习题参考答案

第 1 章绪论

1.1 数据挖掘处理的对象有哪些？请从实际生活中举出至少三种。

答：数据挖掘处理的对象是某一专业领域中积累的数据，对象既可以来自社会科学,又可以来自自然科学产生的数据,还可以是卫星观测得到的数据。数据形式和结构也各不相同,可以是传统的关系数据库,可以是面向对象的高级数据库系统,也可以是面向特殊应用的数据库,如空间数据库、时序数据库、文本数据库和多媒体数据库等,还可以是 Web 数据信息。

实际生活的例子：

①电信行业中利用数据挖掘技术进行客户行为分析，包含客户通话记录、通话时间、所开通的服务等，据此进行客户群体划分以及客户流失性分析。

②天文领域中利用决策树等数据挖掘方法对上百万天体数据进行分类与分析，帮助天文学家发现其他未知星体。

③制造业中应用数据挖掘技术进行零部件故障诊断、资源优化、生产过程分析等。

④市场业中应用数据挖掘技术进行市场定位、消费者分析、辅助制定市场营销策略等。

1.2 给出一个例子，说明数据挖掘对商务的成功是至关重要的。该商务需要什么样的数据挖掘功能？它们能够由数据查询处理或简单的统计分析来实现吗？

答：例如，数据挖掘在电子商务中的客户关系管理起到了非常重要的作用。随着各个电子商务网站的建立，企业纷纷地从“产品导向”转向“客户导向”，如何在保持现有的客户同时吸引更多的客户、如何在客户群中发现潜在价值，一直都是电子商务企业重要任务。但是，传统的数据分析处理，如数据查询处理或简单的统计分析，只能在数据库中进行一些简单的数据查询和更新以及一些简单的数据计算操作，却无法从现有的大量数据中挖掘潜在的价值。而数据挖掘技术却能使用如聚类、关联分析、决策树和神经网络等多种方法，对数据库中庞大的数据进行挖掘分析，然后可以进行客户细分而提供个性化服务、可以利用挖掘到的历史流失客户的特征来防止客户流失、可以进行产品捆绑推荐等，从而使电子商务更好地进行客户关系管理，提高客户的忠诚度和满意度。

1.3 假定你是 Big-University 的软件工程师，任务是设计一个数据挖掘系统，分析学校课程数据库。该数据库包括如下信息：每个学生的姓名、地址和状态(例如，本科生或研究生)、所修课程，以及他们的 GPA。描述你要选取的结构，该结构的每个成分的作用是什么？

答：任务目的是分析课程数据库，那么首先需要有包含信息的关系型数据库系统，以便查找、提取每个属性的值；在取得数据后，需要有特征选择模块，通过特征选择，找出要分析的属性；接下来需要一个数据挖掘算法，或者数据挖掘软件，它应该包含像分类、聚类、关联分析这样的分析模块，对选择出来的特征值进行分析处理；在得到结果后，可以用可视化软件进行显示。

1.4 假定你作为一个数据挖掘顾问，受雇于一家因特网搜索引擎公司。通过特定的例子说明，数据挖掘可以为公司提供哪些帮助，如何使用聚类、分类、关联规则挖掘和离群点检测等技术为企业服务。

答：

(1) 使用聚类发现互联网中的不同群体，用于网络社区发现；

(2) 使用分类对客户进行等级划分，从而实施不同的服务；
(3) 使用关联规则发现大型数据集中间存在的关系，用于推荐搜索。如大部分搜索了“广外”的人都会继续搜索“信息学院”，那么在搜索“广外”后会提示是否进进一步搜索“信息学院”。
(4) 使用离群点挖掘发现与大部分对象不同的对象，用于分析针对网络的秘密收集信息的攻击。

1.5 定义下列数据挖掘功能：关联、分类、聚类、演变分析、离群点检测。使用你熟悉的生活中的数据，给出每种数据挖掘功能的例子。

答：关联是指发现样本间或样本不同属性间的关联。例如，一个数据挖掘系统可能发现的关联规则为：major(X, “computing science”)⇒owns(X, “personal computer”) [support=12%, confidence=98%] 其中，X 是一个表示学生的变量。该规则指出主修计算机科学并且拥有一台个人计算机的学生所占比例为 12%，同时，主修计算机专业的学生有 98%拥有个人计算机。

分类是构造一系列能描述和区分数据类型或概念的模型(或功能)，分类被用作预测目标数据的类的标签。例如，通过对过去银行客户流失与未流失客户数据的分析，得到一个预测模型，预测新客户是否可能会流失。

聚类是将数据划分为相似对象组的过程，使得同一组中对象相似度最大而不同组中对象相似度最小。例如，通过对某大型超市客户购物数据进行聚类，将客户聚类细分为低值客户、高值客户以及普通客户等。

数据演变分析描述和模型化随时间变化的对象的规律或趋势，尽管这可能包括时间相关数据的特征化、区分、关联和相关分析、分类、或预测，这种分析的明确特征包括时间序列数据分析、序列或周期模式匹配、和基于相似性的数据分析 。

离群点检测就是发现与众不同的数据。可用于发现金融领域的欺诈检测。

1.6 根据你的观察，描述一个可能的知识类型，它需要由数据挖掘方法发现，但本章未列出。它需要一种不同于本章列举的数据挖掘技术吗？

答：建立一个局部的周期性作为一种新的知识类型，只要经过一段时间的偏移量在时间序列中重复发生，那么在这个知识类型中的模式是局部周期性的。需要一种新的数据挖掘技术解决这类问题。

