# 数据仓库与数据挖掘复习题完整版

数据仓库与数据挖掘复

习题

Document serial number【NL89WT-NY98YT-NC8CB-NNUUT-NUT108】

2014-2015-1《数据仓库与数据挖掘》

期末考试题型

一、单项选择题（每小题2分，共20分）

二、填空题（每空1分，共20分）

三、简答题(每题6分，共30分)

四、析题与计算题(共30分)

请同学们在考试时不要将复习资料带入考场！！！

单选题

1. 某超市研究销售纪录数据后发现，买啤酒的人很大概率也会购买尿布，这种属于数据挖掘的哪类问题？(A)

A. 关联规则发现

B. 聚类

C. 分类

D. 自然语言处理

2. 以下两种描述分别对应哪两种对分类算法的评价标准？ (A)

(a)警察抓小偷，描述警察抓的人中有多少个是小偷的标准。

(b)描述有多少比例的小偷给警察抓了的标准。

A. Precision, Recall

B. Recall, Precision

A. Precision, ROC D. Recall, ROC

3. 将原始数据进行集成、变换、维度规约、数值规约是在以下哪个步骤的任务？(C)

A. 频繁模式挖掘

B. 分类和预测

C. 数据预处理

D. 数据流挖掘

4. 当不知道数据所带标签时，可以使用哪种技术促使带同类标签的数据与带其他标签的数据相分离？(B)

A. 分类

B. 聚类

C. 关联分析

D. 隐马尔可夫链

5. 什么是KDD (A)

A. 数据挖掘与知识发现

B. 领域知识发现

C. 文档知识发现

D. 动态知识发现

6. 使用交互式的和可视化的技术，对数据进行探索属于数据挖掘的哪一类任务（

A）

A. 探索性数据分析

B. 建模描述

C. 预测建模

D. 寻找模式和规则

7. 为数据的总体分布建模；把多维空间划分成组等问题属于数据挖掘的哪一类任务？(B)

A. 探索性数据分析

B. 建模描述

C. 预测建模

D. 寻找模式和规则

8. 建立一个模型，通过这个模型根据已知的变量值来预测其他某个变量值属于数据挖掘的哪一类任务？(C)

A. 根据内容检索

B. 建模描述

C. 预测建模

D. 寻找模式和规则

9. 用户有一种感兴趣的模式并且希望在数据集中找到相似的模式，属于数据挖掘哪一类任务？(A)

A. 根据内容检索

B. 建模描述

C. 预测建模

D. 寻找模式和规则

11.下面哪种不属于数据预处理的方法？ (D)

A变量代换 B离散化 C 聚集 D 估计遗漏值

12. 假设12个销售价格记录组已经排序如下：5, 10, 11, 13, 15, 35, 50, 55, 72, 92, 204, 215 使用如下每种方法将它们划分成四个箱。等频（等深）划分时，15在第几个箱子内？ (B)

A 第一个

B 第二个

C 第三个

D 第四个

13.上题中，等宽划分时（宽度为50），15又在哪个箱子里？ (A)

A 第一个

B 第二个

C 第三个

D 第四个

14.下面哪个不属于数据的属性类型：(D)

A 标称

B 序数

C 区间 D相异

15. 在上题中，属于定量的属性类型是：(C)

A 标称

B 序数 C区间 D 相异

16. 只有非零值才重要的二元属性被称作：( C )

A 计数属性

B 离散属性 C非对称的二元属性 D 对称属性

17. 以下哪种方法不属于特征选择的标准方法： (D)

A嵌入 B 过滤 C 包装 D 抽样

18.下面不属于创建新属性的相关方法的是： (B)

A特征提取 B特征修改 C映射数据到新的空间 D特征构造

19. 考虑值集{1、2、3、4、5、90}，其截断均值（p=20%）是 (C)

A 2

B 3

C 3.5

D 5

20. 下面哪个属于映射数据到新的空间的方法？ (A)

A 傅立叶变换 B特征加权 C 渐进抽样 D维归约

21. 熵是为消除不确定性所需要获得的信息量，投掷均匀正六面体骰子的熵是： (B)

A 1比特

B 2.6比特

C 3.2比特

D 3.8比特

22. 假设属性income的最大最小值分别是12000元和98000元。利用最大最小规范化的方法将属性的值映射到0至1的范围内。对属性income的73600元将被转化为：(D)

