数据挖掘
课程设计指导书

计算机科学与工程学院

内部使用

湖南科技大学

2023 年 12 月

一、课程设计目的

通过 Python/Java/VC 或自己熟悉的语言编程实现知识发现、关联规则、分类聚类领域一些重要的算法，让学生掌握 C4.5、k-mean、Apriori、EM、PageRank、DBSCAN 等算法基本原理，能够运用数据挖掘方法解决具体问题。

二、课程设计要求

1. 本课程设计有十个备选实验，请自选五个，每个实验 20 分，共 100 分。

2. 遵守课程设计在线教学的规章制度，按时签到，不得在课程设计期间做与课程设计无关的事情。对于课程设计过程中出现的问题，及时向指导老师汇报，并接受指导老师的定期检查。

3. 课程设计报告要求格式规范，语言通顺，每个算法都要写明使用的数据结构，画出算法的程序流程图、关键的源程序代码、仔细记录实验结果。同时，将实验中遇到的问题和疑惑及思考解答过程，作为重点内容写入实验报告。

4. 及时提交课程设计报告及源程序文件。每个算法的报告篇幅建议大约 1-2 页，总课程设计报告页面数大约 10 页左右。实验报告格式请参见附录 1，其中封面单面打印，其他内容双面打印。

实验一 Apriori 算法设计与应用

1、背景介绍

Apriori 算法是一种挖掘关联规则的频繁项集算法，其核心思想是通过候选集生成和向下封闭检测两个阶段来挖掘频繁项集。

2、实验内容

有如下的交易记录，请编写程序实现 Apriori 算法，挖掘该交易记录里的频繁项集，并挖掘出所有的强关联规则，设最小支持度为 50%，最小置信度为 50%。

交易记录	所购买的商品	交易记录	所购买的商品
1	A、B、C、D、E、F、G、	8	A、B、C、E、G、H、
2	A、B、C、D、E、H、	9	A、B、C、D、E、F、H、
3	A、B、C、D、E、F、G、H、	10	C、D、E、F、G、H、
4	A、B、C、G、H、	11	A、B、C、D、G、H、
5	A、B、C、D、G、H、	12	A、C、D、E、F、G、H、
6	A、B、C、D、E、F、G、H、	13	A、B、C、E、F、G、H、
7	A、B、C、D、E、F、G、	14	B、C、E、F、G、H、

3、实验要求

画出程序流程图、给出主要代码，并给出实验结果。

实验二 Close 算法设计与应用

1、背景介绍

一个频繁闭合项目集的所有闭合子集一定是频繁的；一个非频繁闭合项目集的所有闭合超集一定是非频繁的。

2、实验内容

有如下的交易记录，请编写程序实现 Close 算法，挖掘该交易记录里的频繁项集及强关联规则。设最小支持度为 60%，最小置信度为 60%。

序号	商品
1	短裤、帽子、长裤、裙子、棉衣、短袖、衬衫、袜子
2	帽子、长裤、裙子、棉衣、短袖、衬衫、袜子
3	短裤、帽子、裙子、棉衣、短袖、衬衫、袜子
4	短裤、帽子、棉衣、短袖、衬衫、袜子
5	短裤、帽子、长裤、裙子、棉衣、短袖、袜子
6	短裤、帽子、长裤、裙子、棉衣、短袖、衬衫、袜子
7	短裤、帽子、长裤、裙子、棉衣、短袖、
8	帽子、长裤、裙子、棉衣、衬衫、袜子
9	短裤、帽子、长裤、裙子、棉衣、短袖、衬衫、袜子
10	短裤、帽子、长裤、裙子、短袖、衬衫、袜子
11	短裤、帽子、长裤、裙子、棉衣、短袖、衬衫、袜子
12	短裤、帽子、长裤、棉衣、短袖、衬衫、袜子
13	短裤、长裤、裙子、棉衣、短袖、衬衫、袜子
14	帽子、长裤、裙子、棉衣、短袖、衬衫、袜子
15	短裤、帽子、长裤、裙子、棉衣、短袖、衬衫、袜子