1.7 讨论下列每项活动是否是数据挖掘任务：

(1)根据性别划分公司的顾客。
(2)根据可赢利性划分公司的顾客。
(3)计算公司的总销售额。
(4)按学生的标识号对学生数据库排序。
(5)预测掷一对骰子的结果。
(6)使用历史记录预测某公司未来的股票价格。
(7)监视病人心率的异常变化。
(8)监视地震活动的地震波。
(9)提取声波的频率。

答： (1) 不是，这属于简单的数据库查询。
(2) 不是，这个简单的会计计算；但是新客户的利润预测则属于数据挖掘任务。
(3) 不是，还是简单的会计计算。

(4) 不是，这是简单的数据库查询。
(5) 不是，由于每一面都是同等概率，则属于概率计算；如概率是不同等的，根据历史数据预测结果则更类似于数据挖掘任务。
(6) 是，需要建立模型来预测股票价格，属于数据挖掘领域中的预测模型。可以使用回归来建模，或使用时间序列分析。
(7) 是，需要建立正常心率行为模型，并预警非正常心率行为。这属于数据挖掘领域的异常检测。若有正常和非正常心率行为样本，则可以看作一个分类问题。
(8) 是，需要建立与地震活动相关的不同波形的模型，并预警波形活动。属于数据挖掘领域的分类。
(9) 不是，属于信号处理。

第 2 章数据处理基础

2.1 将下列属性分类成二元的、分类的或连续的，并将它们分类成定性的(标称的或序数的)或定量的(区间的或比率的)。
例子：年龄。回答：分类的、定量的、比率的。
(a)用 AM 和 PM 表示的时间。
(b)根据曝光表测出的亮度。
(c)根据人的判断测出的亮度。
(d)医院中的病人数。
(e)书的 ISBN 号。
(f)用每立方厘米表示的物质密度。

答：(a)二元，定量，比率；
(b)连续，定量 ，比率；
(c)分类，定性，标称；
(d)连续，定量，比率；
(e)分类，定性，标称；
(f)连续，定量，比率。

2.2 你能想象一种情况，标识号对于预测是有用的吗？

答：学生的 ID 号可以预测该学生的毕业日期。

2.3 在现实世界的数据中，元组在某些属性上缺失值是常有的。请描述处理该问题的各种方法。

答：处理遗漏值问题的策略有如下几种。
(1) 删除数据对象或属性。一种简单而有效的策略是删除具有遗漏值的数据对象。然而，即使部分给定的数据对象也包含一些信息，并且，如果许多对象都有遗漏值，则很难甚至不可能进行可靠的分析。尽管如此，如果一个数据集只有少量的对象具有遗漏值，则忽略他们可能是合算的。一种相关的策略是删除具有遗漏值的属性。然而，做这件事要小心，因为被删除的属性可能对分析是至关重要的。
(2) 估计遗漏值。有时，遗漏值可以可靠地估计。例如，在考虑以较平滑的方式变化的具有少量但大大分散的遗漏值的时间序列，遗漏值可以使用其他值来估计(插值)。作为另一个例子，考虑一个具有许多相似数据点的数据集。在这种情况下，与具有遗漏值的点邻近的点的属性值常常可以用来估计遗漏的值。如果属性是连续的，则可以使用最近邻的平均属性值；如果属性是分类的，则可以取最近邻中最常出现的

属性值。

(3) 在分析时忽略遗漏值。许多数据挖掘方法都可以修改，忽略遗漏值。例如。假定正在对数据对象聚类，需要计算数据对象间的相似性；如果对于某属性，两个对象之一或两个对象都有遗漏值，则可以仅使用没有遗漏值的属性来计算相似性。当然，这种相似性只是紧邻的，但是除非整个属性数目很少，或者遗漏值的数量很大，否则这种误差影响不大。同样的，许多分类方法都可以修改，处理遗漏值。

2.4 以下规范方法的值域是什么？

(a) min-max 规范化。

(b) z-score 规范化。

(c) 小数定标规范化。

答：(a)[new_min,new_max]；

(b)(-∞,+ ∞)；

(c)(-1.0，1.0)。

2.5 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。解释你的步骤。评论对于给定的数据，该技术的效果。