A 0.821

B 1.224

C 1.458

D 0.716

23.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）：13，15，16，16，19，20，20，21，22，22，25，25，25，30，33，33，35，35，36，40，45，46，52，70, 问题：使用按箱平均值平滑方法对上述数据进行平滑，箱的深度为3。第二个箱子值为：(A)

A 18.3

B 22.6

C 26.8

D 27.9

24. 考虑值集{12 24 33 2 4 55 68 26}，其四分位数极差是：(A)

A 31

B 24

C 55

D 3

25. 一所大学内的各年纪人数分别为：一年级200人，二年级160人，三年级130人，四年级110人。则年级属性的众数是： (A)

A 一年级 B二年级 C 三年级 D 四年级

26. 下列哪个不是专门用于可视化时间空间数据的技术： (B)

A 等高线图 B饼图 C 曲面图 D 矢量场图

27. 在抽样方法中，当合适的样本容量很难确定时，可以使用的抽样方法是：

(D)

A 有放回的简单随机抽样 B无放回的简单随机抽样 C分层抽样 D 渐进抽样

28. 数据仓库是随着时间变化的,下面的描述不正确的是 (C)

A. 数据仓库随时间的变化不断增加新的数据内容;

B. 捕捉到的新数据会覆盖原来的快照;

C. 数据仓库随事件变化不断删去旧的数据内容;

D. 数据仓库中包含大量的综合数据,这些综合数据会随着时间的变化不断地进

行重新综合.

29. 关于基本数据的元数据是指: (D)

A. 基本元数据与数据源,数据仓库,数据集市和应用程序等结构相关的信息;

B. 基本元数据包括与企业相关的管理方面的数据和信息;

C. 基本元数据包括日志文件和简历执行处理的时序调度信息;

D. 基本元数据包括关于装载和更新处理,分析处理以及管理方面的信息.

30. 下面关于数据粒度的描述不正确的是: (C)

A. 粒度是指数据仓库小数据单元的详细程度和级别;

B. 数据越详细,粒度就越小,级别也就越高;

C. 数据综合度越高,粒度也就越大,级别也就越高;

D. 粒度的具体划分将直接影响数据仓库中的数据量以及查询质量.

31. 有关数据仓库的开发特点,不正确的描述是: (A)

A. 数据仓库开发要从数据出发;

B. 数据仓库使用的需求在开发出去就要明确;

C. 数据仓库的开发是一个不断循环的过程,是启发式的开发;

D. 在数据仓库环境中,并不存在操作型环境中所固定的和较确切的处理流,数据仓库中数据分析和处理更灵活,且没有固定的模式

32. 在有关数据仓库测试,下列说法不正确的是: (D)

A. 在完成数据仓库的实施过程中,需要对数据仓库进行各种测试.测试工作中要包括单元测试和系统测试.

B. 当数据仓库的每个单独组件完成后,就需要对他们进行单元测试.

C. 系统的集成测试需要对数据仓库的所有组件进行大量的功能测试和回归测试.

D. 在测试之前没必要制定详细的测试计划.

33. OLAP技术的核心是: (D)

A. 在线性;

B. 对用户的快速响应;

C. 互操作性.

D. 多维分析;

34. 关于OLAP的特性,下面正确的是: (D)

(1)快速性 (2)可分析性 (3)多维性 (4)信息性 (5)共享性

A. (1) (2) (3)

B. (2) (3) (4)

C. (1) (2) (3) (4)

D. (1) (2) (3) (4) (5)

35. 关于OLAP和OLTP的区别描述,不正确的是: (C)

A. OLAP主要是关于如何理解聚集的大量不同的数据.它与OTAP应用程序不同.

B. 与OLAP应用程序不同,OLTP应用程序包含大量相对简单的事务.

C. OLAP的特点在于事务量大,但事务内容比较简单且重复率高.

D. OLAP是以数据仓库为基础的,但其最终数据来源与OLTP一样均来自底层的数据库系统,两者面对的用户是相同的.

36. OLAM技术一般简称为”数据联机分析挖掘”,下面说法正确的是: (D)

A. OLAP和OLAM都基于客户机/服务器模式,只有后者有与用户的交互性;

B. 由于OLAM的立方体和用于OLAP的立方体有本质的区别.

C. 基于WEB的OLAM是WEB技术与OLAM技术的结合.

D. OLAM服务器通过用户图形借口接收用户的分析指令,在元数据的知道下,对超级立方体作一定的操作.

37. 关于OLAP和OLTP的说法,下列不正确的是: (A)

