第一次作业实验报告

一、 题目；

使用 Python 实现对数几率回归模型

二、 问题描述；

2-1 使用数据集：鸢尾花数据集（Iris）： http://archive.ics.uci.edu/ml/datasets/Iris
2-2 完成步骤：
1 下载 Iris 数据集，并读懂数据集的相关说明， 了解数据格式。
2 编程实现对数几率回归模型，并对 Iris 数据集进行分类以验证模型的效能：
(1). 将数据集的 50%作为训练集，50%作为测试集，检验模型在测试集上的分类正确率
(2). 将数据集的 70%作为训练集，30%作为测试集，检验模型在测试集上的分类正确率
(3). 将数据集的 90%作为训练集，10%作为测试集，检验模型在测试集上的分类正确率

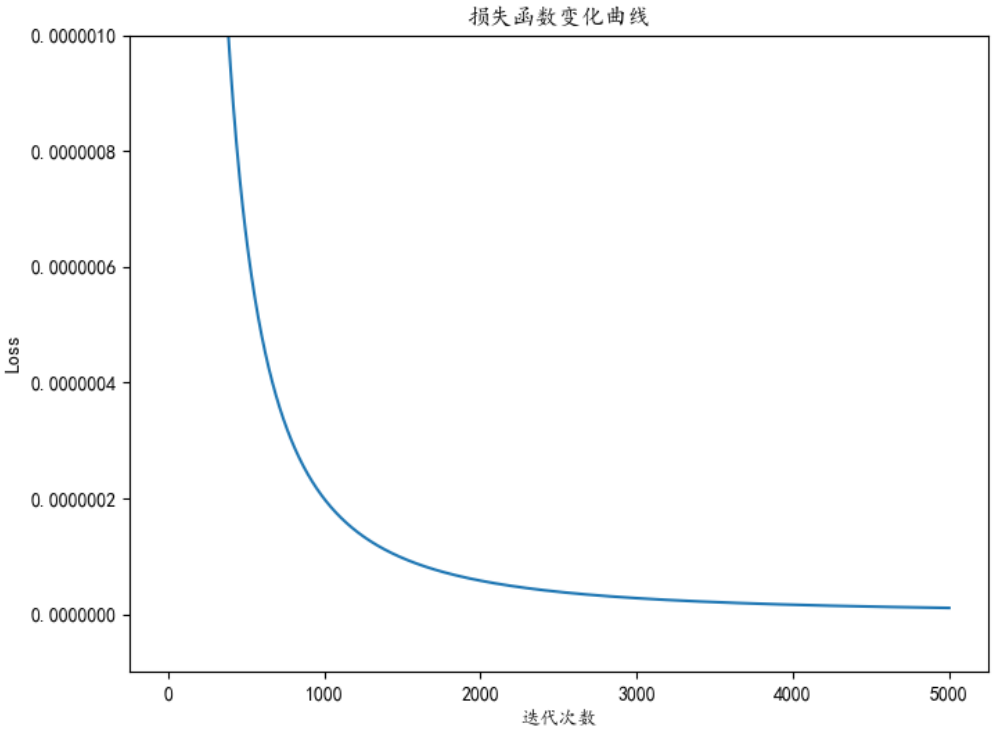
三、 数据集描述；

1. 数据集名称: Iris Plants Database
2. 实例数量: 150 (其中每类 50 个实例)
3. 特征数量：4
4. 具体特征信息：
    1. sepal length in cm
    2. sepal width in cm
    3. petal length in cm
    4. petal width in cm
5. 类别:
    1. Iris Setosa
    2. Iris Versicolour
    3. Iris Virginica