(b) 对于数据平滑，还有哪些其它方法？

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)对于数据平滑，其它方法有：

(1)回归：可以用一个函数(如回归函数)拟合数据来光滑数据；

(2)聚类：可以通过聚类检测离群点，将类似的值组织成群或簇。直观地，落在簇集合之外的值视为离群点。

2.6 使用习题 2.5 给出的 age 数据，回答以下问题：

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

(b) 使用 z-score 规范化转换 age 值 35，其中，age 的标准偏差为 12.94 年。

(c) 使用小数定标规范化转换 age 值 35。

(d) 指出对于给定的数据，你愿意使用哪种方法。陈述你的理由。

答：(a)已知最大值为 70，最小值为 13，则可将 35 规范化为：$\frac{|35-13|}{70-13}=0.386$；

(b)已知均值为 30，标准差为 12.94，则可将 35 规范化为：$\frac{35-30}{12.94}=0.386$；

(c)使用小数定标规范化可将 35 规范化为：$\frac{35}{100}=0.35$；

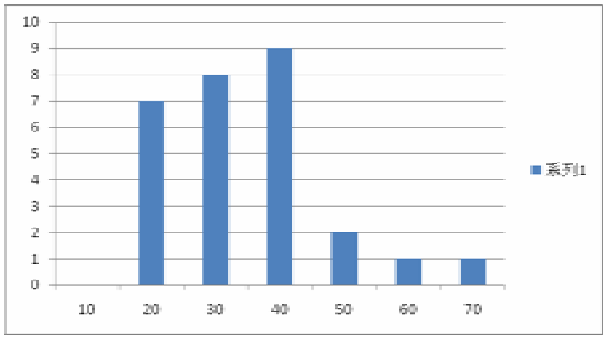
(d)对于给定的数据，你愿意使用 min-max 规范化。理由是计算简单。

2.7 使用习题 2.5 给出的 age 数据

(a) 画一个宽度为 10 的等宽的直方图。

(b) 为以下每种抽样技术勾画例子：有放回简单随机抽样，无放回简单随机抽样，聚类抽样，分层抽样。使用大小为 5 的样本和层“青年”,“中年”和“老年”。

答：(a)如下为宽度为 10 的等宽的直方图：



(b)已知样本大小为 5 和层“ 青年 ”,“ 中年 ”和“ 老年 ”,

(1)有放回简单随机抽样：30,33,30,25,30

(2)无放回简单随机抽样：30,33,33,35,25

(3)聚类抽样：16,25,33,35,46

(4)分层抽样：25,35,52

2.8 以下是一个商场所销售商品的价格清单(按递增顺序排列，括号中的数表示前面数字出现次数)1(2)、5(5)、8(2)、10(4)、12、14(3)、15(5)、18(8)、20(7)、21(4)、25(5)、28、30(3)。请分别用等宽的方法和等高的方法对上面的数据集进行划分。

答：(1)等宽方法：划分为 3 个数据集，每个数据集的宽度为价格 10。价格在 1—10 之间出现次数为 13；价格在 11—20 之间出现的次数为 24；价格在 21—30 之间出现的次数为 13。

(2)等高方法：划分为 2 个数据集，每个数据集的高度为出现的次数 4。出现次数 1—4

之间的价格为 1、8、10、12、14、21、28、30，共 8 个数据；出现次数 5—8 之间的价格为 5、15、18、20、25，共 5 个数据。

2.9 讨论数据聚合需要考虑的问题。

答：数据聚合需要考虑的问题有：

(1)模式识别：这主要是实体识别问题；

(2)冗余：一个属性是冗余的，即它能由另一个表导出，如果属性或维的命名不一致，也可能导致冗余，可以用相关分析来检测；

(3)数据值冲突的检测与处理：有些属性因表示比例或编码不同，会导致属性不同。

2.10 假定我们对一个比率属性 x 使用平方根变换，得到一个新属性 x*。作为分析的一部分，你识别出区间(a, b)，在该区间内，x*与另一个属性 y 具有线性关系。

(a)换算成 x, (a, b)的对应区间是什么？

(b)给出 y 关联 x 的方程。

答：(a)(a^2，b^2)；

(b)Y=kx^0.5 +C (k, C 是常数)。

2.11 讨论使用抽样减少需要显示的数据对象个数的优缺点。简单随机抽样(无放回)是一种好的抽样方法吗？为什么是，为什么不是？

答：抽样减少需要显示的数据对象个数的优点是减少处理数据的费用和时间。缺点是不能利用总体的已知信息和代表总体数据的信息。简单随机抽样(无放回)不是一种好的抽样方法，不能充分地代表不太频繁出现的对象类型和每个对象被选中的概率不一样。

2.12 给定 m 个对象的集合，这些对象划分成 K 组，其中第 i 组的大小为 m_i。如果目标是得到容量为 $n<m$ 的样本，下面两种抽样方案有什么区别？(假定使用有放回抽样)

(a)从每组随机地选择 $n\times m_i/m$ 个元素。

(b)从数据集中随机地选择 n 个元素，而不管对象属于哪个组。

答：(a)组保证了可以在每个组里面得到等比例的样本，而(b)组在每个组里面抽取的样本的个数是随机的，不能保证每个组都能抽到样本。

2.13 一个地方公司的销售主管与你联系，他相信他已经设计出了一种评估顾客满意度的方法。他这样解释他的方案：“ 这太简单了，我简直不敢相信，以前竟然没有人想到，我只是记录顾客对每种产品的抱怨次数，我在数据挖掘的书中读到计数具有比率属性，因此，我的产品满意度度量必定具有比率属性。但是，当我根据我的顾客满意度度量评估产品并拿给老板看时，他说我忽略了显而易见的东西，说我的度量毫无价值。我想，他简直是疯了，因为我们的畅销产品满意度最差，因为对它的抱怨最多。你能帮助我摆平他吗？ ”

(a)谁是对的，销售主管还是他的老板？如果你的答案是他的老板，你做些什么来修正满意度度量？

(b)对于原来的产品满意度度量的属性类型，你能说些什么？

答: (a) 老板是对的。更好的衡量方法应该如下：

不满意率(产品)=每种产品的抱怨次数/该产品的总销售量

(b) 原来衡量方法的属性类型是没有意义的。例如，两件商品有相同的顾客满意度可能会有不同的抱怨次数，反之亦然。

2.14 考虑一个文档-词矩阵，其中 tf_{ij} 是第 i 个词(术语)出现在第 j 个文档中的频率，而 m 是文档数。考虑由下式定义的变量变换：$tf'_{ij} = tf_{ij} \cdot \log\frac{m}{df_i}$

