浙江水利水电学院

实验报告

（ 18-19 学年 2 学期）

|  |  |
| --- | --- |
| 课 程 名 称： | 大数据基础 |
| 班 级： | 软工S18-1 |
| 学 号： | 2018b31007 |
| 姓 名： | 曹愉 |
| 实验室（中心）名 称： | 系统开发分室 |
| 教学单位： | 信艺学院 |

2019 年 3 月 28 日

实 验 名 称：线性回归与逻辑回归 指导教师：卢克

实 验 日 期：2019.3.28 地 点：现南401

同组学生姓名：无

|  |
| --- |
| 实验内容及要求：  采用Python语言实现以下问题：   1. 线性回归：银行贷款额度预测，根据给定的样本银行贷款数据来构建线性回归模型，通过年龄和收入来预测银行的贷款额度。 2. 逻辑回归：信用卡欺诈检测，根据给定的信用卡样本数据来选择构建最优逻辑回归模型。 |
| 主要仪器名称及型号：  （1） 计算机：Windows 7，i5 3Ghz,4GB内存  （2） 开发工具：Jupyter Notebook |
| 实验过程：（可附页）  问题1  1.先利用pandas库从loan.txt中取出数据。  import pandas as pd  import numpy as np  data=pd.read\_csv("loan.txt")  data.head()  2.创建线性回归模型，并将数据传入模型进行训练。  from sklearn import linear\_model  from mpl\_toolkits.mplot3d import Axes3D  import matplotlib.pyplot as plt  regr = linear\_model.LinearRegression()  X=data.loc[:,data.columns!='limit']  y=data.loc[:,data.columns=='limit']  regr.fit(X,y)  a,b=regr.coef\_,regr.intercept\_  x=np.array([[5000,28]])  print(regr.predict(x))  最后传入年龄和收入来预测银行的贷款额度。  问题2   1. 导入数据   import pandas as pd  import matplotlib.pyplot as plt  import numpy as np  data = pd.read\_csv(".\data\creditcard.csv")  data.head()   1. 发现Amount数据与其他组的数据大小差异过大进行数据标准化处理。   利用StandardScaler库函数  from sklearn.preprocessing import StandardScaler  data['normAmount'] = StandardScaler().fit\_transform(data['Amount'].reshape(-1, 1))  data = data.drop(['Time','Amount'],axis=1)  data.head()   1. Class类数据为0和为1的数据量差异过大，因此需要平衡样本。   平衡样本有两种方式，一种是下采样策略，让0和1样本一样少；另外一种是过采样策略，让1样本生成到与0同样多。  因为过采样计算数据量很大，所以先采用下采样进行实验。  X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size = 0.3, random\_state = 0)   1. 用70%的样本进行训练，剩下30%用于测验模型准确度。   X\_train, X\_test, y\_train, y\_test = train\_test\_split(X,y,test\_size = 0.3, random\_state = 0)   1. 添加正则化惩罚项参数，提高模型的泛化能力，避免过拟合。并计算Recall值判断模型训练结果，将采样数据分为5份进行交叉验证，并显示Recall值。     6.使用Recall值最高的模型进行准确度计算，生成混淆矩阵分别计算采样数据和所有数据的准确度,并可以尝试调整逻辑回归阈值从而对结果造成影响。 |
| 实验数据记录或图片：（可附页）  以五千工资和二十八岁代入模型后得出的结果    数据标准化后的normAmount    下采样后样本平衡后的数量    添加惩罚参数后对模型的训练实验结果 |
| 实验结论：  这次试验我们第一次接触了较大样本的大数据挖掘，通过这次实验我们知道了大数据技术中的线性回归和逻辑回归这两种最基础的算法，以及他们相对应的模型的操作和使用。  因为上节课已经基础讲解了两种回归算法的原理，所以这节课我们直接使用python库中的相关函数来进行回归模型的构建，这样要比一步一步的构建模型来的方便得多，但是基础的模型并不能做出很好的拟合以及泛化效果。  所以这次实验我们更多的是调参以及对数据进行归一化标准化这种前期的数据处理，而这些步骤也是在线性回归和逻辑回归这种大数据技术中很重要的部分，虽然都是构建模型，但是不同的参数对整个实验结果的影响则是十分的巨大。这次实验也给我们带来了很多的启发，对这些新技术也有了更多的了解。  希望以后也能在这种实验中探究更加有趣的技术，学习到更多的解决现实问题的方法。利用python语言和大数据技术实现更多的利用方式。 |
| 教师评语：  *从观察能力、实验能力、分析和解决问题的能力、实验结果、实验报告表达能力、实验结论等方面评价。*  *可手写或打印，若为电子稿则首行缩进两字符，宋体小四，1.5倍行距*  *字数不少于30字*  成绩： 批阅教师：  年 月 日 |