## Logistic回归

学号：Y30180631 姓名：杨昊晨

实验内容为，对iris数据采用logistics回归方法进行分类。

1. 数据集简介

iris数据集的中文名是安德森鸢尾花卉数据集，英文全称是Anderson’s Iris data set。iris包含150个样本，对应数据集的每行数据。每行数据包含每个样本的四个特征和样本的类别信息，所以iris数据集是一个150行5列的二维表。

通俗地说，iris数据集是用来给花做分类的数据集，每个样本包含了花萼长度、花萼宽度、花瓣长度、花瓣宽度四个特征（前4列），我们需要建立一个分类器，分类器可以通过样本的四个特征来判断样本属于山鸢尾、变色鸢尾还是维吉尼亚鸢尾（这三个名词都是花的品种）。

iris的每个样本都包含了品种信息，即目标属性（第5列，也叫target或label）。

样本局部截图如下：

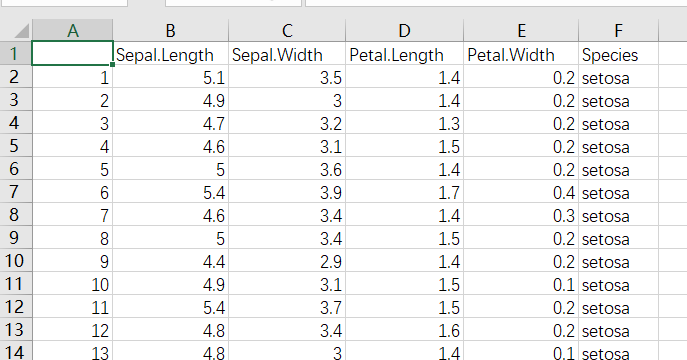


图1-1 Iris数据集局部截图

前两类物种数据完全不重合，后两类数据有部分重合。本次实验采用全部的四维特征，两类物种进行分类。

1. 解决过程

1、PCA降维

首先读取Iris数据集，数据集来源于python的sklearn库，并使用sklearn库的PCA函数进行降维，将其思维的特征数据降至二维平面，绘制平面上如下图：

