课堂练习2

姓名：骆禹松　　班级：2018211129　　学号：2018210071

1 实验目的

- 设计用于三分类的线性模型并编程实现。
- 使用sklearn中提供的鸢尾花数据集作为训练和测试数据。
- 通过可视化方式呈现分类的决策边界。

2 实验原理

> 注：本实验我采用了multinomial logistic regression模型。

在统计学中，多项逻辑回归是一种将逻辑回归推广到多类别问题的分类方法。当所讨论的因变量具有两个以上的类别时，就会考虑使用多项逻辑回归模型。

假设我们有一组数据集$\{(x_i, y_i)\}_{i=1}^{n}$，其中每个x_i具有类标签y_i。回归问题是将预测的类别向量作为输入的线性函数，使均方误差最小，即

$$min \sum_{i=1}^{n} ||y_i - Wxi||^2$$

其中$W \in R^{k \times d}$是权重矩阵，$||\cdot||^2$是L_2范数的平方。

通过公式$a = Wx$可以预测输入的x的所属类别，其中a_i是输入在W的第i行上的投影（第i类的权重）。

3 实验步骤

3.1 实验数据

本次实验采用鸢尾花数据集作为训练和测试数据(只取数据集的前两列，即花萼长度和花萼宽度)。

```
import numpy as np
import matplotlib.pyplot as plt
from sklearn import datasets

iris = datasets.load_iris()
# 取前两个特征值
X = iris.data[:,:2]
Y = iris.target
```

如上代码块中，X表示提取的鸢尾花数据集，Y表示鸢尾花的种类（分别用0，1，2表示三种鸢尾花）。

3.2 数据处理

为了方便进行逻辑回归，此处需要对数组Y进行处理，将分类转换为二进制矩阵形式表示，即$y_i \in \{0,1\}^k$，假设某条鸢尾花数据属于第j类，则若$i=j$，$y_i=1$，否则为0.

```
Y_binary = np.zeros([X.shape[0],3])
for i in range(X.shape[0]):
    Y_binary[i, Y[i]] = 1
```

此时X有两列数据，而Y有三列数据，为方便使用最小二乘法求解矩阵方程$a=Wx$，还需要给X增加一列数据。

```
X0 = X[:,0]
X1 = X[:,1]
dot_xy = [X0,X1]
dot_xy = np.array(dot_xy)
dot_xy = dot_xy.T
dot_xy = np.insert(dot_xy,2,values=1,axis=1)
```

3.3 最小二乘法求解矩阵方程（创建模型）

本次实验使用最小二乘法来求权重矩阵。

```
W = np.linalg.lstsq(dot_xy, Y_binary, rcond=None)
W = W[0]
```

3.4 生成测试数据

为检验模型效果，这里需要先生成一组测试数据。分别以鸢尾花数据集中每列数据的最大最小值作为边界生成一组等间隔数组，并转换为合适的尺寸。

```
[X_test2, X_test1] = np.meshgrid(np.arange(X1.min()-1, X1.max()+1, 0.02), \
np.arange(X0.min()-1, X0.max()+1, 0.02))
m = X_test1.shape[0]
n = X_test1.shape[1]
X_test = np.ones([m*n,3])
X_test1 = X_test1.reshape([m*n,1])
X_test2 = X_test2.reshape([m*n,1])
for i in range(m*n):
    X_test[i,0] = X_test1[i]
    X_test[i,1] = X_test2[i]
```

3.5 预测分类

利用方程$a = Wx$来预测每个测试数据的类别。

```
W_test = X_test.dot(W)
Y_predict = np.argmax(W_test,axis=1)
Y_predict = Y_predict.reshape(m,n)
Y_predict = Y_predict.T
```

3.6 绘制图像

首先绘制提取出的鸢尾花特征数据图。

```
plt.scatter(X0, X1, c=Y, cmap=plt.cm.coolwarm, s=20, edgecolors="k")
```