A. OLAP事务量大,但事务内容比较简单且重复率高.

B. OLAP的最终数据来源与OLTP不一样.

C. OLTP面对的是决策人员和高层管理人员.

D. OLTP以应用为核心,是应用驱动的.

38. 设X={1，2，3}是频繁项集，则可由X产生\_\_(C)\_\_个关联规则。

A、4

B、5

C、6

D、7

40. 概念分层图是\_\_(B)\_\_图。

A、无向无环

B、有向无环

C、有向有环

D、无向有环

41. 频繁项集、频繁闭项集、最大频繁项集之间的关系是： (C)

A、频繁项集频繁闭项集 =最大频繁项集

B、频繁项集 = 频繁闭项集最大频繁项集

C、频繁项集频繁闭项集最大频繁项集

D、频繁项集 = 频繁闭项集 = 最大频繁项集

42. 考虑下面的频繁3-项集的集合：{1，2，3}，{1，2，4}，{1，2，5}，{1，3，4}，{1，3，5}，{2，3，4}，{2，3，5}，{3，4，5}假定数据集中只有5个项，采用合并策略，由候选产生过程得到4-项集不包含（C）

A、1，2，3，4

B、1，2，3，5

C、1，2，4，5

D、1，3，4，5

43.下面选项中t不是s的子序列的是 ( C )

A、s=<{2,4},{3,5,6},{8}> t=<{2},{3,6},{8}>

B、s=<{2,4},{3,5,6},{8}> t=<{2},{8}>

C、s=<{1,2},{3,4}> t=<{1},{2}>

D、s=<{2,4},{2,4}> t=<{2},{4}>

44. 在图集合中发现一组公共子结构，这样的任务称为 ( B )

A、频繁子集挖掘

B、频繁子图挖掘

C、频繁数据项挖掘

D、频繁模式挖掘

45. 下列度量不具有反演性的是 (D)

A、系数

B、几率

C、Cohen度量

D、兴趣因子

46. 下列\_\_(A)\_\_不是将主观信息加入到模式发现任务中的方法。

A、与同一时期其他数据对比

B、可视化

C、基于模板的方法

D、主观兴趣度量

47. 下面购物篮能够提取的3-项集的最大数量是多少（C）

ID 购买项

1 牛奶，啤酒，尿布

2 面包，黄油，牛奶

3 牛奶，尿布，饼干

4 面包，黄油，饼干

5 啤酒，饼干，尿布

6 牛奶，尿布，面包，黄油

7 面包，黄油，尿布

8 啤酒，尿布

9 牛奶，尿布，面包，黄油

10 啤酒，饼干

A、1

B、2

C、3

D、4

48. 以下哪些算法是分类算法，A，DBSCAN B，C4.5 C,K-Mean D,EM （B）

49. 以下哪些分类方法可以较好地避免样本的不平衡问题， A，KNN B，SVM C，Bayes D，神经网络（A）

50. 决策树中不包含一下哪种结点，A,根结点（root node) B,内部结点（internal node） C,外部结点（external node） D,叶结点（leaf node）

(C)

51. 不纯性度量中Gini计算公式为（其中c是类的个数） (A)

A, B, C, D, （A）

53. 以下哪项关于决策树的说法是错误的 (C)

A. 冗余属性不会对决策树的准确率造成不利的影响

B. 子树可能在决策树中重复多次

C. 决策树算法对于噪声的干扰非常敏感

D. 寻找最佳决策树是NP完全问题

54. 在基于规则分类器的中，依据规则质量的某种度量对规则排序，保证每一个测试记录都是由覆盖它的“最好的”规格来分类，这种方案称为 (B)

A. 基于类的排序方案

B. 基于规则的排序方案

C. 基于度量的排序方案

D. 基于规格的排序方案。

55. 以下哪些算法是基于规则的分类器 (A)

A. C4.5

B. KNN

C. Nave Bayes

D. ANN

56. 以下关于人工神经网络（ANN）的描述错误的有 (A)