其中，df_i 是出现 i 个词的文档数，称作词的文档频率(document frequency)。该变换称作逆文档频率变换(inverse document frequency)。

(a)如果出现在一个文档中，该变换的结果是什么？如果术语出现在每个文档中呢？

(b)该变换的目的可能是什么？

答：(a) 如果该词出现在每一个文档中，它的词权就会为 0，但是如果这个词仅仅出现在一个文档中，它就有最大的词权，例如，log m 。

(b) 这个变换反映了以下一个现象：当一个词出现在每一个文档中，对于文档与文档之间，该词没有区分能力，但是那些只是某一两篇文档出现的词，其区分文档的能力就较强。

2.15 对于下面的向量 x 和 y，计算指定的相似性或距离度量。

(a)x=(1，1，1，1)，y=(2，2，2，2) 余弦相似度、相关系数、欧几里得。

(b) x=(0，1，0，1)，y=(1，0，1，0) 余弦相似度、相关系数、欧几里得、Jaccard 系数。

(c) x=(2，-1,0，2，0，-3)，y=(-1，1，-1，0，0，-1) 余弦相似度、相关系数。

答：(a) 余弦相似度、相关系数、欧几里得分别是 0.5，0，2；

(b) 余弦相似度、相关系数、欧几里得、Jaccard 系数分别是 0，1，2，0；

(c) 余弦相似度、相关系数分别是 0，0。

2.16 简单地描述如何计算由以下类型的变量描述的对象间的相异度：

(a) 不对称的二元变量

(b) 分类变量

(c) 比例标度型(ratio-scaled)变量

(d) 数值型变量

答：

(a) 使用 Jaccard 系数计算不对称的二元变量的相异度；

(b) 采用属性值匹配的方法(属性值匹配，相似度为 1，否则为 0)可以计算用分类变量描述的对象间的相异度；

(c) 对比例标度变量进行对数变换，对变换得到的值采用与处理区间标度变量相同的方法来计算相异度；

(d) 可采用欧几里得距离公式或曼哈顿距离公式计算。

2.17 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

$$d_{12} = \sqrt{(22-20)^2 + (1-0)^2 + (42-36)^2 + (10-8)^2} = \sqrt{45}$$

(b) 计算两个对象之间的曼哈顿距离

$$d_{12} = |22-20| + |1-0| + |42-36| + |10-8| = 11$$

(c) 计算两个对象之间的闵可夫斯基距离，其中参数 r=3

$$d_{12} = \sqrt[3]{|22-20|^3 + |1-0|^3 + |42-36|^3 + |10-8|^3} = \sqrt[3]{233}$$

2.18 以下表格包含了属性 name，gender，trait-1，trait-2，trait-3，及 trait-4，这里的 name 是对象的 id，gender 是一个对称的属性，剩余的 trait 属性是不对称的，描述了希望找到的笔友的个人特点。假设有一个服务是试图发现合适的笔友。

name	gender	trait-1	trait-2	trait-3	trait-4
Keavn	M	N	P	P	N
Caroline	F	N	P	P	N
Erik	M	P	N	N	P

对不对称的属性的值，值 P 被设为 1，值 N 被设为 0。

假设对象(潜在的笔友)间的距离是基于不对称变量来计算的。

(a) 计算对象间的简单匹配系数；

(b) 计算对象间的 Jaccard 系数；

(c) 你认为哪两个人将成为最佳笔友？哪两个会是最不能相容的？

(d) 假设我们将对称变量 gender 包含在我们的分析中。基于 Jaccard 系数，谁将是最和谐的一对？为什么？

答：

(a) 计算对象间的简单匹配系数

SMC (Keavn, Caroline) = (2+2)/(0+0+2+2) = 1

SMC(Keavn, Erik) = (0+0)/(2+2+0+0) = 0

SMC(Caroline,Erik) = (0+0)/(2+2+0+0) = 0

(b) 计算对象间的 Jaccard 系数

Jaccard (Keavn, Caroline) = 2/(2+0+0) = 1

Jaccard (Keavn, Erik) = 0/(0+2+2) = 0

Jaccard (Caroline,Erik) = 0/(0+2+2) = 0

(c) 根据属性的匹配程度，Keavn 和 Caroline 将成为最佳笔友，Caroline 和 Erik 会是最不能相容的。

(d) 若将对称变量 gender 包含在分析中，设值 M 被设为 1，值 F 被设为 0，

Jaccard (Keavn, Caroline) = 2/(2+1+0) = 2/3

Jaccard (Keavn, Erik) = 1/(1+2+2) = 1/5

Jaccard (Caroline,Erik) = 0/(0+2+3) = 0

因为 Jaccard (Keavn, Caroline)最大，因此，Keavn 和 Caroline 是最和谐的一对。

2.19 给定一个在区间[0，1]取值的相似性度量，描述两种将该相似度变换成区间[0，∞]中的相异度的方法。

答：取倒数减一：$d(p,q) = \frac{1}{s(p,q)} - 1$

取对数：$d(p,q) = -\log(s(p,q))$

第 3 章分类与回归

3.1 简述决策树分类的主要步骤。

答：决策树生成的过程如下：

(1)对数据源进行数据预处理，得到训练集和测试集；

(2)对训练集进行训练；

(3)对初始决策树进行树剪枝；

(4)由所得到的决策树提取分类规则；

(5)使用测试数据集进行预测，评估决策树模型；

3.2 给定决策树，选项有：(1)将决策树转换成规则，然后对结果规则剪枝，或(2)对决策树剪枝，然后将剪枝后的树转换成规则。相对于(2)，(1)的优点是什么？

答：相对于(2)，(1)的优点是：由于第一种方法已经将决策树转换成规则，通过规则，可以很快速的评估决策树以及其子树紧凑程度，不能提高规则的估计准确率的任何条件都可以减掉，从而泛化规则；

