太原理工大学
TAIYUAN UNIVERSITY OF TECHNOLOGY

《大数据挖掘与分析》实验指导书

学院：大数据学院

时间：2019 年 12 月

实验一 线性回归和 logistic 回归应用

一、实验目的

1. 理解线性回归模型

2. 会编写代码使用线性回归模型

3. 理解 logistic 回归模型

4. 会编写代码使用 logistic 回归模型

二、实验要求

1. 利用给定数据集独立完成实验

2. 书写实验报告书

3. 拓展任务：自己编写代码实现线性回归模型和 logistic 回归模型

三、实验内容

1.使用线性回归模型（用 sklearn 包和 matplotlib.pyplot 包）

（1）用点图展示训练集（数据集为 data.csv）

（2）创建并拟合模型

（3）用不同的颜色分别表示训练集点和预测值线（用 data 数据集第一列作为输入来预测）

示例：

```
from sklearn.linear_model import LinearRegression
import numpy as np
import matplotlib.pyplot as plt
```

```
# 画图
plt.plot(x_data, y_data, 'b.')
plt.plot(x_data, model.predict(x_data), 'r')
plt.show()
```

2.使用 logistic 回归模型（用 sklearn 包和 classification_report 包）

（1）将 21 维测试数据和第 22 列标签分别放入训练集和标签集(训练集为 logistictraining.txt，训练集为 logistictest.txt)

（2）创建并拟合模型

（3）用测试集进行测试，并输出正确率

示例：

```
from sklearn.linear_model import LogisticRegression

def colicSklearn():
    frTrain = open('horseColicTraining.txt')                    #打开训练集
    frTest = open('horseColicTest.txt')                         #打开测试集
    trainingSet = []; trainingLabels = []
    testSet = []; testLabels = []
    for line in frTrain.readlines():
        currLine = line.strip().split('\t')
        lineArr = []
```

3.拓展任务

（1）自己编写代码实现线性回归模型用

（2）自己编写代码实现 logistic 回归模型

示例：

```
# 载入数据
data = np.genfromtxt("data.csv", delimiter=",")
```

```
# 学习率learning rate
lr = 0.0001
# 截距
b = 0
# 斜率
k = 0
# 最大迭代次数
epochs = 50

# 最小二乘法
def compute_error(b, k, x_data, y_data):
```

```
def gradient_descent_runner(x_data, y_data, b, k, lr, epochs):
    # 计算总数据量
    m = float(len(x_data))
    # 循环epochs次
    for i in range(epochs):
```