A，神经网络对训练数据中的噪声非常鲁棒 B，可以处理冗余特征 C，训练ANN是一个很耗时的过程 D，至少含有一个隐藏层的多层神经网络

57. 通过聚集多个分类器的预测来提高分类准确率的技术称为 (A)

A,组合(ensemble) B,聚集(aggregate) C，合并(combination) D，投票(voting)

58. 简单地将数据对象集划分成不重叠的子集，使得每个数据对象恰在一个子集中，这种聚类类型称作（ B ）

A、层次聚类

B、划分聚类

C、非互斥聚类

D、模糊聚类

59 在基本K均值算法里，当邻近度函数采用（ A ）的时候，合适的质心是簇中各点的中位数。

A、曼哈顿距离

B、平方欧几里德距离

C、余弦距离

D、Bregman散度

60.（ C ）是一个观测值，它与其他观测值的差别如此之大，以至于怀疑它是由不同的机制产生的。

A、边界点

B、质心

C、离群点

D、核心点

填空题

第一章

（1）数据库中的知识挖掘(KDD)包括以下七个步

骤：、、

、、、和

（2）数据挖掘的性能问题主要包括：、和

（3）当前的数据挖掘研究中，最主要的三个研究方向是：、

和

（4）在万维网(WWW)上应用的数据挖掘技术常被称为：

（5）孤立点是指：

答案：

（1）数据清理，数据集成，数据选择，数据变换，数据挖掘，模式评估，知识表示

（2）算法的效率、可扩展性和并行处理

（3）统计学、数据库技术和机器学习

（4）WEB挖掘

（5）一些与数据的一般行为或模型不一致的孤立数据

第二章

（1）进行数据预处理时所使用的主要方法包

括：、、

和

（2）处理噪声数据的方法主要包括：、、

和

（3）模式集成的主要问题包括：和

（4）数据概化是指：

（5）数据压缩可分为：和两种类型。

（6）进行数值归约时，三种常用的有参方法是：、

和

（7）数据离散度的最常用度量是、和

答案：

（1）数据清理、数据集成、数据变换、数据规约

（2）分箱、聚类、计算机和人工检查结合、回归

（3）整合不同数据源中的元数据，实体识别问题

（4）沿概念分层向上概化

（5）有损压缩，无损压缩

（6）线性回归方法，多元回归，对数线性模型

（7）五数概括、中间四分位数区间、标准差

第三章

（1）概念分层有四种类型，分别是：、、

和

（2）常用的四种兴趣度的客观度量是：、、

和

（3）同时满足和的关联规则称为强关联规则。

答案：

（1）模式分层，集合分组分层，操作导出的分层，基于规则的分层

（2）简单性、确定性、实用性、新颖性

（3）最小置信度临界值、最小支持度临界值

第四章

（1）关联规则挖掘中，两个主要的兴趣度度量是：和

（2）Aprior算法包括和两个基本步骤

（3）项集的频率是指

（4）大型数据库中的关联规则挖掘包含两个过程：和

（5）根据规则中所处理的值类型，关联规则可分为：和

（6）Apriori性质是指：

（7）挖掘多维关联规则的技术可以根据量化属性的处理分为三种基本方法：、

和

（8）对于频繁项集挖掘，在挖掘过程中使用的约束包括以下五种类

型：、

、、和

（9）在多维关联规则挖掘中，我们搜索的不是频繁项集，而是

答案：

（1）支持度和置信度

（2）连接和剪枝

（3）包含项集的事务数

（4）找出所有频繁项集、由频繁项集产生强关联规则

（5）布尔关联规则、量化关联规则

（6）频繁项集的所有非空子集也必须是频繁的

（7）量化属性的静态离散化、量化关联规则、基于距离的关联规则

（8）反单调的、单调的、简洁的、可转变的、不可转变的

（9）频繁谓词集

第五章

（1）通过对数据进行预处理，可以提高分类和预测过程的、

和

（2）防止分类中的过分适应的两种方法分别是：和

答案

（1）准确性、有效性和可伸缩性

（2）先剪枝、后剪枝

第六章

（1）在数据挖掘中，常用的聚类算法包

括：、、、基于网格的方法和基于模型的方法。

（2）聚类分析常作为一个独立的工具来获得

（3）一个好的聚类分析方法会产生高质量的聚类，具有两个特征：

和

（4）许多基于内存的聚类算法所常用的两种数据结构是和（5）基于网格的聚类方法的优点是：

（6）孤立点产生的主要原因包括：和

（7）在基于统计的孤立点检测中，常用于不一致性检验的参数包

括：、

和

答案：

（1）划分方法、层次的方法、基于密度的方法

（2）数据分布的情况

（3）高类内相似度、低类间相似度

（4）数据矩阵、相异度矩阵

（5）处理数度快

（6）度量或执行错误、数据变异的结果

（7）数据分布、分布参数、预期的孤立点数

问答题

1.何谓数据挖掘它有哪些方面的功能

2.

