云南大学软件学院
实 验 报 告

姓名：王子陈 学号：20171050008 专业：电子科学与技术 日期：2019/10/23 成绩：_____
任课教师：谢仲文

数据挖掘技术 实验三

一、实验目的

1. 掌握线性回归的直线拟合方法。
2. 选择一种编程语言实现最小二乘法的直线拟合。

二、实验内容

1．在一个简单的、虚拟的数据集（训练集）上应用最小二乘法进行拟合。该数据集如下：

	自变量 1	自变量 2	自变量 3	因变量
1	1	3	-7	-7.5
2	2	5	4	5.2
3	-3	-7	-2	-7.5
4	1	4	-12	-15

2．计算误差。

3. 谈谈你对线性回归的认识。

三、实验要求

1. 完成实验内容，源码作为实验报告附件一起打为一个压缩包提供。该压缩包要包含实验报告、代码文件。
2. 关键部分要求有注释，注释量不低于 20%
3. 要求独立完成，不得抄袭代码。

四、关键实验步骤（请粘贴关键步骤、代码、实验结果）

使用 Numpy 库中对矩阵运算的函数,得到系数矩阵:

根据最小二乘法的结论：$B=(X'X)^{-1}X'Y$

```
import numpy as np
x_list=np.reshape(np.array([[1,3,-7],[2,5,4],[-3,-7,-2],[1,4,-12]]),(4,3))#得到 4x3 的矩阵
y_list=[-7.5,5.2,-7.5,-15]
trans_x=np.transpose(x_list)   #X 矩阵转置
mat_mul=np.mat(trans_x@x_list)   #X 矩阵与其转置矩阵叉乘
inverse=mat_mul.I   #求矩阵的逆
B=np.mat(inverse@trans_x@y_list)
```

得到 B:
[[12.01208791 -4.35934066 0.82527473]]

各个矩阵的直观显示:
x_list: [[1 3 -7]
 [2 5 4]
 [-3 -7 -2]
 [1 4 -12]]

X'=[[1 2 -3 1]
 [3 5 -7 4]
 [-7 4 -2 -12]]

y_list: [-7.5, 5.2, -7.5, -15]

X'X= [[15 38 -5]
 [38 99 -35]
 [-5 -35 213]]

$(X'X)^{-1}$=[[16.78951817 -6.69399831 -0.70583263]
 [-6.69399831 2.67962806 0.28317836]
 [-0.70583263 0.28317836 0.03465765]]

所以根据上面得到的系数矩阵 B，拟合出来的直线为：
y=12.012*x1-4.359*x2+0.825*x3
误差：

y 观测值	y 理论值	误差
-7.5	-6.84	0.66
5.2	5.529	0.329
-7.5	-7.173	0.327
-15	-15.324	0.324

ε =0.8693365

使用 sklearn 的 LinearRegression 模型做拟合

1. 导入包:

```
import pandas as pd   #用 pandas 和 numpy 对数据进行操作
import numpy as np
import matplotlib.pyplot as plt   #用 matplotlib 进行图像化
from pandas import DataFrame,Series
from sklearn.model_selection import train_test_split   #使用 sklearn 进行数据集训练与模型导入
from sklearn.linear_model import LinearRegression
```

import seaborn as sns

2. 通过 read_csv 导入数据,返回的数据类型是 DataFrame

data_set=pd.read_csv("C:/Users/MY/Desktop/数据挖掘实验/实验 3：最小二乘法/data_set.csv",engine='python')　#文件名含有中文,指定 engine 为 python

```
   自变量1  自变量2  自变量3   因变量
0     1     3    -7  -7.5
1     2     5     4   5.2
2    -3    -7    -2  -7.5
3     1     4   -12 -15.0
<class 'pandas.core.frame.DataFrame'>
```