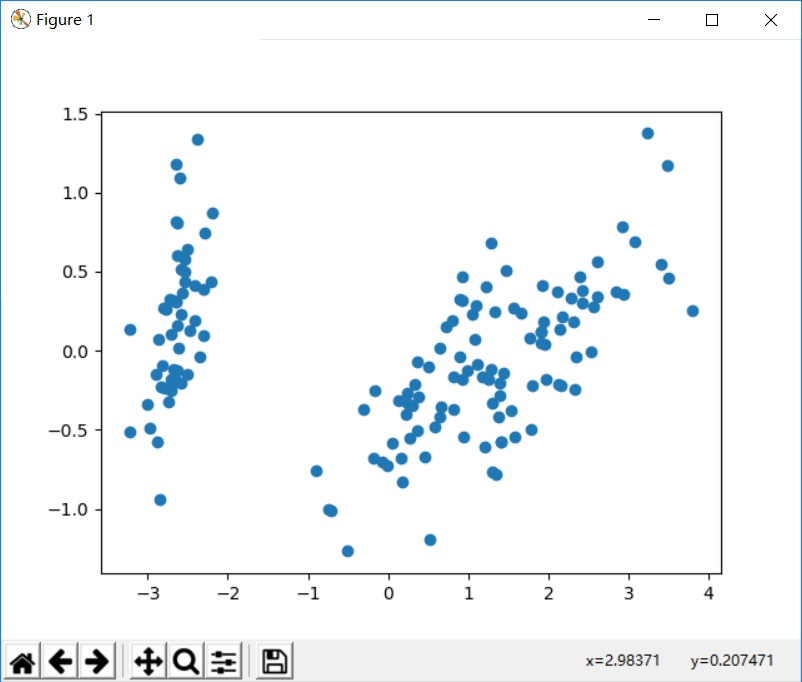


图2-1 降维图像

2、梯度下降法求logistic回归

接下来采用梯度下降法求解Logistic Regression。

由于没有给出结束的条件，这里简单迭代5000次，学习率采用0.1，并且绘制收敛曲线。

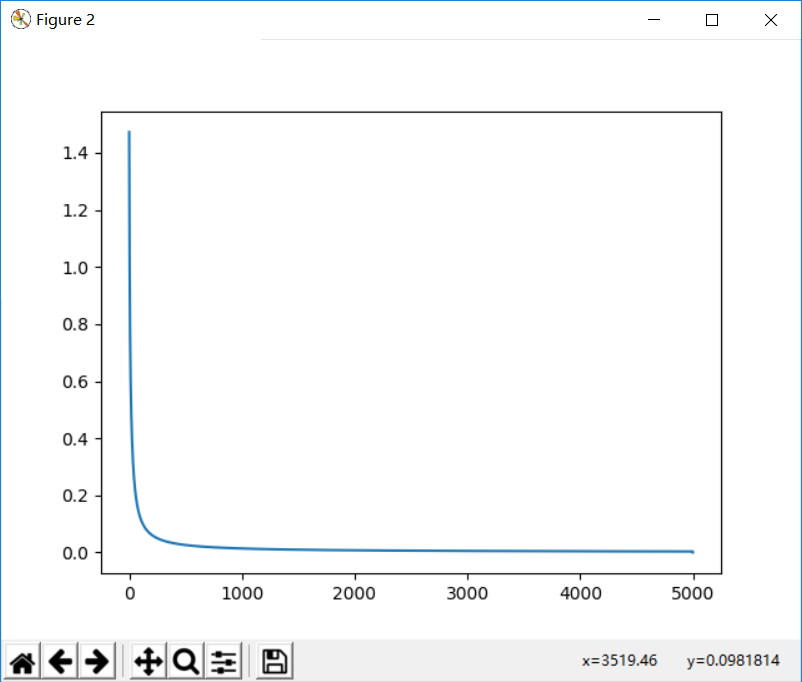


图2-2 收敛曲线

3、绘制后验概率图

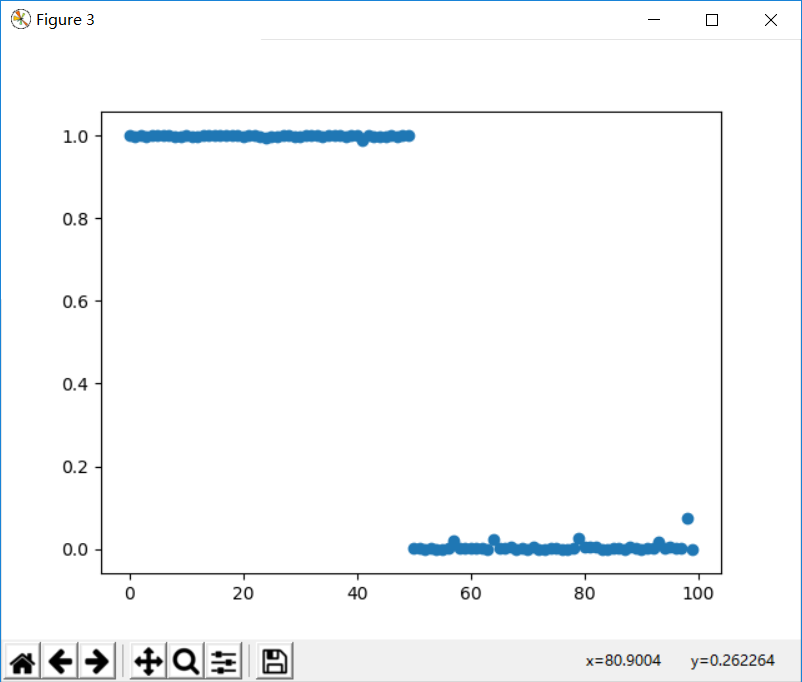


图2-3 后验概率图