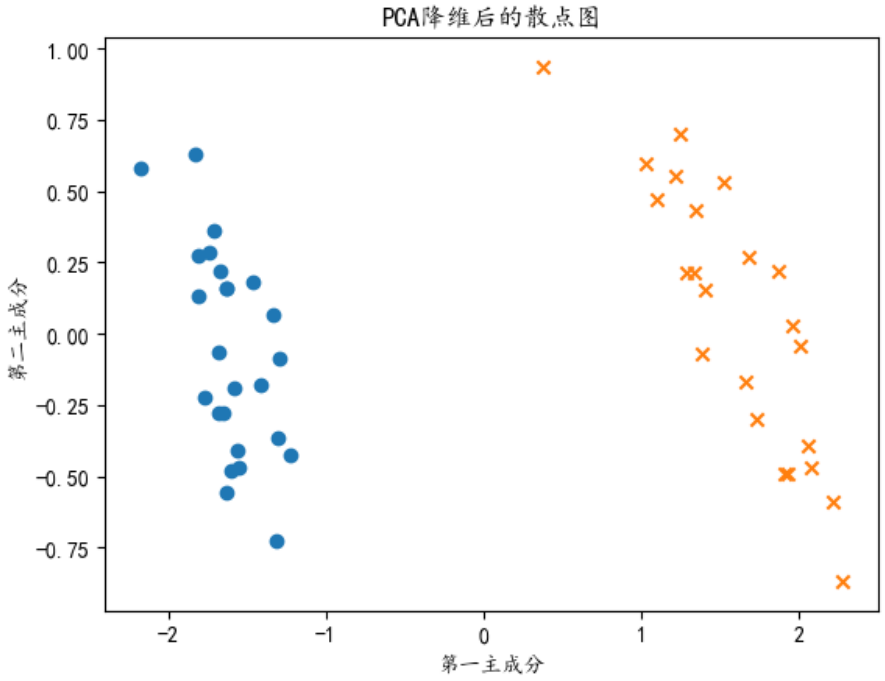
四、 实验结果图；

1) **训练集 : 测试集 = 1 : 1、**
其中,

$$loss = \frac{1}{2m}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



图表 1 损失函数变化曲线



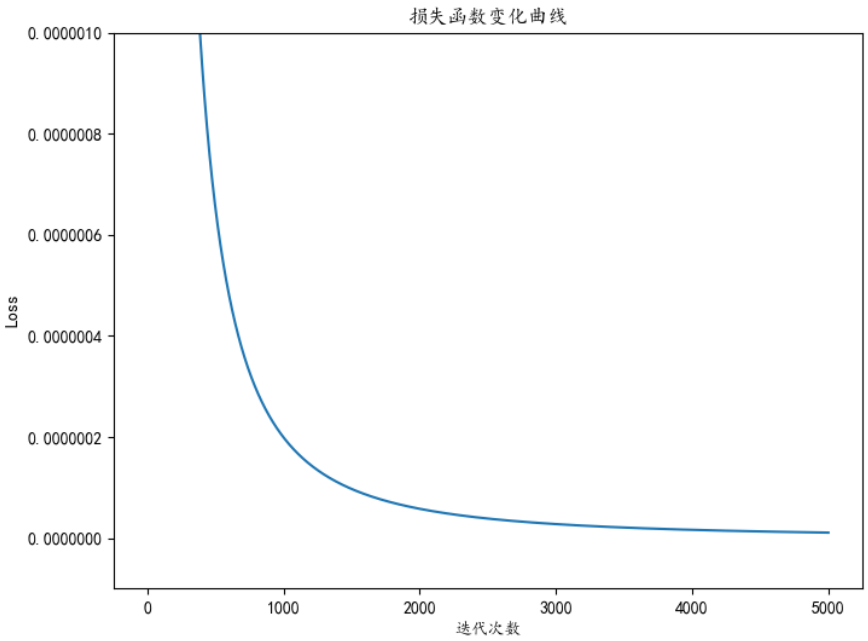
图表 2 特征向量从四维降到二位的散点图

附：选取的数据集类为 Iris Setosa ，Iris Versicolour

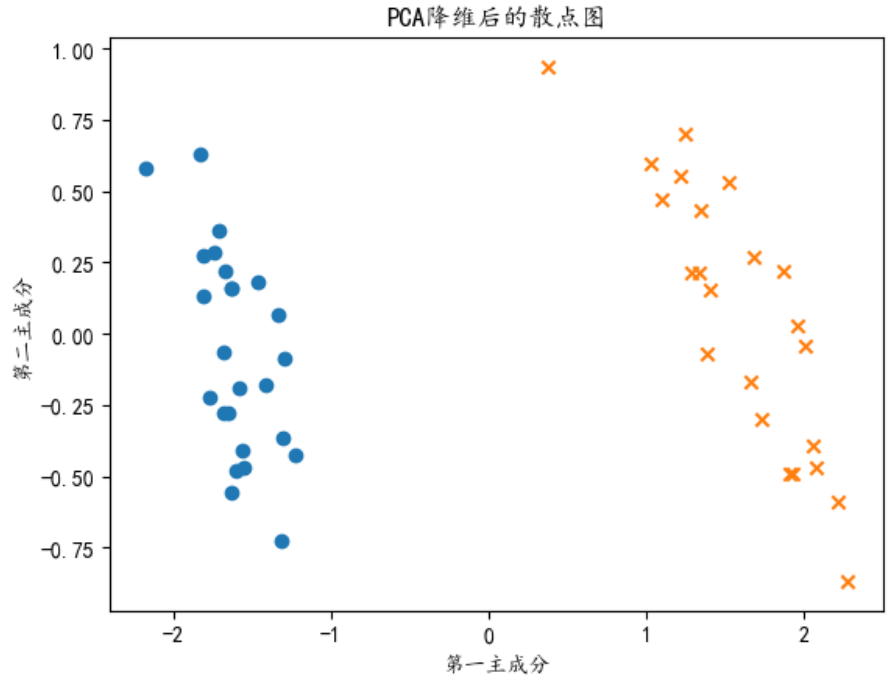
2)训练集 ： 测试集 = 7 ： 3

其中，

$$loss = \frac{1}{2m}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



图表 3 损失函数变化曲线



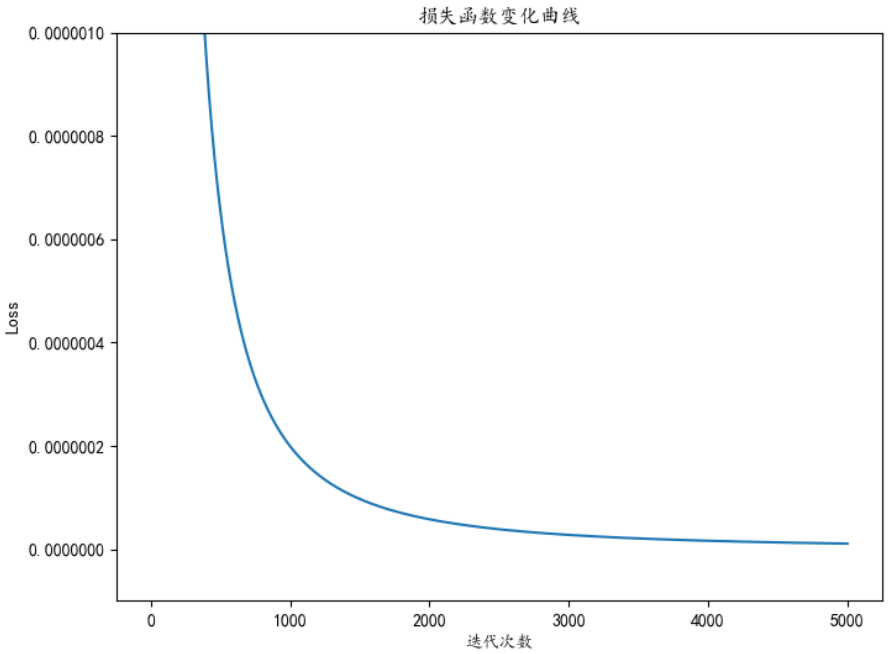
图表 4 特征向量从四维降到二位的散点图

附：选取的数据集类为 Iris Setosa ，Iris Versicolour

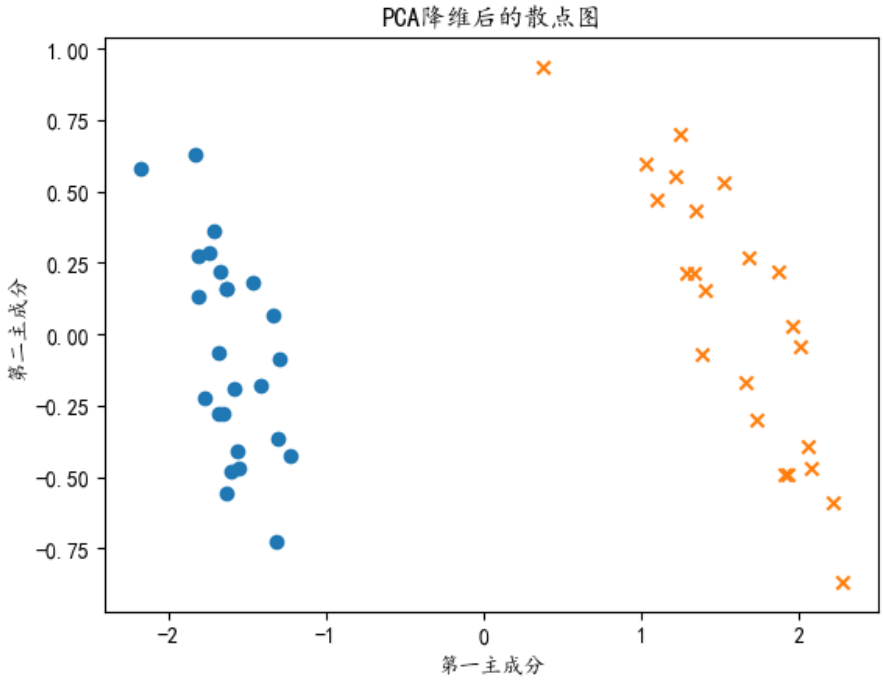
3)训练集 ： 测试集 = 9 ： 1

其中，

$$loss = \frac{1}{2m}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$



