北 京 交 通 大 学

课程设计开题报告表

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 课程名称 | | 大数据概论 | 学院 | 软件学院 软件工程专业 |
| 小组成员 | | 贺思超、陈杰、韩熔、解世超 | | |
| 课程设计题目 | | 基于往年气温数据预测未来天气气温 | | |
| 课程设计研究内容及  选题依据 | 天气预测一直以来就是人们的愿景以及痛点之一，有效的天气预测有利于政府对资源的统筹调度。  本实验拟采用Python语言，从www.meteomanz.com网页爬取气温数据并存储为CSV文件，使用panda读取CSV文件并进行数据预处理，并采用Sklearn框架进行机器学习处理。最后使用随机数森林模型进行建模预测。 | | | |
| 课程设计关键问题预期目标 | 基于天气预测的广泛应用和重要意义，以及其与大数据极强的关联性，本小组选择气温预测进行课程研究，旨在运用大数据分析与挖掘的理论与技术解决气温预测的相关问题。  由于大气运动的不确定性等因素，天气预测需要收集大量的数据用以分析处理，大数据分析与挖掘的相关理论与技术对天气预测有着重要的作用。随着社会发展，国家防灾减灾、保障社会经济发展和人民安康福祉对天气预测提出了更高的需要，发展以“数字化、精细化”为主要特征的天气预测技术是实现天气业务现代化的必然要求，大数据分析必然在其中起到关键作用。 | | | |
| 课程设计的研究计划及预期解决方案 | 要包括数据来源、存储（excel或是数据库都可以）、模型、评估、可视化等。以上内容是小组内部讨论过的，是暂时的方案，后续可以更改，我需要了解你们的思路是否正确。  数据来源：http://www.meteomanz.com/  存储：CSV文件  模型：随机森林模型  评估：MAE  可视化：Matplot绘制折线图 | | | |
| 技术难点及  解决方案 | 随机森林在解决回归问题时，并没有像它在分类中表现的那么好，这是因为它并不能给出一个连续的输出。当进行回归时，随机森林不能够做出超越训练集数据范围的预测，这可能导致在某些特定噪声的数据进行建模时出现过度拟合。  对于许多统计建模者来说，随机森林给人的感觉就像一个黑盒子，你无法控制模型内部的运行。只能在不同的参数和随机种子之间进行尝试。  解决方案：  1.增加数据集数量 | | | |
| 组内分工及  开题  当前进展 | 数据挖掘  数据处理  数据建模  主程序构建 | | | |