暑期课程设计授课笔记

关于 2022 年暑期计算机科学与技术专业《信息系统集成与开发》课程的基本信息是，共有 8 个自然班，组合成 3 个教学班。每个教学班的课程主要（计科 2019 级）你的腾讯会议系统的用*户名必须包括学号，建议改为：班级+姓名+学号。例如，计科 1901 杨浩然 19851076。确信自己加入我们课程的微信群（按照不同授课班，共有 3 个不同的群，不要加*错群）。课程设计报告格式规定。

1 课前基本要求

2022 年暑期小学期《信息系统集成与开发》的学生是计算机科学与技术专业 2019 级，共 8 个自然班，3 个教学班。课前主要说明与要求如下：

（1）每次上课之前必须保证参会者姓名正确，否则需要修改。你的腾讯会议系统的用户名必须包括学号，建议改为：班级+姓名+学号。例如，计科 1901 杨浩然 19851076。

（2）确信自己加入我们课程的微信群（按照不同授课班，共有 3 个不同的群，不要加错群）。

（3）严格遵守课程设计报告格式规定。

4 机器学习应用系统与大数据分析。

5 6 月 24 日 13:00PM 提交课程报告给各班班长。

6 学习 Weka 系统（网站）。

7

课程要求与注意事项

（1）改名：你的腾讯会议系统用户名必须包括学号，建议改为：班级+姓名+学号。例如，计科 1904 刘紫琦 19851078。

（2）确信自己加入我们课程的微信群（任课教师按照不同授课班，共有 3 个不同的群，不要加错群）。

（3）课程设计报告要求严格遵循格式规定（这是评定成绩的依据之一）。

（4）建立人工智能应用系统：机器学习应用系统与大数据分析。

（5）2022 年 6 月 24 日 13:00PM 提交课程报告给各班班长。

（6）阅读 Weka 系统网站（或者其他中文网站），并撰写网站简介。注意。必须有参考文献著录与引用[2]。为保障校外实践教学工作顺利进行，现将学生校外实践过程中应承担的安全责任明确如下：

7 20220616A

3　课程设计报告的写作规范 2　确信自己加入我们课程的微信群（按照不同授课班，共有 3 个不同的群，不要加错群）。

（7）安装 Weka 系统，并尝试运行。

2　课程成绩评定依据

课程成绩评定依据划分为主要依据与次要依据。

2.1　主要考核依据

2022 年 6 月 24 日下午 13:00PM 提交给各班班长两个完全一样的课程设计报告（Word 版与 PDF 版），具体要求如下：

（1）格式正确；

（2）页数要求：25-35 页：

（3）结构合理，文笔流畅；

（4）设计内容合乎逻辑。

2.2　次要考核手段

我们的课程共 24 学时，从 2022 年 6 月 16 日上午开始，至 2022 年 6 月 23 日晚上结束。次要考核手段是你们在此期间的表现与阶段性成果。

（1）课堂活跃程度（腾讯会议聊天）；

（2）成员登记表（腾讯会议系统导出）；

（3）课程设计报告的过程性版本；

（4）补充性作业，例如，写一个程序，从一个纯文本文件顺序读取，并按照学号进行记数。

3　课程设计报告基本要求与撰写规范

7 20220616A 3　*课程设计报告的写作规范 2　确信自己加入我们课程的微信群（按照不同授课班，共有3个不同的群，不要加错群）*。

3.1　关于字体的规定

中文必须采用中文字体，一般使用宋体，特殊情况下使用楷体；英文一般采用 Times New Roman，涉及程序、代码、数据库及其字段、计算机命令等一律采用等宽字体 Consolas，其他特殊情况所采用的字体另行说明。注意，最常见的错误是英文使用中文字体。

3.2　课程设计报告大致结构

课程设计报告的组织由学生自己决定。建议大致结构如下：

（1）报告题目之下应简单介绍本课程设计报告的选题内容与意义，报告的基本内容，以及达到的目的等等。

（2）第 1 节要求介绍 Weka 网站的基本内容，也可以介绍一下其他人工智能应用、机器学习、或数据挖掘技术的其他网站。

（3）第 2 节要求介绍机器学习与数据挖掘的最基本的概念，以及 Weka 软件系统的基本结构与功能。

（4）第 3 节要求介绍数据集合的格式规范，以及转化方法，要求举例说明，并且每个教学班的不同同学必须至少选择一个 UCI datasets 网站之上的不同的数据集合。