从大量的、不完全的、有噪声的、模糊的、随机的数据中，提取隐含在其中的、人们事先不知道的、但又是潜在有用的信息和知识的过程称为数据挖掘。相关的名称有知识发现、数据分析、数据融合、决策支持等。

数据挖掘的功能包括：概念描述、关联分析、分类与预测、聚类分析、趋势分析、孤立点分析以及偏差分析等。

3.何谓数据仓库为什么要建立数据仓库

4.

数据仓库是一种新的数据处理体系结构，是面向主题的、集成的、不可更新的(稳定性)、随时间不断变化(不同时间)的数据集合，为企业决策支持系统提供所需的集成信息。

建立数据仓库的目的有3个：

一是为了解决企业决策分析中的系统响应问题，数据仓库能提供比传统事务数据库更快的大规模决策分析的响应速度。

二是解决决策分析对数据的特殊需求问题。决策分析需要全面的、正确的集成数据，这是传统事务数据库不能直接提供的。

三是解决决策分析对数据的特殊操作要求。决策分析是面向专业用户而非一般业务员，需要使用专业的分析工具，对分析结果还要以商业智能的方式进行表现，这是事务数据库不能提供的。

5.何谓粒度它对数据仓库有什么影响按粒度组织数据的方式有哪些

粒度是指数据仓库的数据单位中保存数据细化或综合程度的级别。粒度影响存放在数据仓库中的数据量的大小，同时影响数据仓库所能回答查询问题的细节程度。按粒度组织数据的方式主要有：

①简单堆积结构

②轮转综合结构

③简单直接结构

④连续结构

6.何谓聚类它与分类有什么异同

7.

聚类是将物理或抽象对象的集合分组成为多个类或簇(cluster)的过程，使得在同一个簇中的对象之间具有较高的相似度，而不同簇中的对象差别较大。

聚类与分类不同，聚类要划分的类是未知的，分类则可按已知规则进行；聚类是一种无指导学习，它不依赖预先定义的类和带类标号的训练实例，属于观察式学习，分类则属于有指导的学习，是示例式学习。

8.分类知识的发现方法主要有哪些分类过程通常包括哪两个步骤

9.

分类规则的挖掘方法通常有：决策树法、贝叶斯法、人工神经网络法、粗糙集法和遗传算法。分类的过程包括2步：首先在已知训练数据集上，根据属性特征，为每一种类别找到一个合理的描述或模型，即分类规则；然后根据规则对新数据进行分类。

10.什么是决策树如何用决策树进行分类

11.

决策树是用样本的属性作为结点，用属性的取值作为分支的树结构。

它是利用信息论原理对大量样本的属性进行分析和归纳而产生的。决策树的根结点是所有样本中信息量最大的属性。树的中间结点是以该结点为根的子树所包含的样本子集中信息量最大的属性。决策树的叶结点是样本的类别值。