3、实验要求

画出程序流程图、给出主要代码，并正确给出实验结果。

实验三 FP-tree 算法设计与应用

1、背景介绍

Frequent Pattern Tree：不使用侯选集，直接压缩数据库成一个频繁模式树，通过频繁模式树可以直接得到频集。进行 2 次数据库扫描：一次对所有 1-项目的频度排序；一次将数据库信息转变成紧缩内存结构。

2、实验内容

有如下的交易记录，请编写程序实现 FP-tree 算法，挖掘该交易记录里的频繁项集及强关联规则。设最小置信度为 60%，最小支持度为 60%。

交易记录	所购买的商品
1	A、B、C、D、E、F、H、
2	C、D、E、F、H、
3	A、D、E、F、H、
4	A、B、C、D、F、H、
5	A、B、C、F、H、
6	A、B、C、D、E、、
7	A、B、C、D、E、F
8	A、B、C、D、E、F、H、
9	A、B、C、D、E、
10	A、C、D、E、F、H、
11	A、B、C、D、
12	A、B、C、D、E、F、H、
13	A、D、E、F、H、
14	A、B、C、D、E、F、H、
15	A、B、C、D、F、H、
16	A、B、C、D、E、F、H、
17	A、B、D、E、F、H、
18	A、B、C、D、E、F、H、
19	A、B、E、F、H、
20	A、B、C、D、E、F、H、

3、实验要求

画出程序流程图、给出主要代码，并正确给出实验结果。

实验四 EM 算法设计与应用

1、背景介绍

编程实现最大期望算法：在概率模型中寻找参数最大似然预计或者最大后验预计的算法。用于寻找，依赖于不可观察的隐性变量的概率模型中，参数的最大似然预计。最大期望算法经过两个步骤交替进行计算，第一步是计算期望（E），利用对隐藏变量的现有预计值，计算其最大似然预计值；第二步是最大化（M）。最大化在 E 步上求得的最大似然值来计算参数的值。M 步上找到的参数预计值被用于下一个 E 步计算中，这个过程不断交替进行。

2、实验内容

这是一个抛硬币的例子，H 表示正面向上，T 表示反面向上，，硬币有两个，A 和 B，硬币是有偏的。本次实验总共做了 5 组，每组随机选一个硬币，连续抛 10 次。请用 EM 算法，推测 A、B 硬币正面朝上的概率。

实验	结果	结果	结果	结果	结果	结果	结果	结果	结果	结果
1	正面	反面	反面	反面	正面	正面	正面	反面	反面	正面
2	正面	正面	正面	正面	反面	正面	正面	正面	正面	正面
3	正面	反面	正面	正面	正面	正面	正面	正面	正面	正面
4	正面	反面	正面	反面	反面	反面	反面	正面	反面	反面
5	反面	正面	正面	正面	正面	正面	正面	正面	正面	正面

3、实验要求

画出程序流程图、给出主要代码，并正确给出实验结果。

实验五 KNN 算法设计与应用

1、背景介绍

k-近邻（kNN, k-NearestNeighbor）是在训练集中选取离输入的数据点最近的 k 个邻居，根据这个 k 个邻居中出现次数最多的类别（最大表决规则），作为该数据点的类别。。

2、实验内容

某班有 14 个同学，已登记身高及等级，新同学易昌，身高 1.74cm，等级是什么。请用 knn 算法进行分类识别，其中 k=5。

序号	姓名	身高（cm）	等级
1	李丽	1.5	矮
2	吉米	1.92	高
3	马大华	1.7	中等
4	王晓华	1.73	中等
5	刘敏	1.6	矮
6	张强	1.75	中等
7	李秦	1.6	矮
8	王壮	1.9	高
9	刘冰	1.68	中等
10	张喆	1.78	中等
11	杨毅	1.70	中等
12	徐田	1.68	中等
13	高杰	1.65	矮
14	张晓	1.78	中等

