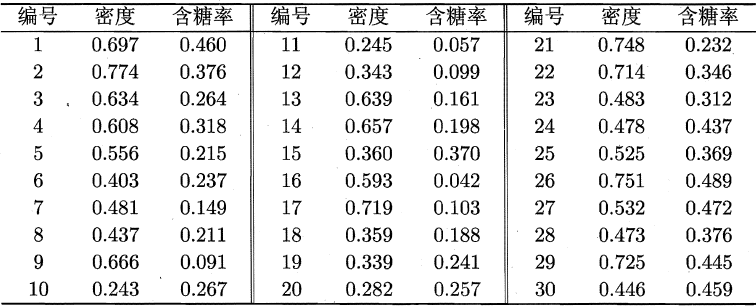
****

**第7章**

**实验内容：**

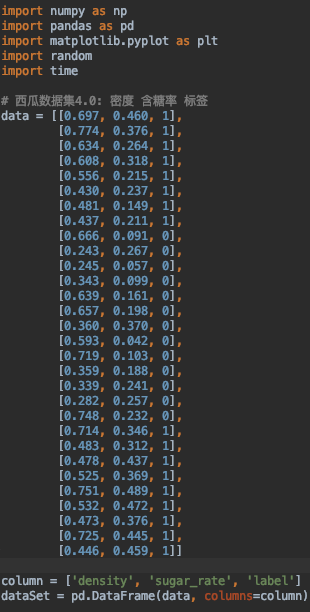
试编程实现K-means算法，设置三组不同的k值、三组不同的初始中心点，在西瓜数据集4.0上进行实验比较，并讨论什么样的初始中心有利于取得好结果。