3.3 计算决策树算法在最坏情况下的时间复杂度是重要的。给定数据集 D，具有 m 个属性和|D|个训练记录，证明决策树生长的计算时间最多为 $m \times |D| \times \log(|D|)$。

答：假设训练集拥有|D|实例以及 m 个属性。我们需要对树的尺寸做一个假设，假设树的深度是由 log |D| 决定，即 O(log |D|)。考虑一个属性在树的所有节点上所要做的工作量。当然不必在每一个节点上考虑所有的实例。但在树的每一层，必须考虑含有|D|个实例的整个数据集。由于树有 log |D|个不同的层，处理一个属性需要的工作量是 $|D| \times \log(|D|)$。

在每个节点上所有属性都要被考虑，因此总的工作量为 $m \times |D| \times \log(|D|)$。

3.4 考虑表 3-23 所示二元分类问题的数据集。

表 3-23 习题 3.4 数据集

A	B	类标号
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
T	F	-

(1) 计算按照属性 A 和 B 划分时的信息增益。决策树归纳算法将会选择那个属性？

(2) 计算按照属性 A 和 B 划分时 Gini 系数。决策树归纳算法将会选择那个属性？

答：

按照属性 A 和 B 划分时，数据集可分为如下两种情况：

	A=T	A=F
+	4	0
-	3	3

	B=T	B=F
+	3	1
-	1	5

(1)

划分前样本集的信息熵为 $E=-0.4\log_2 0.4-0.6\log_2 0.6=0.9710$

按照属性 A 划分样本集分别得到的两个子集(A 取值 T 和 A 取值 F)的信息熵分别为:

$$\mathrm{E}_{\mathrm{A=T}}=-\frac{4}{7}\log_2\frac{4}{7}-\frac{3}{7}\log_2\frac{3}{7}=0.9852$$

$$\mathrm{E}_{\mathrm{A=F}}=-\frac{3}{3}\log_2\frac{3}{3}-\frac{0}{3}\log_2\frac{0}{3}=0$$

按照属性 A 划分样本集得到的信息增益为：$\Delta=E-\frac{7}{10}E_{A=T}-\frac{3}{10}E_{A=F}=0.2813$

按照属性 B 划分样本集分别得到的两个子集(B 取值 T 和 B 取值 F)的信息熵分别为:

$$\mathrm{E}_{\mathrm{B=T}}=-\frac{3}{4}\log_2\frac{3}{4}-\frac{1}{4}\log_2\frac{1}{4}=0.8113$$

$$\mathrm{E}_{\mathrm{B=F}}=-\frac{1}{6}\log_2\frac{1}{6}-\frac{5}{6}\log_2\frac{5}{6}=0.6500$$

按照属性 B 划分样本集得到的信息增益为：$\Delta=E-\frac{4}{10}E_{B=T}-\frac{6}{10}E_{B=F}=0.2565$

因此，决策树归纳算法将会选择属性 A。

(2)

划分前的 Gini 值为 $G=1-0.4^2-0.6^2=0.48$

按照属性 A 划分时 Gini 指标:

$$\mathrm{G}_{\mathrm{A=T}}=1-\left(\frac{4}{7}\right)^2-\left(\frac{3}{7}\right)^2=0.4898$$

$$\mathrm{G}_{\mathrm{A=F}}=1-\left(\frac{3}{3}\right)^2-\left(\frac{0}{3}\right)^2=0$$

Gini 增益 $\Delta=G-\frac{7}{10}G_{A=T}-\frac{3}{10}G_{A=F}=0.1371$

按照属性 B 划分时 Gini 指标:

$$\mathrm{G}_{\mathrm{B=T}}=1-\left(\frac{1}{4}\right)^2-\left(\frac{3}{4}\right)^2=0.3750$$

$$G_{B=F}=1-\left(\frac{1}{6}\right)^2-\left(\frac{5}{6}\right)^2=0.2778$$

Gini 增益 $\Delta=G-\frac{4}{10}G_{B=T}-\frac{6}{10}G_{B=F}=0.1633$

因此，决策树归纳算法将会选择属性 B。

3.5 证明：将结点划分为更小的后续结点之后，结点熵不会增加。

证明：根据定义可知，熵值越大，类分布越均匀；熵值越小，类分布越不平衡。假设原有的结点属于各个类的概率都相等，熵值为 1，则分出来的后续结点在各个类上均匀分布，此时熵值为 1，即熵值不变。假设原有的结点属于个各类的概率不等，因而分出来的

