**线性回归模型实验指导书**

一、实验目的

1. 熟悉如何通过Python实现机器学习算法；
2. 了解线性回归模型相关知识。

二、实验环境

系统：windows

软件：Python

示例代码版本：seaborn==0.11.1

Python 3.8.3

numpy==1.19.5

三、实验内容

1. 一元线性回归

数据集ex1data1.txt,有97个样本数据，每个样本有两列：人口和经济效益，通过实验，熟悉Python的一些基本函数和数据可视化方法。用最小二乘“参数估计”方法对模型进行求解。

2.多元线性回归

数据集ex1data2.txt中，有47个样本数据，每个数据有3列：房子大小，房间数目，房价。通过房子大小和房间数目作为样本特征，房价作为样本输出标签，通过线性回归模型来预测房价。请预测一个两间房，1300平方尺的房子的房价。

3. 标准数据集分析

数据集BJ\_AIR\_data来自于机器学习数据集仓库UCI Machine Learning Repository（链接：http://archive.ics.uci.edu/ml/datasets.php）。数据集中包含有2010-2014年北京美国大使馆收集的空气污染PM2.5数据及对应的气象数据。原始数据集中的缺失数据已经剔除。为了简化数据，部分属性值也已经省略。请根据给出的数据集简单分析这期间影响北京空气质量的主要因素。

四、实验步骤

（1，2参考.ipynb文件。）

1. 一元线性回归：
2. 导入数据集1中的数据，画出数据集1中数据的散点图。

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

df = pd.read\_csv('ex1data1.txt',names=['人口','利润'])

fig=plt.figure()

plt.rcParams['font.sans-serif'] = ['SimHei']

plt.rcParams['axes.unicode\_minus'] = False

sns.set(font='SimHei',font\_scale=1.5)

sns.lmplot('人口','利润',data=df,size=6,fit\_reg = False)

plt.show()

文本

描述已自动生成

1. 用最小二乘参数估计进行模型求解。

编程实现：

1. 根据教材P54页公式(3.7),(3.8)，求出w和b的最优解。
2. 根据公式(3.3)计算出y的预测值。

import numpy as np

def computeCost (X,y,theta):

inner=np.power((X\*theta.T)-y,2)

return np.sum(inner)/(2\*len(X))

df.insert(0,'ONE',1)

cols=df.shape[1]

X=df.iloc[:,0:cols-1]

y=df.iloc[:,cols-1:cols]

X = np.matrix(X.values)

y = np.matrix(y.values)

theta = np.matrix(np.array([0,0]))

computeCost(X,y,theta)

def gradientDescent(X,y,theta,alpha,iters):

temp=np.matrix(np.zeros(theta.shape))

parameters= int(theta.ravel().shape[1])

cost=np.zeros(iters)

for i in range(iters):

error=(X\*theta.T)-y

for j in range(parameters):

term = np.multiply(error, X[:,j])

temp[0,j] = theta[0,j] - ((alpha / len(X)) \* np.sum(term))

theta=temp

cost[i] = computeCost(X,y,theta)

return theta,cost

alpha=0.01

iters=1500

g, cost = gradientDescent(X, y, theta, alpha, iters)

computeCost(X, y, g)

1. 作图画出原始数据和预测结果，进行对比（请通过Matplotlib，熟悉plot语句用法，比如横轴、纵轴名称，用不同的颜色和形状区分原始数据和预测结果等）

x = np.linspace(df.人口.min(),df.人口.max(),100)#以人口最小值为起点，最大值为终点，创建元素个数为100的等差数列

f = g[0,0] + (g[0,1] \* x) #f是假设函数H

fig, ax = plt.subplots(figsize=(12,8))#以其他关键字参数\*\*fig\_kw来创建图

#figsize=(a,b):figsize 设置图形的大小,b为图形的宽,b为图形的高,单位为英寸

ax.plot(x, f, 'r', label='Prediction') #设置点的横坐标，纵坐标，用红色线，并且设置Prediction为关键字参数

ax.scatter(df.人口, df.利润, label='Traning Data') #以人口为横坐标，利润为纵坐标并且设置Traning Data为关键字参数

ax.legend(loc=2) #legend为显示图例函数，loc为设置图例显示的位置，loc=2即在左上方

ax.set\_xlabel('Population') #设置x轴变量

ax.set\_ylabel('Profit') #设置x轴变量

ax.set\_title('Predicted Profit vs. Population Size') #设置表头

plt.show()

ax=sns.lineplot(data=cost,err\_style='band',ci=[10,50,90],n\_boot=iters) #函数参数解释见笔记

ax.set\_xlabel("inters")

ax.set\_ylabel("cost")

plt.show()

#让我们换一种画图的方法

fig, ax = plt.subplots(figsize=(12,8)) #以其他关键字参数\*\*fig\_kw来创建图

#figsize=(a,b):figsize 设置图形的大小,b为图形的宽,b为图形的高,单位为英寸

ax.plot(np.arange(iters), cost, 'b') #作图:以迭代次数为x，代价函数值为y,线条颜色为红色

ax.set\_xlabel('Iterations') #设置x轴变量

ax.set\_ylabel('Cost') #设置y轴变量

ax.set\_title('Error vs. Training Epoch') #设置表头

plt.show()

多元线性回归：

1. 根据教材P55页多元线性回归原理，用最小二乘法对参数进行估计，根据公式(3.11)求出最优解，学得模型。
2. 预测房价，比如：一个两间房，1300平方尺的房子的房价。

请参考.ipynb文件

3. 标准数据集分析

请同学们基于回归模型进行简单的分析。

第三问需同学们自行完成

四、实验报告

实验报告要求：

1. 使用深圳大学实验报告模板

2. 要有运行结果截图

3. 建议附上代码，并以注释方式记录写代码的思考过程

4. 截止日期2024年10月15日11:59pm, [发送电子版至TA邮箱iamironmancc@qq.com，格式：2024](mailto:发送电子版至TA邮箱jecdlinjw@126.com，格式：2022)机器学习第一次实验报告-姓名。