（5）前 3 节内容不能超过你的正文的一半！

（6）第 4 节介绍你选择或指定的问题及其数据集合。

（7）以后的课程设计报告的内容自己决定。选题及其要求详见第 4 部分。

3.3 标点符号使用的基本要求

标点符号使用的最主要要求是一个段落结束之后有一个句号！一般地，中文之间使用中文标点符号，英文之间使用英文标点符号，中文和英文之间使用中文标点符号；在参考文献著录之中尽量全部使用英文标点符号。

3.4 关于表格的要求

关于表格的要求及其注意事项如下：

（1）表格的上、下各空一行。

（2）表格必须有编号和标题，并且居中排版于表格的**上方**。

（3）表格全文统一编号，或按节编号。

（4）表中文字（表头和表内容）：宋体，5 号字体（10.5 磅）；段落：居中，不要段前段后，行距为固定值 18 磅。

（5）表格属性：不指定表格宽度，对齐方式选择居中，文字环绕选择无。

（6）列：每列宽度根据需要设置，尽量小。

（7）在顶层菜单的【布局】下选择“查看网格线”按钮，并且选择表格为无边框与底纹。

（8）一律采用国际通行的三线格：将表格的最顶线和最低线改为 1.5 磅，选中表格属性，第 1 行之下的线为 0.5 磅。

4 课程设计题目及其分配原则

分配原则按照优先次序如下：

（1）鼓励学生**自由选题**，但是不能与其他同学重复，同时需要老师确认；

（2）**第 1 名**选择老师推荐题目的学生（即现场已确认），注意，相应题目的学生全部标明在脚注之中（务必请相关同学自己确认！）；

（3）其他大多数学生按照你自己的 8 位学号除以 29 的余数，就是老师指定分派给你的需要分析的数据集合，具体课程设计报告的题目**据此自己拟定**。

4.1 常见的数据集合网站

常见的数据集合网站如下：

（1）数据源泉网站：datafountain 网站地址 https://www.datafountain.cn/dataset。

（2）阿里云天池网站：aliyun.com。

（3）https://www.kaggle.com。

（4）其他也可以。

4.2 我们课程选择的数据集合

数据集合具体如下：

（1）美国金县二手房价数据集[1]：选自 datafountain 网站，共包含 21,614 个二手房价信息，数据集下载网址为 https://www.datafountain.cn/datasets/67。计科 1905 谷平阔 19851226 已经下载好数据集合，没有问题。

（2）美国居民收入数据集[2]：选自阿里云天池网站，共包含 30163 条数据。可从平台下载数据，数据集都来自于美国 1994 年人口普查数据库当中抽取的。数据集下载网址为：数据集-阿里云天池(aliyun.com)。

（3）认证日志与风险日志数据集[3]：来源于 DataFountain，该数据集提供了系统用户在访问应用系统产生的行为数据，共有 150000 个样本，具体包括 18 个变量。数据集下载网址为 https://www.datafountain.cn。计科 1904 路倩 19851207 第三题：https://www.datafountain.cn/datasets/6327，已经下载好数据集合，没有问题。

（4）信用卡获准的数据集[4]：选自 DataFountain 网站，共包含两个 excel 数据表，都来自于生活中真实的情况。提供各领域公开数据集的下载服务，数据集下载网址为 https://www.datafountain.cn/datasets/4599。

（5）澳大利亚降雨数据集（weatherAUS）5：选自 Kaggle 网站，该数据集包含来自众多澳大利亚气象站的约 10 年每日天气观测数据，共 145460 天的数

---

[1] 计科 1905 雷云凤 19851047；计科 1907 张起源 19851180。
[2] 计科 1908 汤冬江 19851277
[3] 计科 1904 路倩 19851207；计科 1908 刘炫纶 19851118。
[4] 计科 1908 崔义凡 19851077。
[5] 计科 1902 李学渊 19851228；计科 1908 周圣策 19851208；计科 1904 冯佳楠 19851167。

据。数据集下载网址为 https://www.kaggle.com/jsphyg/weather-dataset-rattle-package。

（6）国际象棋[6]：影响国际象棋获胜因素分析的数据集合使用大数据大赛 DF（Data fountain）中的 20058 条有关国际象棋线上赛事的信息，数据集下载网址为 https://www.datafountain.cn/。计科 1906 李铭雨 19851284 已经下载好数据集合，没有问题。

