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 4 月 4 日

实 验 名 称：K近邻分析 指导教师：卢克

实 验 日 期：2019.4.4 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：  K近邻分析：  对于给定的租房数据，进行房间出租价格预测。  数据特征：  \* accommodates: 可以容纳的旅客  \* bedrooms: 卧室的数量  \* bathrooms: 厕所的数量  \* beds: 床的数量  \* price: 每晚的费用  \* minimum\_nights: 客人最少租了几天  \* maximum\_nights: 客人最多租了几天  \* number\_of\_reviews: 评论的数量 要求：对样本数据进行标准化或归一化处理；  1. 要求采用sklearn.neighbors的KNeighborsRegressor对训练数据进行建模； 2. 采用root mean squared error (RMSE)均方根误差来对模型进行评价。 |
| 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）   1. 先用pandas库取出保存在csv文件中的数据      1. 对实验数据进行标准化处理，使训练结果更加精确      1. 对数据区分训练集和测试集      1. 引入Sklearn库，进行K近邻算法计算      1. 首先使用单一特征进行预测      1. 加入更多特征进行预测      1. 对特征进行重组，选出最优特征组      1. 从两项开始，不同特征混合分组      1. 选出最佳特征组，及平均预测偏差 |
| 实验数据记录或图片：（可附页）   1. 数据标准化处理后      1. 两项特征计算      1. 三项特征计算      1. 四项特征计算      1. 五项特征计算      1. 六项特征计算      1. 七项特征计算      1. 最优特征组 |
| 实验结论：  这次试验我们再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K近邻预测这种较为基础的预测算法，以及与它相对应的模型的操作和使用。  因为课上已经基础讲解了K近邻算法的原理，所以实验中我们直接使用python库中的相关函数来进行K近邻模型的构建，这样要比一步一步的手动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，所以我们对基础的数据进行了二次加工，通过标准化使数据集中的标准方差均为1，这样做出来的模型才可以更加精确。  这次实验我们更多的是调参以及对数据进行标准化这种前期的数据处理，最后才在有了用五项特征值得出最为接近的模型，而这些步骤也是在K近邻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 |