表1 西瓜数据集4.0

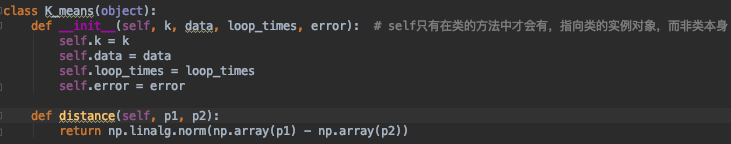


**实验步骤:**

1. **导入相应的包和利用dataframe导入数据集**

****

1. **定义k-means类和距离类方法**

****

1. **定义类方法fit根据k-means算法，选取k个初始化，计算每一个点到簇心的距离，更新簇心，知道簇心不变**

****

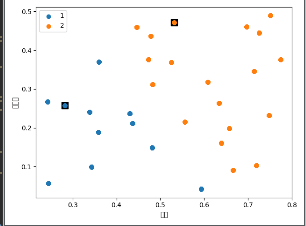
****

1. **定义不同的k值**

**截屏2020-06-27 01.11.13**

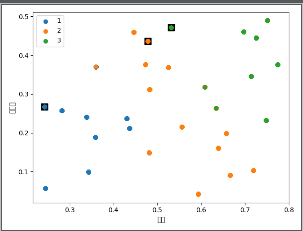
1. **当k=2时**

**截屏2020-06-27 01.07.48**

****

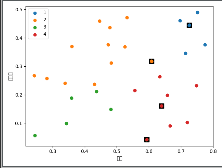
**当k=3时**

**截屏2020-06-27 01.08.48**

****

**当k=4时**

**截屏2020-06-27 01.10.33**

****

**实验结果：**

**发现当k=3时有更好的分类效果**

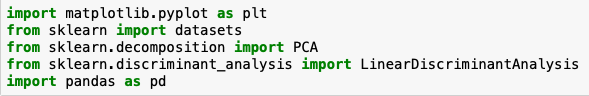
**第八章**

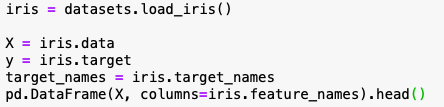
**实验内容：**

请通过对Iris data的可视化，比较PCA、LDA和ICA的可视化效果。

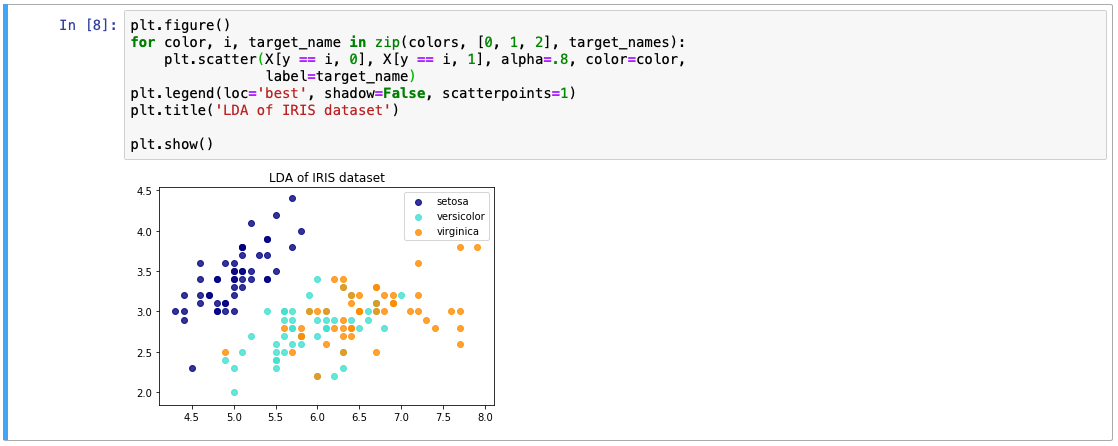
**实验步骤：**

**导入iris数据集合：**

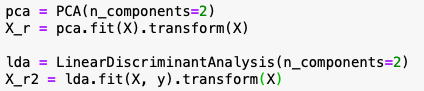
****

****

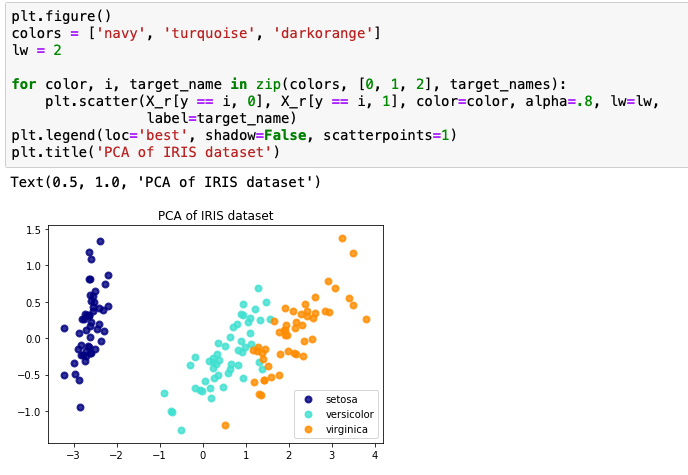
**iris数据原始数据点：**

****

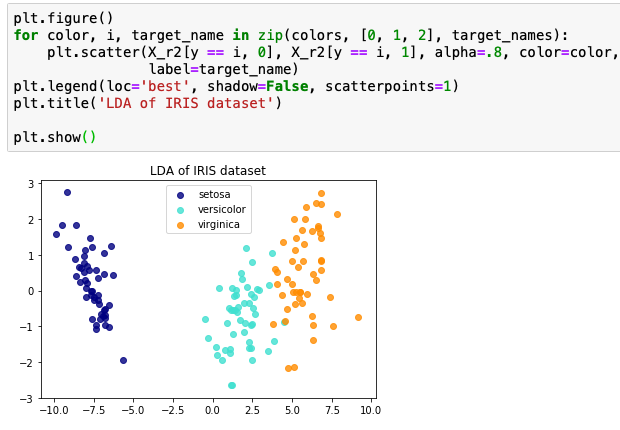
**进行pca和lda：**

****

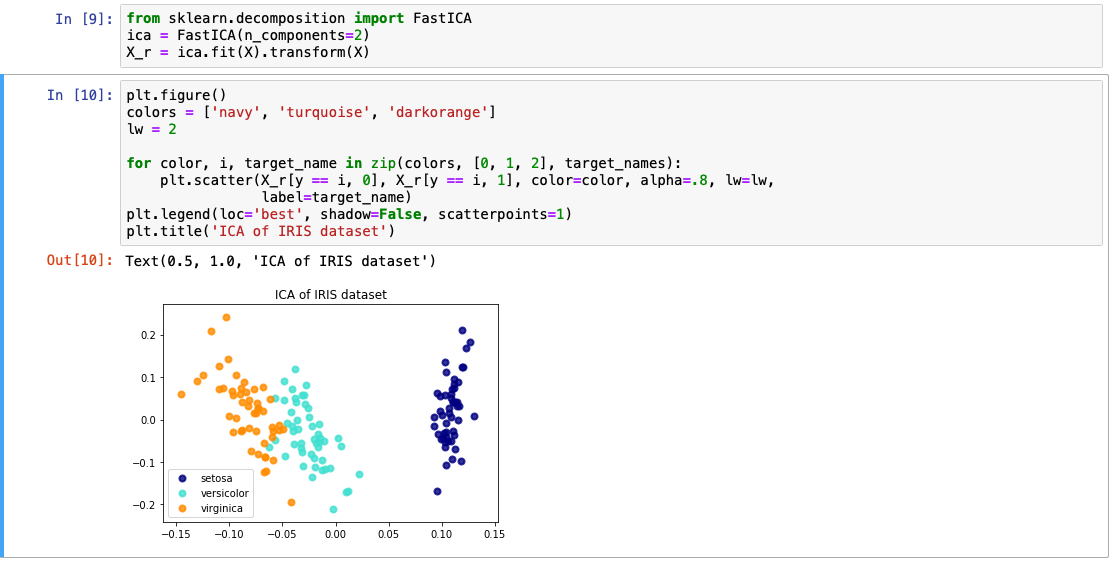
**画出pca变化后的数据点：**

****

**画出LDA变化：**

****

**ICA变化：**

****

**总结：**

利用ICA和LDA可以让两个本来看起来比较重叠的数据集合划分开来，而pca重叠部分较多