# 《数据挖掘》课程大作业

## 2020 年 11 月 18 日

**注：正文 16 页以上，请按内容要求完成、命名规则整理，请分班收集后交到学活 207，另外将电子档整理后打包发至邮箱：18838910786@163.com；作业提交截至日期：2020 年 12 月 31 日，逾期不接收。**

## 1 作业目的

**(1)** 考查学生配置机器学习环境的能力。

**(2)** 考查学生数据建模的能力。

**(3)** 考查学生数据分析的能力。

## 2 参考资源

**(1)** 数据集参考资源：

- `https://www.kaggle.com/tags/data-cleaning`
- `https://makingnoiseandhearingthings.com/2018/04/19/datasets-for-data-cleaning-practice/`
- ……

**(2)** 代码及案例参考资源：

- `http://cookdata.cn/`
- ……

## 3 内容要求

**(1)** 给出配置机器学习环境的步骤；

**(2)** 自选问题，给出问题背景的介绍；

**(3)** 收集相关数据，并对数据集进行描述；

**(4)** 对数据进行清洗，并做可视化处理 (可选，看数据集是否需要)；

**(5)** 对数据进行建模，至少选择两种模型 (**核心内容**)；

**(6)** 对模型结果进行分析 **(核心内容)**；

**(7)** 核心代码模块请放入文档，所有代码请给出注释，并作为附件打包提交 **(重点内容)**；

# 4 结构要求

**(1)** 封面：包含个人信息，请参照模板填写个人及相关信息；

**(2)** 摘要：简述文章的内容，包括应用（研究）背景和意义、方法、主要结果；

**(3)** 目录；

**(4)** 正文：请参考如下结构

- 第一章内容包括：机器学习环境的配置；
- 第二章内容包括：选题背景及现状分析；
- 第三章内容包括：数据采集与清洗；
- 第四章内容包括：数据分析与建模；
- 第五章内容包括：总结和展望；

**(5)** 参考文献；

**(6)** 附件：包括代码。

# 5 排版格式要求

## 5.1 Latex 模板

`https://github.com/yuanhaizhuan/xtuthesis`

## 5.2 Word 模板

将群分享给大家。

## 5.3 命名规则

**(1)** 主目录命名规则：班级 + 学号 + 姓名 + 论文名称（**如：应数 1 班 +2017750312+ 小明 +COVID-19 疫情的数据科学实践)**；主目录下包括文档和附件目录；（**注：班级名称请统一为：韶峰班、应数 1 班、应数 2 班、统计 1 班、信计 1 班、信计 2 班、信计 3 班**）

**(2)** 文档命名规则：班级 + 学号 + 姓名 + 论文名称；

**(3)** 附件目录统一命名为: 附件；其附件目录文件请自行合理命名，具有可读性即可。

# 6 评分标准

| 项目 | 分值 | 优秀<br>90-100 | 良好<br>80-89 | 中等<br>70-79 | 及格<br>60-69 | 不及格<br>0-59 | 评分 |
|---|---|---|---|---|---|---|---|
| 调研论证 | 10 | 能独立查阅文献以及从事其它形式的调研，能较好地理解课题任务提出实施方案，有各类信息分析整理、从中获取新知识的能力 | 能阅读教师给出参考资料、文献外，还能阅读一些自选资料，能较好地分析各类信息，并提出较合理的实施方案 | 除阅读教师指定参考资料、文献外，能分析各类信息，有实施方案 | 阅读教师指定参考资料、文献，有实施方案 | 未完成教师指定的参考资料及文献，无信息分析整理，实施方案不合理 | |
| 技术水平与实际能力 | 20 | 设计（研究方法）合理，理论分析与计算正确，实验数据准备可靠，有较强的实际动手能力、分析能力和应用能力 | 设计（研究方法）比较合理，理论分析与计算正确，实验数据比较准确，有一定的实际动手能力、分析能力和应用能力 | 设计（研究方法）比较合理，理论分析与计算基本正确，实验数据基本准确，实际动手能力尚可 | 设计（研究方法）基本合理，理论分析与计算无大错 | 设计（研究方法）不合理，理论分析与计算有原则性或原理性错误，实验数据不可靠，实际动手能力差 | |
| 研究成果 | 20 | 对设计，研究的问题有较深刻分析或有独到之处，成果突出 | 对设计，研究的问题能正确分析或有新见解，成果比较突出 | 对研究的问题能提出自己的见解，成果有一定意义 | 对某些问题提出个人见解，并得出设计，研究结果 | 缺乏设计，研究能力，未取得任何成果 | |
| 研究创新 | 10 | 有重大改进或独特见解，有一定理论价值或实用价值 | 有较大改进或新颖的见解，理论性或实用性尚可 | 有一定改进或新的见解 | 有一定见解 | 观念陈旧，已有内容的重复 | |
| 论文撰写 | 15 | 结构严谨，逻辑性强，论述层次清晰，语言准确，文字流畅，完全符合规范化要求 | 结构合理，符合逻辑，文章层次分明，语言准确，文字流畅，达到规范化要求 | 结构基本合理，层次较为分明，文理通顺，基本达到规范化要求 | 结构基本合理，论证基本清楚，文字尚通顺，勉强达到规范化要求 | 内容空泛，结构混乱，文字表达不清，错别字较多，达不到规范化要求 | |
| 答辩情况 | 15 | 能简明扼要，重点突出地阐述论文的主要内容，能准确流利地回答相关问题 | 能比较流利，清晰地阐述论文的主要内容，能较恰当地回答与论文有关的问题 | 基本能叙述出论文的主要内容，答辩时回答问题基本准确 | 能阐明自己的基本观点，答辩时回答问题无重大错误 | 不能阐明自己的基本观点，主要问题答不出或有原则性，原理性错误 | |
| 学习态度 | 10 | 满勤，作业完成情况很好 | 满勤，作业完成情况较好 | 有缺课记录，作业完成情况一般 | 无故缺课较多，作业完成情况较差 | 无故缺课很多，作业完成情况糟糕 | |