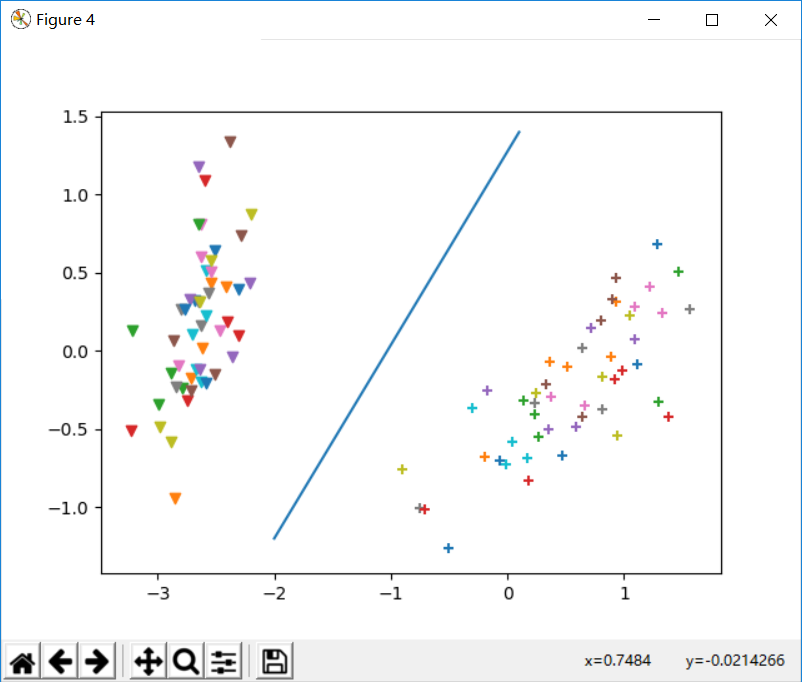


图2-4 分类面

4、对后两类数据进行分类

若对后两类即数据集中50-150的数据进行分类，绘制收敛曲线。并绘制其后验概率图，可以看出后两类之间是有重合的。

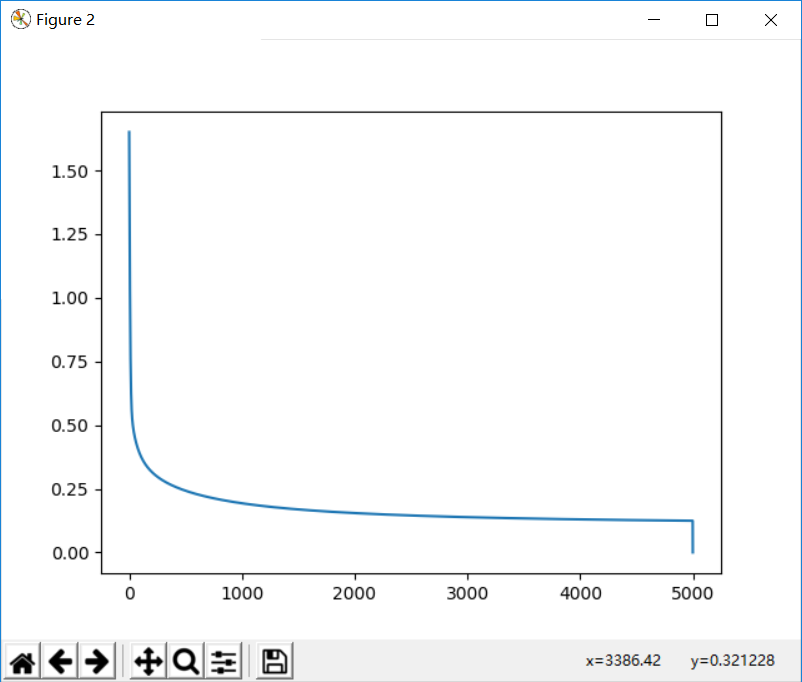


图2-4-1 收敛曲线

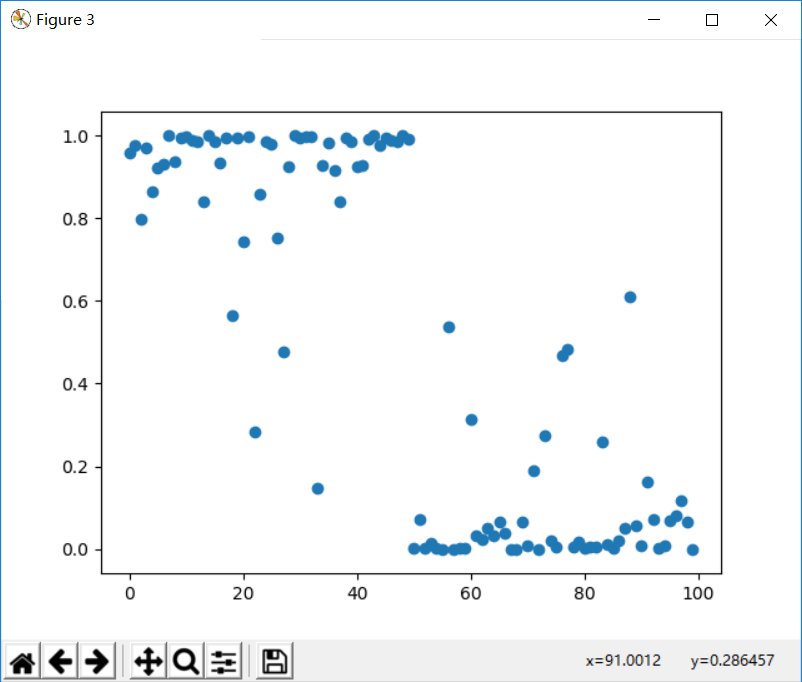


图2-4-2 后验概率图

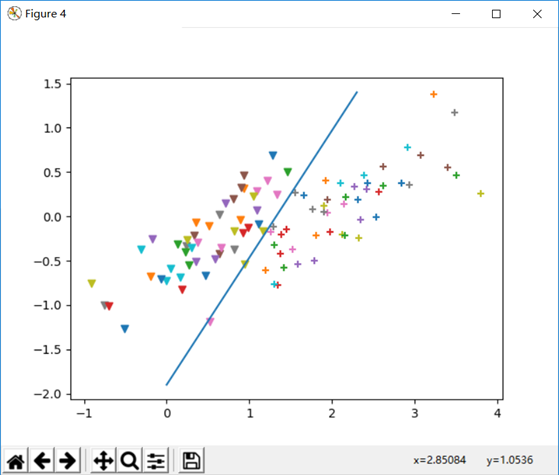


图2-4-3 分类面