# 数据挖掘应用技术与实践实验指导书

随着数据库技术的发展，形成了数据丰富，知识缺乏的严重局面。针对如何有效利用这些海量的数据信息的挑战，数据挖掘技术应运而生，并显示出强大的生命力。数据挖掘技术是数据处理技术进入了一个更高级的阶段，是对未来人类产生重大影响的十大新兴技术之一。因此加强数据挖掘领域的理论与实践学习也已成为专业学生的必修内容。

本实验指导书通过大量的实例，循序渐进地引导学生做好各章的实验。根据实验教学大纲，我们编排了4个实验。在实验中学生根据实验指导中的内容进行验证总结。实验完成后，学生按照要求完成实验报告。整个实验中，我们强调培养学生动手实践能力，掌握数据挖掘的基本方法。

# 实验一 ID3算法实现

1. 实验目的

通过编程实现决策树算法，信息增益的计算、数据子集的划分、决策树的构建过程。加深对相关算法的理解过程。

实验类型：验证

计划学时：4学时

1. 实验内容
2. 分析决策树算法的实现流程
3. 分析信息增益的计算、数据子集的划分、决策树的构建过程
4. 根据算法描述编程实现算法，调试运行；
5. 实验方法

根据ID算法对数据进行分析。

1. 思考题
2. 信息增益的计算
3. 选择相关字段后根据相关字段的取值对数据集进行划分
4. 决策树构建的终止条件

## 实验二 贝叶斯算法

1. 实验目的

通过贝叶斯算法的编程实现，加深对贝叶斯算法理解，同时利用贝叶斯算法对简单应用实现预测分类。

实验类型：验证

计划学时：4学时

1. 实验内容
2. 分析贝叶斯算法；
3. 计算条件概率；
4. 预测精度的计算与评估；
5. 编程实现贝叶斯分类算法，并对简单应用样本数据进行预测分类。
6. 实验方法
7. 实现贝叶斯算法；

算法流程如下：

1. 输入实验数据
2. 设定训练集数据和测试集数据
3. 计算训练集数据中总各属性在各类中的概率分布情况
4. 利用测试数据计算机贝叶斯分了算法的精度
5. 输出分类结构
6. 利用实验数据对贝叶斯算法进行检测；
7. 求精确度计算；
8. 完成整个分类与评估的过程
9. 思考
10. 如果各属性不是条件独立的，则如何进行分类。

## 实验三 先验算法

1. 实验目的
2. 掌握Apriori算法在关联规则挖掘中频繁项集的产生以及关联规则集合的产生过程中的作用；
3. 根据算法描述编程实现算法，调试运行。并结合相关实验数据进行应用，分析实验结果。

实验类型：验证

计划学时：4学时

1. 实验内容
2. 频繁项集的生产与Apriori算法实现；
3. 关联规则的生成过程算法实现；
4. 结合实验数据对算法进行分析；
5. 实验过程

编程完成以下算法：

1、频繁项集的产生

2、规则的产生

1. 思考题
2. 集合的表示以及相关操作的实现方法。
3. 项集的数据结构描述。

## 实验四 K-means算法

1. 实验目的

通过分析K-means聚类算法的聚类原理，利用编程工具编程实现K-means聚类算法，并通过对样本数据的聚类过程，加深对该聚类算法的理解与应用过程。

实验类型：验证

计划学时：4学时

1. 实验内容
2. 分析K-means聚类算法；
3. 分析距离计算方法；
4. 分析聚类的评价标准；
5. 编程完成K-means聚类算法，并 基于相关实验数据实现聚类过程；
6. 实验方法

K-means聚类算法原理：K-means聚类算法以k为参数，把n个对象分为k个簇，使簇内的对象具有较高的相似度。相似度的计算根据簇中对象的平均值来进行计算。

算法描述：

输入：簇的书面k和包含n个对象的数据库

输出：是平方误差准则最小的k个簇

算法流程：

1. 给定需要生成的簇的数目k
2. 随机选择k个对象作为初始的簇中心
3. 余下的对象，根据其与各个簇中心的距离，将他们赋给最近的簇
4. 重新计算每个簇的平均值
5. 以所得到的簇的平均值，返回第3步重新分配对象，直到准则函数收敛。

准则函数E：常常采用平方误准则来度量。

1. 思考题
2. 考虑选择不同的距离函数，对结果的影响。
3. 评价函数的计算。

## 考核与成绩评定

1、考核内容：考勤情况、完成情况、实验报告书撰写情况

2、成绩评定方法：学生做完实验后，教师依据学生考勤情况、完成情况、实验报告书撰写情况，给出优、良、中、及格、不及格的等级成绩评定，并转换为实验成绩的一部分。

实验报告格式

西南科技大学

计算机实验报告

课程名称：

实验名称：

学 号：

学生姓名：

班 级：

指导教师：

评 分：

实验日期：

实验目的：

实验内容：

实验要求：

实验步骤：

实验思考：

实验体会：