后续结点不均匀地分布在各个类上，则此时的分类比原有的分类更不均匀，故熵值减少。

3.6 为什么朴素贝叶斯称为“朴素”？简述朴素贝叶斯分类的主要思想。

答：朴素贝叶斯之所以称之为朴素是因为，它假设属性之间是相互独立的。

朴素贝叶斯分类的主要思想为：利用贝叶斯定理，计算未知样本属于某个类标号值的概率，根据概率值的大小来决定未知样本的分类结果。

(通过某对象的先验概率，利用贝叶斯公式计算出其后验概率，即该对象属于某一类的概率，选择具有最大后验概率的类作为该对象所属的类。)

3.7 考虑表 3-24 数据集，请完成以下问题：

表 3-24 习题 3.7 数据集

记录号	A	B	C	类
1	0	0	0	+
2	0	0	1	-
3	0	1	1	-
4	0	1	1	-
5	0	0	1	+
6	1	0	1	+
7	1	0	1	-
8	1	0	1	-
9	1	1	1	+
10	1	0	1	+

(1) 估计条件概率 $P(A|+)$， $P(B|+)$， $P(C|+)$， $P(A|-)$， $P(B|-)$， $P(C|-)$。

(2) 根据(1)中的条件概率，使用朴素贝叶斯方法预测测试样本(A=0，B=1，C=0)的类标号；

(3) 使用 Laplace 估计方法，其中 p=1/2，l=4，估计条件概率 $P(A|+)$，$P(B|+)$，$P(C|+)$，$P(A|-)$， $P(B|-)$， $P(C|-)$。

(4) 同(2)，使用(3)中的条件概率

(5) 比较估计概率的两种方法，哪一种更好，为什么？

答：(1) $P(A|+)=3/5$

$P(B|+)=1/5$

$P(A|-)=2/5$

$P(B|-)=2/5$

$P(C|-)=1$

(2) 假设 P(A=0,B=1,C=0)=K

则 K 属于两个类的概率为：

P(+|A=0,B=1,C=0)=P(A=0,B=1,C=0)×P(+)/K

=P(A=0|+)P(B|+)P(C=0|+)×P(+)/K=0.4×0.2×0.2×0.5/K=0.008/K

P(-|A=0,B=1,C=0)=P(A=0,B=1,C=0)×P(-)/K

=P(A=0|-)P(B|-)P(C=0|-)×P(-)/K=0.4×0.2×0×0.5/K=0/K

则得到，此样本的类标号是+。

(3) P(A|+)=(3+2)/(5+4)=5/9

P(A|-)=(2+2)/(5+4)=4/9

P(B|+)=(1+2)/(5+4)=1/3

P(B|-)=(2+2)/(5+4)=4/9

P(C|-)=(0+2)/(5+4)=2/9

(4) 假设 P(A=0,B=1,C=0)=K

则 K 属于两个类的概率为：

P(+|A=0,B=1,C=0)=P(A=0,B=1,C=0)×P(+)/K

=P(A=0|+)P(B|+)P(C=0|+)×P(+)/K

=(4/9) ×(1/3) ×(1/3) ×0.5/K=0.0247/K

P(-|A=0,B=1,C=0)=P(A=0,B=1,C=0)×P(-)/K

=P(A=0|-)P(B|-)P(C=0|-)×P(-)/K

=(5/9) ×(4/9) ×(2/9) ×0.5/K=0.0274/K

则得到，此样本的类标号是-。

(5) 当条件概率为 0 的时候，条件概率的预测用 Laplace 估计方法比较好，因为我们不想整个条件概率计算结果为 0.

3.8 考虑表 3-25 中的一维数据集。

表 3-25 习题 3.8 数据集

X	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
Y	-	-	+	+	+	-	-	+	-	-

根据 1-最近邻、3-最近邻、5-最近邻、9-最近邻，对数据点 x=5.0 分类，使用多数表决。

答： 1-最近邻：+

3-最近邻：-

5-最近邻：+

9-最近邻：-

3.9 表 3-26 的数据集包含两个属性 X 与 Y，两个类标号“ + ”和“ - ”。每个属性取三个不同值策略：0,1 或 2。“ + ”类的概念是 Y=1，“ - ”类的概念是 X=0 and X=2。

表 3-26 习题 3.9 数据集

X	Y	实例数	
		+	-
0	0	0	100
1	0	0	0
2	0	0	100
1	1	10	0
2	1	10	100
0	2	0	100
1	2	0	0
2	2	0	100

(1) 建立该数据集的决策树。该决策树能捕捉到“+”和“-”的概念吗？

(2) 决策树的准确率、精度、召回率和 F1 各是多少？(注意，精度、召回率和 F1 量均是对“+”类定义)

(3) 使用下面的代价函数建立新的决策树，新决策树能捕捉到“+”的概念么？

$$C(i,j)=\begin{cases}0 & \text{如果} i=j\\ 1 & \text{如果} i=+, j=-\\ \dfrac{-\text{实例个数}}{+\text{实例个数}} & \text{如果} i=-, j=+\end{cases}$$

(提示：只需改变原决策树的结点。)

答：(1)在数据集中有 20 个正样本和 500 个负样本，因此在根节点处错误率为

$$E=1-\max\left(\frac{20}{520},\frac{500}{520}\right)=\frac{20}{520}$$

如果按照属性 X 划分，则：

	X=0	X=1	X=2
+	0	10	10
-	200	0	300

$E_{X=0}$=0/310=0

$E_{X=1}$=0/10=0

$E_{X=2}$=10/310

$$\Delta_X=E-\frac{200}{520}\times 0-\frac{10}{520}\times 0-\frac{310}{520}\times\frac{10}{310}=\frac{10}{520}$$