（7）Thera 银行的针对用户的电话营销活动记录数据集[7]：该数据集中总计共有数据样本 30,478 条。请计科 1903 刘佳红 19871039 等同学补充网址。

（8）马里兰交通违法行为数据集8：选自 DataFountain 网站，提供 AI 竞赛/大数据竞赛、人工智能数据集等，共包含 21 个行业、1,555 个数据集，都来自于社会中真实的数据信息。该网站平台包括实训、竞赛、社区、数据集四个方面。数据集下载网址为 https://www.datafountain.cn。

（9）中国联通套餐数据集合9：套餐主要指运营商推出的针对各种需要的客户而推出的服务产品和附加业务。计科 1905 娄建鹏 19851002 已经下载好数据集合，没有问题。所选数据集来自中国计算机协会与中国联通研究院发布的公开比赛数据集（https://www.datafountain.cn/competitions/311/datasets）。

（10）美国车祸数据集10：涵盖美国 49 个州，数据收集了 2016 年 2 月至 12 月的交通事故。数据的数据量超过十万条，数据来源：https://www.kaggle.com/sobhanmoosavi/us-accidents。计科 1904 冀强 19851177 已经下载好数据集合，没有问题。

（11）和鲸社区“某红书销售情况”数据集11：中国知名的第三方数据科学社区之一，也为数据科学的人才提供交流、切磋的机会。计科 1907 郑璇 19851091 已经下载好数据集合，没有问题。数据集共包含 29,452 个样本，7 个变量，网址为：https://www.heywhale.com/mw/dataset/5f1e3fde94d484002d2f6715/file。7 个变量：购买的金额、性别、年龄、活跃度、生命周期、R 指标、M 指标、F 指标；用于预测用户可能购买的金额。

---

[6] 计科 1907 任泽江 19851213。
[7] 计科 1907 赖宇华 19851269；计科 1906 尤子涵 19851125。
[8] 计科 1903 郑新灿 19851121；计科 1908 霍冠酉 19851215；计科 1906 马鸣怡 19851097。
[9] 计科 1903 袁海东 19851238；计科 1908 刘琳洁 19851295；计科 1905 付金伟 19851182。
[10] 计科 1902 赵君哲 19851072；计科 1907 贾尧 19851156；计科 1904 武德钊 19843044。
[11] 计科 1903 杨帆 19851129；计科 1908 马嘉荣 19851158。

（12）搜狗新闻数据集[12]：文本分类，论文数据集保存为.txt 格式，共有 5,000 篇文章，数据集来源地址为：http://www.sogou.com/labs/resource/ca.php。

（13）葡萄酒数据集合[13]：原始出处是 DataFountain 网站，此数据集是关于葡萄酒的一些评论。其中包含了 130K 条葡萄酒评论，包括了品种，位置，酒庄，价格和描述等因素。网址为 https://www. datafountain.cn/datasets/。数据为不同的国家消费者对许多红葡萄酒的评论。包括生产国、省份、城市、叙述、制造商、种类、价格等的一些重要细节。计科 1906 李笑瑞 19851212 已经下载好数据集合，**没有问题**。

（14）比特币历史价格数据集合[14]：数据来源于中国最大的人工智能开发者社区——阿里云天池，阿里云天池网址为：https://tianchi.aliyun.com。以比特币历史价格为基础分析未来的走势，采用了阿里云天池中的 4857377 条有关比特币历史价格的数据，数据中包含 8 个相关变量。计科 1904 马浩然 19851007 已经下载好数据集合，没有问题。

（15）广告点击率预估挑战赛数据集合[15]：此数据集为科大讯飞 2021 广告点击率预估挑战赛数据，网址：http://challenge.xfyun.cn/topic/info?type=Ad-click-through。也可来源于 DataFountain，包含训练集和测试集。其中测试集一共提供了 71,466 个样本，12 个特征字段；训练集一共提供了 391,825 个样本和 13 个特征。计科 1908 李嘉玉 19851066 已经下载好数据集合，没有问题。

（16）车辆贷款违约预测数据集合[16]：天池大数据大赛是由阿里巴巴集团赞助，是面向全球科研人员的高端算法大赛。通过开放大量数据和分布式计算资源，使所有参与者都有机会应用他们的算法来解决各种存在社会或商业中的问题。计科 1904 贺宇乐 19851096 已经下载好数据集合，没有问题。