3、实验要求

（1）画出程序流程图、给出主要代码，并正确给出实验结果。

（2）思考：k 取不同值，对结果有什么影响？

实验六 ID3 算法设计与实现

1、背景介绍

信息增益是针对一个一个特征（属性）而言的，就是看一个特征，系统有它和没有它时的信息量各是多少，两者的差值就是这个特征给系统带来的信息量，即信息增益。ID3 算法：Iterative Dichotomiser 3，迭代二叉树 3 代，通过计算每个属性的信息增益，认为信息增益高的是好属性，每次划分选取信息增益最高的属性为划分标准，重复这个过程，直至生成一个能完美分类训练样例的决策树。

2、实验内容

下表中共有 1024 条用户信息，分别表示不同年龄、收入、身份（是否为学生）、信誉的用户是否会购买电脑。请用 ID3 算法，构造相应的决策树。

计数	年龄	收入	学生	信誉	是否买电脑
64	青	高	否	良	不买
64	青	高	否	优	不买
128	中	高	否	良	买
60	老	中	否	良	买
64	老	低	是	良	买
64	老	低	是	优	不买
64	中	低	是	优	买
128	青	中	否	良	不买
64	青	低	是	良	买
132	老	中	是	良	买
64	青	中	是	优	买
32	中	中	否	优	买
32	中	高	是	良	买
63	老	中	否	优	不买
1	老	中	否	优	买

3、实验要求

画出程序流程图、给出主要代码，并正确给出实验结果。

实验七 DBSCAN 算法设计与应用

1、背景介绍

DBSCAN 算法：如果一个点 q 的区域内包含多于 MinPts 个对象，则创建一个 q 作为核心对象的簇。然后，反复地寻找从这些核心对象直接密度可达的对象，把一些密度可达簇进行合并。当没有新的点可以被添加到任何簇时，该过程结束。

2、实验内容

有如下二维数据集，取 ε =2，minpts=3，请使用 DBSCAN 算法对其聚类（使用曼哈顿距离）

序号	横坐标	纵坐标	序号	横坐标	纵坐标	序号	横坐标	纵坐标
1	20	45	22	20	45	43	6	36
2	34	23	23	34	23	44	78	37
3	53	67	24	53	67	45	7	38
4	54	85	25	54	85	46	70	39
5	67	4	26	67	4	47	6	36
6	33	67	27	33	67	48	45	45
7	24	78	28	24	78	49	67	67
8	37	90	29	37	90	50	84	6
9	67	34	30	67	34	51	23	78
10	34	56	31	78	32	52	13	7
11	89	78	32	23	33	53	45	70
12	65	23	33	45	34	54	67	76
13	45	45	34	67	35	55	4	54
14	67	67	35	67	76	56	68	60
15	84	6	36	4	54	57	45	45
16	23	78	37	68	60	58	34	67
17	13	7	38	7	38	59	35	67
18	45	70	39	70	39	60	36	4
19	67	76	40	76	40	61	37	68
20	4	54	41	45	70	62	38	45
21	68	60	42	67	76	63	20	34

3、实验要求

画出程序流程图、给出主要代码，并正确给出实验结果。

实验八 K-mean 算法设计与应用

1、背景介绍

k-means 算法，也被称为 k-平均或 k-均值，是一种得到最广泛使用的聚类算法。相似度的计算根据一个簇中对象的平均值来进行。算法首先随机地选择 k 个对象，每个对象初始地代表了一个簇的平均值或中心。对剩余的每个对象根据其与各个簇中心的距离，将它赋给最近的簇。然后重新计算每个簇的平均值。这个过程不断重复，直到准则函数收敛。

2、实验内容

请对下表中的数据进行 k-mean 聚类，距离为欧氏距离，k=3.