如果按照属性 Y 划分，则：

	Y=0	Y=1	Y=2
+	0	20	0
-	200	100	200

$E_{Y=0}$=0/200=0

$E_{Y=1}$=20/120

$E_{Y=2}$=0/200=0

$$\Delta_X=E-\frac{120}{520}\times\frac{20}{120}=0$$

因此 X 被选为第一个分裂属性，因为 X=0 和 X=1 都是纯节点，所以使用 Y 属性去分割不纯节点 X=2。

Y=0 节点包含 100 个负样本，Y=1 节点包含 10 个正样本和 100 个负样本，Y=2 节点包含 100 个负样本，所以子节点被标记为“—”。整个结果为：

$$\text{类标记}=\begin{cases}+, X=1\\ -, \text{其他}\end{cases}$$

(2)

		预测类	
		+	-
实际类	+	10	10
	-	0	500

accuracy: $\frac{510}{520}$=0.9808，precision: $\frac{10}{10}$=1.0

recall: $\frac{10}{20}$=0.5 ， F-measure: $\frac{2*0.5*1.0}{1.0+0.5}$=0.6666

(3)由题可得代价矩阵为

		预测类	
		+	-
实际类	+	0	500/20=25
	-	1	0

决策树在(1)之后还有 3 个叶节点，X=2∧Y=0，X=2∧Y=1，X=2∧Y=2。其中 X=2∧Y=1 是不纯节点，误分类该节点为“+”类的代价为：10*0+100*1=100，误分该节点为“—”类的代价为：10*25+100*0=250。所以这些节点被标记为“+”类。

分类结果为：

$$\text{类标记}=\begin{cases}+ & X=1\vee(X=2\wedge Y=1)\\ - & \text{其他}\end{cases}$$

3.10 什么是提升？陈述它为何能提高决策树归纳的准确性？

答：提升是指给每个训练元组赋予权重，迭代地学习 k 个分类器序列，学习得到分类器 M_i 之后，更新权重，使得其后的分类器 M_{i+1}“更关注”M_i 误分的训练元组，最终提升的分类器 M*组合每个个体分类器，其中每个分类器投票的权重是其准确率的函数。在提升的过程中，训练元组的权重根据它们的分类情况调整，如果元组不正确地分类，则它的权重增加，如果元组正确分类，则它的权重减少。元组的权重反映对它们分类的困难程度，权重越高，越可能错误的分类。根据每个分类器的投票，如果一个分类器的误差率越低，提升就赋予它越高的表决权重。在建立分类器的时候，让具有更高表决权重的分类器对具有更高权重的元组进行分类，这样，建立了一个互补的分类器系列。所以能够提高分类的准确性。

3.11 表 3-27 给出课程数据库中学生的期中和期末考试成绩。

表 3-27 习题 3.11 数据集

期中考试	期末考试
X	Y
72	84
50	63
81	77
74	78
94	90
86	75
59	49
83	79
65	77
33	52
88	74

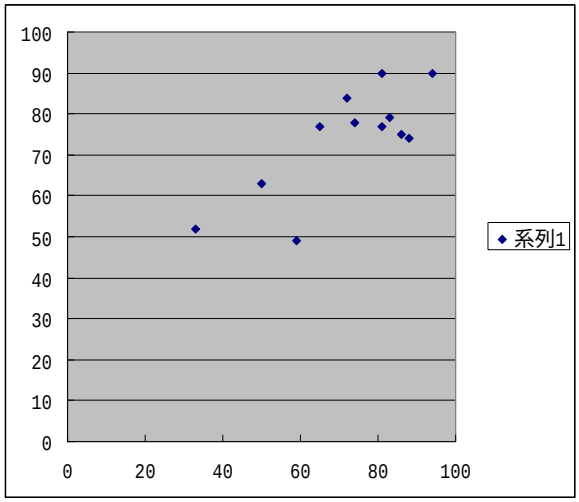
81	90

(1) 绘制数据的散点图。X 和 Y 看上去具有线性联系吗？

(2) 使用最小二乘法，由学生课程中成绩预测学生的期末成绩的方程式。

(3) 预测期中成绩为 86 分的学生的期末成绩。

答：(1)数据图如下所示：



X 和 Y 具有线性联系。

(2)

	X	Y	X*Y	X^2	预测 Y
1	72	84	6048	5184	73.9031
2	50	63	3150	2500	61.1079
3	81	77	6237	6561	79.1375
4	74	78	5772	5476	75.0663
5	94	90	8460	8836	86.6983
6	86	75	6450	7396	82.0455
7	59	49	2891	3481	66.3423
8	83	79	6557	6889	80.3007
9	65	77	5005	4225	69.8319
10	33	52	1716	1089	51.2207
11	88	74	6512	7744	83.2087
12	81	90	7290	6561	79.1375
SUM	866	888	66088	65942	