决策树用于对新样本的分类，即通过决策树对新样本属性值的测试，从树的根结点开始，按照样本属性的取值，逐渐沿着决策树向下，直到

树的叶结点，该叶结点表示的类别就是新样本的类别。决策树方法是数

据挖掘中非常有效的分类方法。

12.简述ID3算法的基本思想及其主算法的基本步骤。

首先找出最有判别力的因素，然后把数据分成多个子集，每个子集又选择最有判别力的因素进一步划分，一直进行到所有子集仅包含同一类

型的数据为止。最后得到一棵决策树，可以用它来对新的样例进行分

类。

主算法包括如下几步：

①从训练集中随机选择一个既含正例又含反例的子集(称为窗口)；

②用“建树算法”对当前窗口形成一棵决策树；

③对训练集(窗口除外)中例子用所得决策树进行类别判定，找出错判

的例子；

④若存在错判的例子，把它们插入窗口，重复步骤②，否则结束。

13.噪声数据的产生原因有哪些？

(1)数据采集设备有问题

(2)在数据录入过程中发生了人为或计算机错误

(3)数据传输过程中发生错误

(4)由于命名规则或数据代码不同而引起的不一致。

14.遗传算法与传统寻优算法相比有什么特点？

①遗传算法为群体搜索，有利于寻找到全局最优解；

②遗传算法采用高效有方向的随机搜索，搜索效率高；

③遗传算法处理的对象是个体而不是参变量，具有广泛的应用领域；

④遗传算法使用适应值信息评估个体，不需要导数或其他辅助信息，

运算速度快，适应性好；

⑤遗传算法具有隐含并行性，具有更高的运行效率。

15.请解释一下在数据挖掘关联规则中什么是支持度和可信度，以及关联规

则。

支持度：规则A→B的支持度指的是所有事件中A与B同地发生的的概率，即P(A∪B)，是AB同时发生的次数与事件总次数之比。支持度是对关联规则重要性的衡量。

可信度：规则A→B的可信度指的是包含A项集的同时也包含B项集的条件概率P(B|A)，是AB同时发生的次数与A发生的所有次数之比。可信度是对关联规则的准确度的衡量。

关联规则：同时满足最小支持度阈值和最小可信度阈值的规则称之为关联规则。

分析与计算题

1. 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）：

13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,33,35,35,35,35,36, 40,45,46,52,70。

(a) 使用按箱平均值平滑对以上数据进行平滑，箱的深度为3。

(b) 该数据的均值是多少，中位数是多少？

(c) 使用 min-max规范化，将age值35转换到[0.0，1.0]区间。

答：(a)已知数据元组中 age 的值如下(按递增序)：

13,15,16,16,19,20,20,21,22,22,25,25,25,25,30,33,33,33,35,35,35,35 ,36,40,45,46,52,70。

且箱的深度为 3，划分为(等频)箱：

箱 1：13,15,16

箱 2：16,19,20

箱 3：20,21,22

箱 4：22,25,25

箱 5：25,25,30

箱 6：33,33,33

箱 7：35,35,35

箱 8：35,36,40

箱 9：45,46,52

箱 10：70

用箱均值光滑：

箱 1：15,15,15

箱 2：18,18,18

箱 3：21,21,21

箱 4：24,24,24

箱 5：27,27,37

箱 6：33,33,33

箱 7：35,35,35

箱 8：37,37,37

箱 9：48,48,48

箱 10：70；

（b）答：

（c）答：

2. 给定两个向量对象，分别表示为 p1(22，1，42，10)，p2(20，0，36，8)：

(a) 计算两个对象之间的欧几里得距离；

(b) 计算两个对象之间的曼哈顿距离；

(c) 计算两个对象之间的切比雪夫距离；

(d) 计算两个对象之间的闵可夫斯基距离，用 x=3。

答：(a) 计算两个对象之间的欧几里得距离：

(b) 计算两个对象之间的曼哈顿距离：

(c)计算两个对象之间的闵可夫斯基距离，其中参数 r=3：

3. 数据库有4笔交易，设minsup=60%， minconf=80%。

用 Apriori 算法找出所有频繁项集，列出所有关联规则。

4. 给定以下数据集（2，4，10，12，15，3，21），进行K-Means聚类，设定聚类数为2个，即k=2。在给出的数据集中随机选择的两个对象作为初始簇中心，分别是m1=2，m2=4，相似度按照欧式距离计算。求：

（1）第一次循环（迭代）结束时，划分所得的两个簇分别是多少？

（2）第一次循环（迭代）结束后，进行下一次循环（迭代）时簇心是多少？解：（1）当m

=2时，样本（2，4，10，12，15，3，21）距离该代表点的距离

1

分别为2，8，10，13，1，19。

=4时，样本（2，4，10，12，15，3，21）距离该代表点的距离分别为-

当m

2

2，6，8，11，-1，17。

最小距离是1或者-1将该元素放入m

=2的聚类中，则该聚类为（2，3）；

1

=4为（4，10，12，15，21）。

另一个聚类m

2

（2）完成数据样本的划分之后，对于每一个聚类，计算其中所有数据样本的均值，并且将其作为该聚类的新的代表点，由此得到k个均值代表点：m1=2.5，m2=12。