### 实验一 数据预处理

**实验目的：**

1. 掌握数据挖掘前期的数据预处理方法；
2. 通过自行编程实现数据预处理。

**实验内容**：

**1.数据平滑**

假定用于分析的数据包含属性age。数据元组中age的值如下（按递增序）：13, 15, 16, 16, 19, 20, 20, 21, 22, 22, 25, 25, 25, 25, 30, 33, 33, 35, 35, 35, 35, 36, 40, 45, 46, 52, 70。使用你所熟悉的程序设计语言进行编程，实现如下功能（要求程序具有通用性）：

(a) 使用按箱平均值平滑法对以上数据进行平滑，箱的深度为3。

(b) 使用按箱中值平滑法对以上数据进行平滑，箱的深度为3。

(c) 使用按箱边界值平滑法对以上数据进行平滑，箱的深度为3。

**2.离群点筛选**

数据同上题，使用你所熟悉的程序设计语言进行编程，找出其中的离群点（要求程序具有通用性）。

**实验步骤**：

1．编程实现数据平滑（要求对任意数据组均使用）。

2．编程实现离群点筛选（找出离群点，即1.5倍IQR之外的点）。

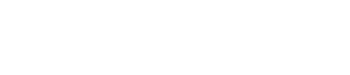
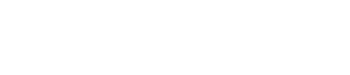
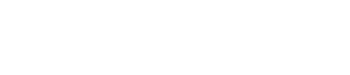
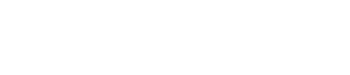
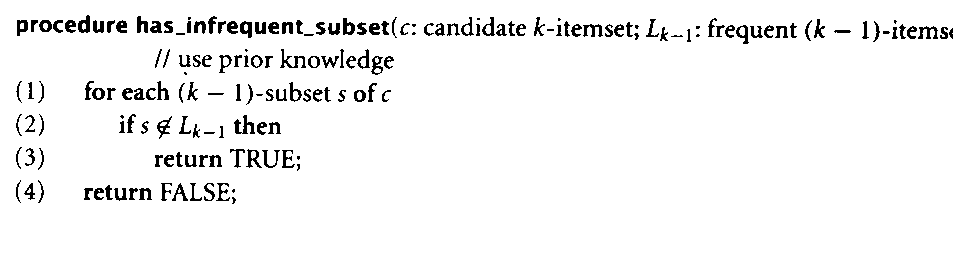
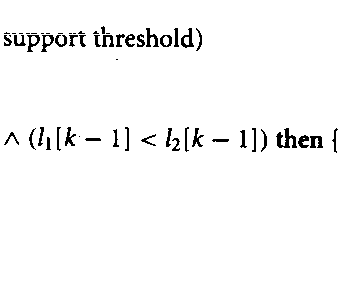
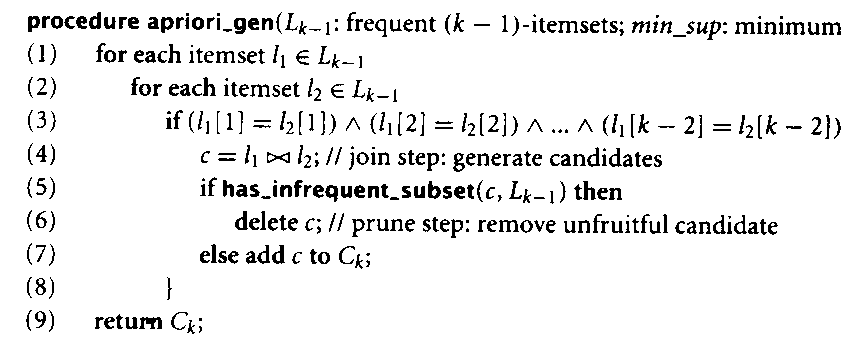
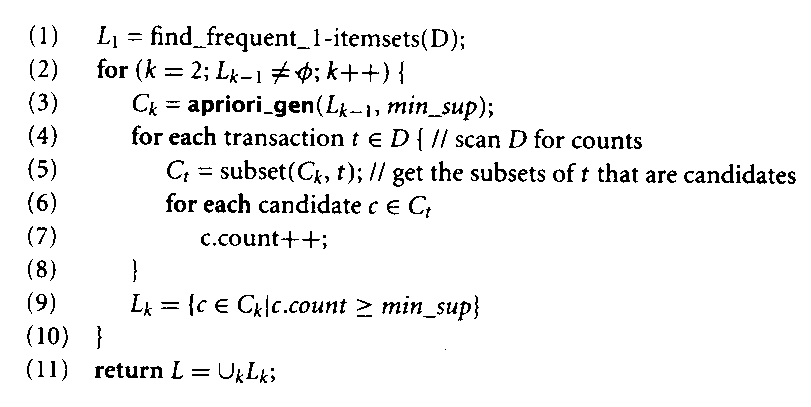
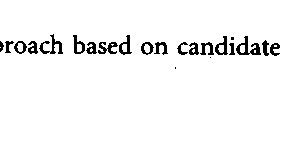
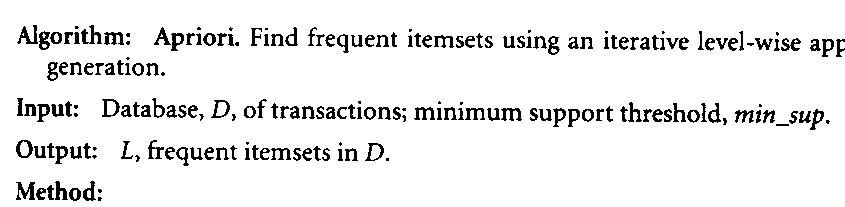
### 实验二 关联规则挖掘

