程序说明

材料学院陆子良516051910046

March 26, 2019

1 程序概述及运行

1.1 程序概述

本程序旨在利用python为工具，通过数据获取，数据处理，可视化与数据拟合，尝试得到上海二手房交易价格与交易房屋属性之间的关系。运用python应用于当下流行的数据处理与分析领域，且与实际问题相结合，具有一定的现实意义。

1.2 程序运行

程序运行环境需要**python3.5**或以上版本！！！

程序需引用的标准库有：sqlite3;time。

需额外的安装第三方库有:requests,bs4,re,numpy,pandas,matplotlib,seaborn,sklearn。

本项目在IDLE中直接运行main.py文件即可，由于数据分析的工作主要体现在过程中，故最后结果以分析报告（数据分析报告.pdf）的形式给出，程序只输出过程中得到的数据和可视化图像。

2 程序结构

本程序主要包括爬虫、清洗、可视化、特征选择、数据拟合五个模块及最后的main.py主程序。

2.1 爬虫模块

爬虫模块对应”class-spider.py”文件，主要功能是通过爬虫通过链家网站爬取上海二手房交易信息，爬取百度地图获取房屋的位置信息，从而完整的获得二手房的房屋单价、面积、房龄等各种信息，再通过python 内置的数据库sqlite3实行实时存储。此模块采用及时存储的方法，在爬取到一条信息后，便立即存储在数据库中，以防止因为各种原因导致的网络中断引起的数据丢失。

2.2 清洗模块

清洗模块对应"class-wash.py"文件,主要功能是将爬虫获取的初始数据进行拆分，提取，过滤归一化等，处理为适宜可视化和拟合的数据，为了便于使用者查看，将最后清洗完成的数据保存为excel文件。

2.3 可视化模块

可视化模块对应"class-paint.py"文件，主要功能是绘制散点图，直方图，箱型图，热力图等将实现数据及数据之间的关系的可视化，便于使用者了解房产数据的特征，并为选取特定的特征值进行拟合做好准备。

2.4 特征选择模块

特征选择模块对应"class-feature-select.py"文件，主要功能是基于相关系数热力图的绘制结果，选取相关系数较高的特征值进行拟合。由于本人学识有限，只对部分变量进行了简单处理，将最后选择的特征值保存在了excel文件中。

2.5 数据拟合模块

特征选择模块对应"class-predict.py"文件，主要功能是使用多种回归算法对房屋价格和相关特征值进行拟合，利用十折交叉验证得到准确率并进行比较，并导出最后结果。

2.6 主程序main.py

主程序main.py通过调用各个模块，实现数据分析的全过程，具体见程序中注释。

3 注意事项

- 本程序采取的数据是已经于2019年3月1日所爬取并存储在data.db和RawData.xls里的数据，故程序运行时不需运行爬虫部分。如需验证爬虫程序，爬虫运行时间较长，约为**2h**左右，请保持网络畅通。且更新后的网页的数据和最后拟合结果可能与原版本不符。**如爬虫运行因问题导致中断，请更改main.py程序第18行，将for循环更改为当前中断页面至300。**

  例：程序显示在爬取第150页第x行数据后报错，则将第18行改为$for\ i\ in\ range$（150：COUNT）:后再次运行，直至爬取完成。

- 本程序推荐于windows系统下运行，如使用Linux系统将导致拟合阶段不明原因的错误。

- 请确保电脑中字体库齐全，若图1的中文显示异常，请参照分析报告中的相关图像。

由于本人学识有限，数据分析及程序编写中多有瑕疵，敬请老师、助教及后续使用者指正！