3. 数据检验：判断是否可以做线性回归

a. 输出相关系数（0-0.3 弱相关；0.3-0.6 中相关；0.6-1 强相关）

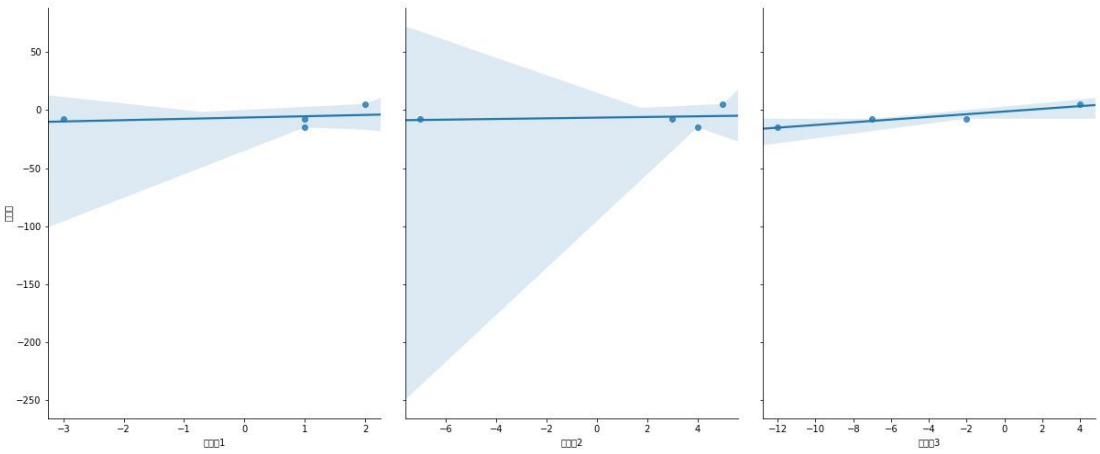
print(data_set.corr())

```
          自变量1       自变量2       自变量3       因变量
自变量1  1.000000  0.993584 -0.016460  0.297713
自变量2  0.993584  1.000000 -0.120343  0.193104
自变量3 -0.016460 -0.120343  1.000000  0.945763
因变量   0.297713  0.193104  0.945763  1.000000
```

从 corr 表中看出，自变量 3 和因变量是有比较强的线性关系的，而自变量 2，3 与因变量相关性较弱

b. 通过 seaborn 添加一条最佳拟合直线和 95%的置信带，直观判断相关关系

sns.pairplot(data_set,x_vars=['自变量 1','自变量 2','自变量 3'],y_vars='因变量',size=7,aspect=0.8,kind='reg')



4. 划分训练集的特征值与标签

df_feat=data_set.ix[:,:3]　#返回前 3 列自变量

df_label=data_set[[‘’]]　#返回的是 DataFrame 类型

5. 把训练集中的特征值与标签放入 LinearRegression()模型中且使用 fit 函数进

行训练,在模型训练完成后会得到对应的线性回归方程式,需要利用函数中的 intercept_与 coef_。

```
model=LinearRegression()
model.fit(df_feat,df_label)
a=model.intercept_
b=model.coef_
```

6.得到结果：

最佳拟合线:截距 [-0.76666667] 回归系数 [[8.42307692 -2.91025641 0.91794872]]

即所得的多元线性回归模型的函数为：y= -0.767+8.423*x_1 – 2.910x_2 +0.918x_3

7. 误差

y 观测值	y 理论值	误差
-7.5	-7.5	0
5.2	5.201	0.001
-7.5	-7.502	0.002
-15	-15	0

ε =0.002236

sklearn 的 LinearRegression 模型使用的参数估计方法应该也是普通最小二乘法,但是用 sklearn 拟合的误差很小.

对线性回归的认识：

线性回归是利用数理统计中回归分析，来确定两种或两种以上变量间相互依赖的定量关系的一种统计分析方法。其表达形式为 y = Bx+ ε， ε 为误差服从均值为 0 的正态分布。如果回归分析中，只包括一个自变量和一个因变量，且二者的关系可用一条直线近似表示，这种回归分析称为一元线性回归分析。如果回归分析中包括两个或两个以上的自变量，且因变量和自变量之间是线性关系，则称为多元线性回归分析。

线性回归拟合出来的因变量值可以和样本数据给定的因变量值存在差异, 即允许模型拟合存在误差。为了达到最佳的拟合效果, 方法是用最小二乘法找到使得理论值与观测值(即误差 ε)的平方和最小的自变量系数。进而得到与观测数据匹配的最佳函数。