**实验目的：**

* 1. 掌握关联规则挖掘的Apriori算法；
  2. 将Apriori算法用具体的编程语言实现。

**实验内容**：

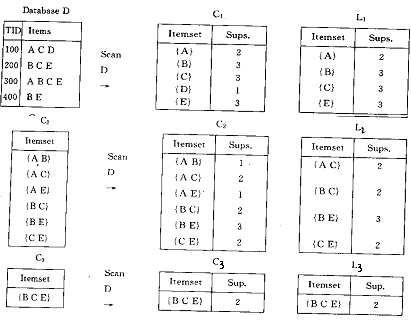
根据下列的Apriori算法进行编程：



**实验步骤**：

1.编制程序。

2.调试程序。可采用下面的数据库D作为原始数据调试程序，得到的候选1项集、2项集、3项集分别为C1、C2、C3，得到的频繁1项集、2项集、3项集分别为L1、L2、L3。



### 实验三 聚类分析

**实验目的：**

1. 掌握k-means聚类方法；
2. 通过自行编程，对三维空间内的点用k-means方法聚类。

**实验内容**：

1.问题描述：

立体空间三维点的聚类.

说明：数据放在数据文件中(不得放在程序中)，第一行是数据的个数，以后各行是各个点的x,y,z坐标。

2.设计要求

读取文本文件数据，并用K-means方法输出聚类中心

3. 需求分析

k-means 算法接受输入量k；然后将n个数据对象划分为 k个聚类以便使得所获得的聚类满足：同一聚类中的对象相似度较高；而不同聚类中的对象相似度较小。聚类相似度是利用各聚类中对象的均值所获得一个“中心对象”（引力中心）来进行计算的。

k-means算法的工作过程说明如下：首先从n个数据对象任意选择k个对象作为初始聚类中心，而对于所剩下的其它对象，则根据它们与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类。然后，再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值），不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数，具体定义如下：

 (1)

其中*E*为数据库中所有对象的均方差之和，*p*为代表对象的空间中的一个点，*mi*为聚类*Ci*的均值(***p***和***m****i*均是多维的)。公式(1)所示的聚类标准，旨在使所获得的*k*个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

**实验步骤**：

Step 1.读取数据组，从N个数据对象任意选择k个对象作为初始聚类中心；

Step 2.循环Step 3到Step 4直到每个聚类不再发生变化为止；

Step 3.根据每个聚类对象的均值（中心对象），计算每个对象与这些中心对象的距离，并根据最小距离重新对相应对象进行划分；

Step 4.重新计算每个（有变化）聚类的均值（中心对象）。