图表 5 损失函数变化曲线



图表 6 特征向量从四维降到二位的散点图

附：选取的数据集类为 Iris Setosa ，Iris Versicolour

五、 实验结果分析；

5-1 训练集 ： 测试集 = 9 ： 1

训练集各个特征各自的方差百分比

	第一主成分	第二主成分	第三主成分	第四主成分
百分比	92.167226 %	5.762558 %	1.706583 %	0.363633 %

可见第一、二主成分可体现原特性信息的 97.93%，故可用第一、二主成分来描述
由图表 2 可见，Iris Setosa && Iris Versicolour 两类数据集**降维**后呈现出一种明显的**线性可分**关系，并且差距都**非常显著**，
在学习率为

$$\alpha = 0.1$$

迭代次数为 5000 次的时候，学习的正确率高达 100%
设

$$\beta = (a, \quad b, \quad c)^T$$

其中：
a 表示 PCA 降维后的第一主成分系数
a 表示 PCA 降维后的第一主成分系数
c 表示常数项

则 Logistics 函数表达式为

$$y = \frac{1}{1 + e^{-\beta \hat{x}}}$$

其中：

$\beta =$ （7.5165302，3.24566008，1.07410887）

$\hat{x} =$ （x；1）

5-2 训练集 ： 测试集 = 7 ： 3

训练集各个特征各自的方差百分比

	第一主成分	第二主成分	第三主成分	第四主成分