P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P13
1	2	2	4	5	6	6	7	9	1	3	5	3
2	1	4	3	8	7	9	9	5	12	12	12	3

3、实验要求

画出程序流程图、给出主要代码，并正确给出实验结果。

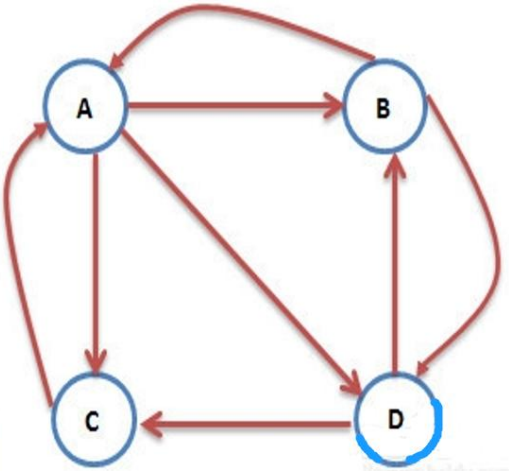
实验九 PageRank 算法设计与应用

1、背景介绍

PageRank 算法：计算每一个网页的 PageRank 值，然后根据这个值的大小对网页的重要性进行排序。

2、实验内容

互联网中的网页的链接可以看作是有向图，如 A、B、C、D 四个网页，请采用 PageRank 算法对网页进行分级排序。



3、实验要求

画出程序流程图、给出主要代码，并正确给出实验结果。

实验十 模型评估指标

1、背景介绍

当模型训练完成后，需要进行模型评估来量化模型性能的好坏。而且模型的性能不单单只有一个维度，所以模型的好坏通常会用多个指标来进行衡量。例如，现在想要衡量一个分类器的性能，可能第一时间会想到用准确率来衡量模型的好坏，但是准确率高并不一定就代表模型的性能高，因此可能会需要使用如 F1 Score、AUC 等指标来衡量。

2、实验内容

请编写程序，实现混淆矩阵、准确率、精确率、ROC 曲线、AUC 面积、F1 Score 的计算，对下列数据进行分析。

（1）某分类器对 66 只动物进行分类，其中 13 只猫，53 只不是猫，分类器判断时这 13 只猫只有 10 只预测对了，其他动物也只预测对了 45 只。

（2）某分类器对图片进行分类，共有 100 张图片。其中 36 张汽车，预测对了 30 张，另外 4 张预测成了船，2 张预测成了房子。40 张船，预测对了 34 张，另外 6 张预测成了汽车。24 张房子，预测对了 24 张。

3、实验要求

画出程序流程图、给出主要代码，并正确给出实验结果。

附件　1：课程设计报告格式

湖南科技大学计算机科学与工程学院

_____课程设计报告

专业班级：_____

姓　　名：_____

学　　号：_____

指导教师：_____

时　　间：_____

地　　点：_____

指导教师评语：	
成绩：　　等级：	签名：_____ 年　月　日

实验报告内容如下：

一、实验题目

二、背景介绍

三、实验内容（包括：实验原理/运用的理论知识、算法/程序流程图、步骤和方法、关键源代码）

四、实验结果与分析

五、小结与心得体会

实验报告排版要求：

1、标题：四号宋体，加粗
2、正文：小四号宋体
3、关键源代码：五号字体
4、流程图：五号字体，并在图下标注图名
5、表格：五号字体，采用三线表格式，并在表上标注表名

附件　2：计算机科学与工程学院课程设计成绩单

计算机科学与工程学院课程设计成绩单

课程设计名称:　　教师:　　上课班级:

开课学期: 2023-2024-1　　考核方式：考查

学号	姓名	平时成绩	测试成绩	创新成绩	报告成绩	总成绩

注：1.此表格前两栏学生信息可从教务系统中导出。

2.平时成绩（0—20 分）、测试成绩（0-40 分）、创新成绩（0-10 分）以及报告成绩（30 分）这四项求和为总成绩（0—100 分）。

3. 总成绩按“≥90 分为优秀、80 分~89 分为良好、70 分~79 分为中等、60~69 分为及格、60 分以下为不及格”的标准转换为等级制。

4. 在课程设计报告封面上附上百分制成绩和等级制成绩。