（17）B 站播放量数据集合[17]：选取 2019 年 1 月-2020 年 3 月，B 站播放量过五万的视频的一些信息，主要特征有行业标签、视频编号、作者名、作者 id、视频的发布日期、视频链接、视频标题、投币数、总弹幕数、收藏数、点赞数、总评论数、分享数以及总播放量，共计 50,130 行数据。该数据集来源于和鲸社

---

[12] 计科 1908 朱晓萌 19851161；计科 1906 吴庭辉 19851270。
[13] 计科 1907 邹时勉 19851251；计科 1903 刘洪 19851036；计科 1906 李笑瑞 19851212。
[14] 计科 1908 刘学峰 19813216；计科 1903 陈锦鑫 17851012；计科 1904 王泽辰 19851068。
[15] 计科 1903 刘馨雅 19851193；计科 1906 张顺程 18851244；计科 1908 李琢 19851203。
[16] 计科 1901 安雳冰 19851250
[17] 计科 1903 高佳玉 19851287；计科 1906 陈子阳 19851082；计科 1907 卢龙腾 19851196。

区，数据集的网址为：https://www.heywhale.com/mw/dataset/5f6dbf6071c70000307f0e60。

（18）茶类产品数据集合[18]：基于用户评论的茶类产品分析，主要研究内容是消费者对于购买茶类产品的相关评论，同时还包括茶类产品的信息。本文所研究数据是从和鲸社区获取的，网址为：https://www.heywhale.com/。计科 1905 齐柄焱 19851243 已经下载好数据集合，没有问题。

（19）5G 用户的分析数据集合[19]：数据是从和鲸社区获取的，对 5G 用户的分析以及预测，该数据集包含 train_set.csv 和 train_label.csv 两张表，记录了用户是否使用 5G 业务。

（20）网上书籍销售数据集合[20]：数据来源于 CSDN 网站，CSDN 是世界上著名的中文 IT 技术交流平台，成立于 1999 年，包含资源下载等，拥有超过 600 万数据库，信息丰富可靠。计科 1908 袁绍斌 19851042 已经下载好数据集合，没有问题。本文所用的数据库网址是：https://download.csdn.net/download/quf2zy/10468265。数据以 CSV 格式存储共有 24,000 多条数据，共有 7 个特征值。

（21）航空系列数据[21]：数据集是从阿里云天池平台官网上下载的航空系列数据集，里面包含三个数据表，分为训练集，测试集，样本集等 3 个数据集，数据在一万条以上，包含众多航空系列因素标签。计科 1904 王盛 19851188 对于数据集合下载没有问题。

（22）电影数据集合[22]：Kaggle 作为数据发掘、算法分析，和数据竞赛的在线平台，它为参赛者免费提供数据集，定期举办竞赛并给予奖励，为数据分析师提供学习交流的平台。大赛网址为 https://www.kaggle.com/。计科 1905 齐安航 19851073 数据集下好了，没问题。

（23）建筑震后受损情况数据集合[23]：自天池大数据科研平台，数据集获取网址为 https://tianchi.aliyun.com/dataset/dataDetail?dataId=63678#1。天池作为阿里云旗下的大数据资源平台，致力于为云计算提供优质的技术人员。

---

[18] 计科 1901 林燕芬 19851255；计科 1904 徐晓庆 19851135；计科 1907 贾梦帆 19851139
[19] 计科 1901 董昭阳 19851220；计科 1904 朱博辉 19851183。
[20] 计科 1901 王迪 19851224；
[21] 计科 1903 薛雅琪 19851095；计科 1904 石雨晴 19851001；计科 1908 赵晗啸 19851137。
[22] 计科 1902 彭庆林 19851299。
[23] 计科 1906 张鹏 19851165；。

（24）企业员工幸福指数数据集（Wellbeing_and_lifestyle_data）[24]：来源于 Data Fountain（简称 DF 平台），此数据集包含 12,757 份数据，22 个属性，均来源于真是的问卷调查。数据集下载网址为：https://www.datafountain.cn/datasets/5038。

（25）宽带服务提供商的数据集[25]：针对研究宽带用户的流失情况，论文采用宽带服务提供商的数据集，包含 50 万条数据，其中主要包括用户在网时长、带宽选择、是否使用电话服务、担保收益、抱怨次数等，下载网址为 https://www.datafountain.cn/datasets/5177。计科 1906 马嘉琦 19851105 已经下载好数据集合，没有问题。