百分比	90.03101 %	7.902239 %	1.647118 %	0.419633 %

可见第一、二主成分可体现原特性信息的 97.93%，故可用第一、二主成分来描述
由图表 4 可见，Iris Setosa && Iris Versicolour 两类数据集**降维**后呈现出一种明显的**线性可分**关系，并且差距都**非常显著**，
在学习率为

$$\alpha = 0.1$$

迭代次数为 5000 次的时候，学习的正确率高达 100%
设

$$\beta = (a, \quad b, \quad c)^T$$

其中：
a 表示 PCA 降维后的第一主成分系数
a 表示 PCA 降维后的第一主成分系数
c 表示常数项

则 Logistics 函数表达式为

$$y = \frac{1}{1 + e^{-\beta \hat{x}}}$$

其中：

$\beta =$（8.88774289，-2.95379474，2.24050438）

$\hat{x} =$ （x；1）

5-3 同理 可得 当 **训练集 ： 测试集 = 9 ： 1** 时

训练集各个特征各自的方差百分比

	第一主成分	**第二主成分**	**第三主成分**	**第四主成分**
百分比	90.83831 %	7.059793 %	1.746783 %	0.355113 %

$$y = \frac{1}{1 + e^{-(8.89002291k_1 - 3.14308308k_2 + 2.44324421)}}$$

学习的正确率高达 100 %