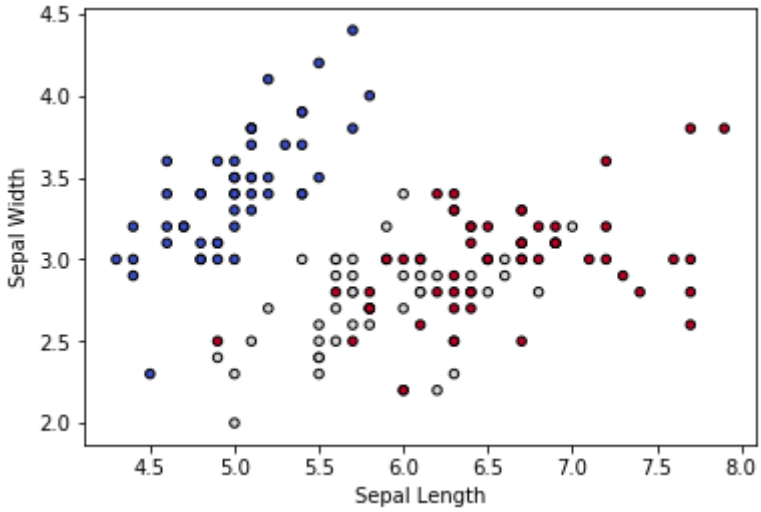
然后绘制预测的测试数据分类图，即通过可视化的形式呈现分类的决策边界。

```
plt.contourf(X_test1, X_test2, Y_predict, cmap=plt.cm.coolwarm, alpha=0.8)
plt.xlim(X_test1.min(), X_test1.max())
plt.ylim(X_test2.min(), X_test2.max())
plt.xlabel("Sepal Length")
plt.ylabel("Sepal Width")
```

4 实验结果

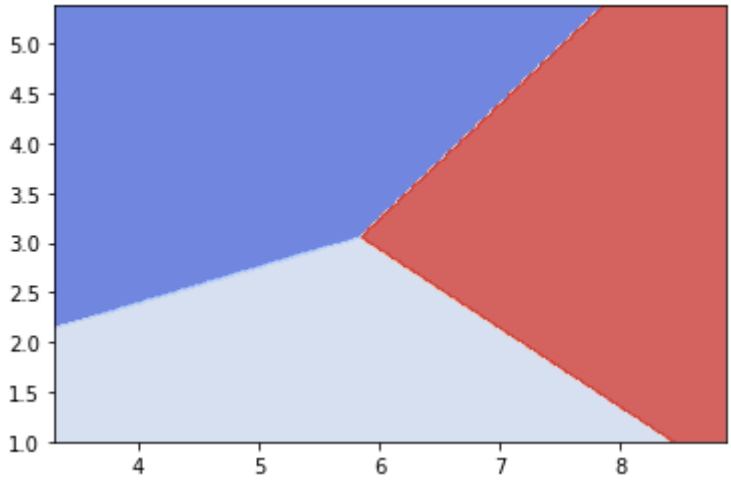
4.1 鸢尾花特征数据图

利用章节3.6中的代码可以绘制提取出的鸢尾花特征数据图，结果如下图所示：



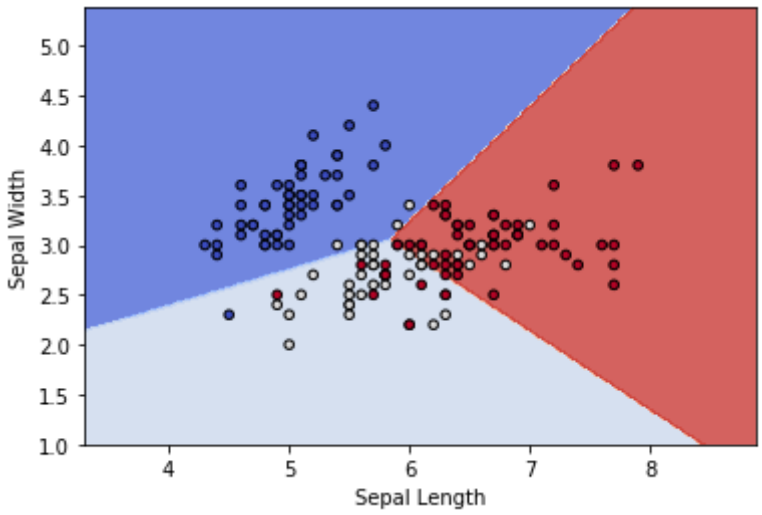
4.2 决策边界绘制

利用章节3.6中的代码可以绘制出分类的决策边界，结果如下图所示：



4.3 图像合成

将章节4.1和4.2中的图像组合在一起展示，结果如下图所示：



5 实验感悟

通过本次实验，我更深入地理解了逻辑回归模型，学会了使用线性模型进行分类决策。