（26）自杀率相关数据集合[26]：来源于 DataFountain（简称 DF 平台）是北京数联众创科技有限公司旗下品牌，DF 平台专业提供赛题，数据分析等，数据集网址：https://www.datafountain.cn/datasets/36。数据集采用的是由世界卫生组织收集的由 1985 年到 2016 年的全球不同人群中的自杀率相关信息，包括 27820 个样本，*12 个特征值*。计科 1905 张阳 19851222 已经下载好数据集合，没有问题。

（27）航班信息数据集合[27]：受疫情影响的航空公司的航班延误和取消情况数据集选自 DF（datafountain.cn, 原 WID 竞赛平台），数据集是由美国运输部（DOT）的运输统计局会跟踪大型航空承运人运营的国内航班的准点运行情况。数据集下载的网址为：https://www.datafountain.cn/datasets/5450。计科 1904 融改弟 19851145 已经下载好数据集合，没有问题。

（28）员工疲劳度数据集合[28]：数据来源于 DataFountain。数据网址为：https://www.datafountain.cn。

（29）驾驶疲劳预测数据集合[29]：来源于阿里云天池福特竞赛，是 2012 年美国企业联合政府发起的，面向全世界的科研人员和大学生，是一个大数据挖掘检测的平台。论文的网址为 https://tianchi.aliyun.com/dataset/dataDetail?sataId=89345。论文采用的数据集分为两部分，一个是测试集，一个是训练集，都是以 CSV 形

[24] 计科 1902 胡海洋 19851037；计科 1904 高锦秋 19851018。
[25] 计科 1902 刘子琛 19851199；计科 1908 费天洪 18851042。
[26] 计科 1902 马新竺 19851252；计科 1906 黄唯哲 19851241。
[27] 计科 1904 融改弟 19851145；计科 1907 黄贺 19813148
[28] 计科 1902 何佳怡 19851075
[29] 计科 1901 郭金泽 19851067。

式存储。测试集共有 120,841 条数据，训练集共有 604,330 条数据，有 33 个数值型的特征值。计科 1908 马长春 19851138 已经下载好数据集合，没有问题。

其他题目：

datafountain 网站，共包含 21,614 个二手房价信息，数据集下载网址为 https://www.datafountain.cn/？

5 补充问题

补充说明几个问题。

5.1 一般思路

关于具体所解决的问题，值得大家注意的是：

（1）明白数据集各属性的含义，从实际出发，理解数据集合的提供者希望去进行什么样的分析。

（2）大多数数据集是不能直接用 Weka 打开，需要在打开文件，对数据集合进行一定的预处理。

（3）毫无疑问，训练集的质量越高，则学习效果就越好；因为我们是依据训练集来进行归纳与监督学习的。一般地，训练集合越大，则学习效果就越好。

5.2 特殊处理

如果你觉得实在不胜任去解决涉及文本的问题，那么你就不得不考虑以下解决方案：

（1）第一个解决方案：一定要看清楚文本分类的过程，相关概念以及文本的向量表示；还需要在网上查找分词、去停用词、以及生成频率的软件工具。

（2）第二种方案是一种**退而求其次**的方案，可以换一个题目。但是题目不能与 29 个题目重复，最好在常用的数据网站上找，无论找什么数据集合都要写一段简短的描述文字，并且给出具体下载的地址，不能只是网站。（

（3）第三种解决方案是**底线方案**：如果前两种解决方案你认为你都不能完成任务的话，那么底线是你在 uci 网站上找一个有一万条以上的数据集，进行分

析。这种情况下，你就不要在第三节中介绍 uci 上的这个数据集，从第四节开始描述你在这个数据集之上的所有工作。

5.3 关于表格的要求

关于图示的要求及注意事项：

（1）从视觉效果看，图示上下各空一行。

（2）图示必须有编号和标题，并且居中排版与图的下方。

（3）图示全文统一编号，或按章节编号。

（4）从视觉效果看，图示中的主体文字必须小于接近于 5 号字体.

（5）表格或图示与正文段落没有上下左右的位置关系。

（6）表序号或图示序号与其相应的标题之间空一个汉字距离，即两个半角距离。

（7）建议采用画布来插入图形，必须确保图示与其标题在同一页面上。

（8）一般的，图与表的篇幅不得超过正文的一般，应该有 60%以上是文字。我们可以利用调整文字的先后次序来在页面上布置图示。

5.4 其他问题

意思是论文第三节之后的内容一个组可以一样，也可以不一样，关键看质量（计科 1907 舒一凡 19851022 有解释能力）。

幻灯片写什么？答：依据自己的课程设计报告，重点突出自己的工作。