$Y = a + b*X$

$a = Y_0 + b*X_0$

$b = (\Sigma x_i y_i - nX_0Y_0)/(\Sigma x_i^2 - nX_0^2)$

$X_0 = (\Sigma x_i)/n$

$Y_0 = (\Sigma y_i)/n$

求得 $a = 32.0279$，$b = 0.5816$。

(3) 由(2)中表可得，预测成绩为 86 分的学生的期末成绩为 82.0455。

3.12 通过对预测变量变换，有些非线性回归模型可以转换成线性模型。指出如何将非线性回归方程 $y = ax^{\beta}$ 转换成可以用最小二乘法求解的线性回归方程。

答：令 $w = x^{\beta}$，对样本数据做变换 $w_i = x_i^{\beta}(i = 1,2,....,n)$，利用($w_i$，$Y_i$)(i=1，2，…，n)

解出 y = aw 中的 a，再代入 $y = ax^{\beta}$ 即得到 y 对 x 的回归方程。

第 4 章聚类分析

4.1 什么是聚类？简单描述如下的聚类方法：划分方法，层次方法，基于密度的方法，基于模型的方法。为每类方法给出例子。

答：聚类是将数据划分为相似对象组的过程，使得同一组中对象相似度最大而不同组中对象相似度最小。主要有以下几种类型方法：

(1)划分方法

给定一个有 N 个元组或者记录的数据集，分裂法将构造 K 个分组，每一个分组就代表一个聚类，K<N。而且这 K 个分组满足下列条件：第一，每一个分组至少包含一条记录；第二，每一条记录属于且仅属于一个分组(注意：这个要求在某些模糊聚类算法中可以放宽)；对于给定的 K，算法首先给出一个初始的分组方法，以后通过反复迭代的方法改变分组，使得每一次改进之后的分组方案都较前一次好，而所谓好的标准就是：同一分组中的记录越近越好，而不同分组中的记录越远越好。

使用这个基本思想的算法有：K-MEANS 算法、K-MEDOIDS 算法、CLARANS 算法。

(2)层次方法

这种方法对给定的数据集进行层次似的分解，直到某种条件满足为止。具体又可分为“自底向上”和“自顶向下”两种方案。例如在“自底向上”方案中，初始时每一个数据记录都组成一个单独的组，在接下来的迭代中，它把那些相互邻近的组合并成一个组，直到所有的记录组成一个分组或者某个条件满足为止。

代表算法有：BIRCH 算法、CURE 算法、CHAMELEON 算法等。

(3)基于密度的方法

基于密度的方法与其它方法的一个根本区别是：它不是基于各种各样的距离，而是基于密度的。这样就能克服基于距离的算法只能发现“类圆形”的聚类的缺点。这个方法的指导思想就是：只要一个区域中的点的密度大过某个阈值，就把它加到与之相近的聚类中去。

代表算法有：DBSCAN 算法、OPTICS 算法、DENCLUE 算法等。

(4)基于模型的方法

基于模型的方法给每一个聚类假定一个模型，然后去寻找能够很好的满足这个模型的数据。这样一个模型可能是数据点在空间中的密度分布函数或者其它。它的一个潜在假定就是：目标数据集是由一系列的概率分布所决定的。

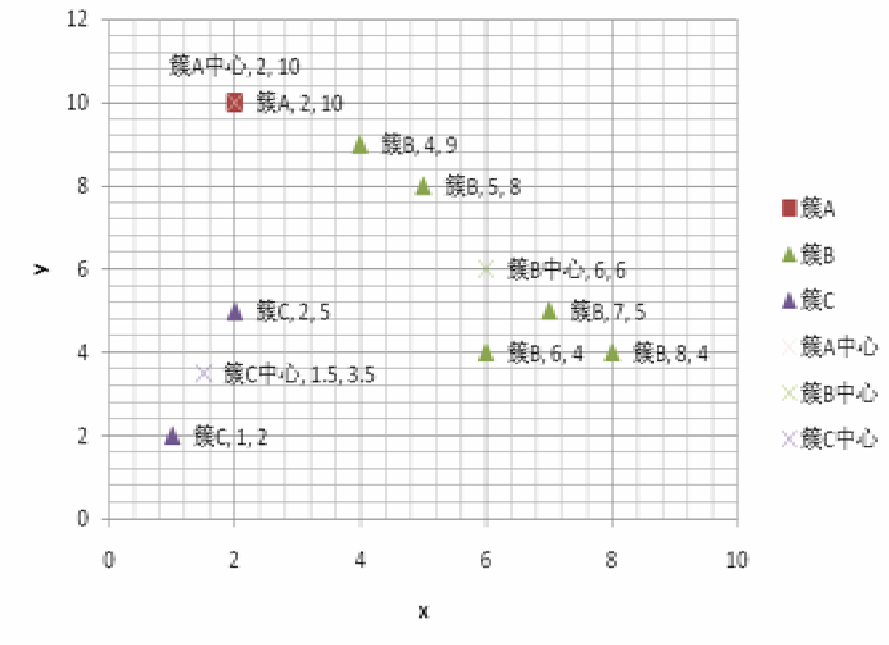
基于模型的方法主要有两类：统计学方法和神经网络方法(SOM)。

4.2 假设数据挖掘的任务是将如下的 8 个点(用(x,y)代表位置)聚类为三个簇。A1(2,10),A2(2,5),A3(8,4),B1(5,8),B2(7,5),B3(6,4),C1(1,2),C2(4,9)。距离函数是 Euclidean 函数。假设初始我们选择 A1,B1 和 C1 为每个簇的中心，用 k-means 算法来给出

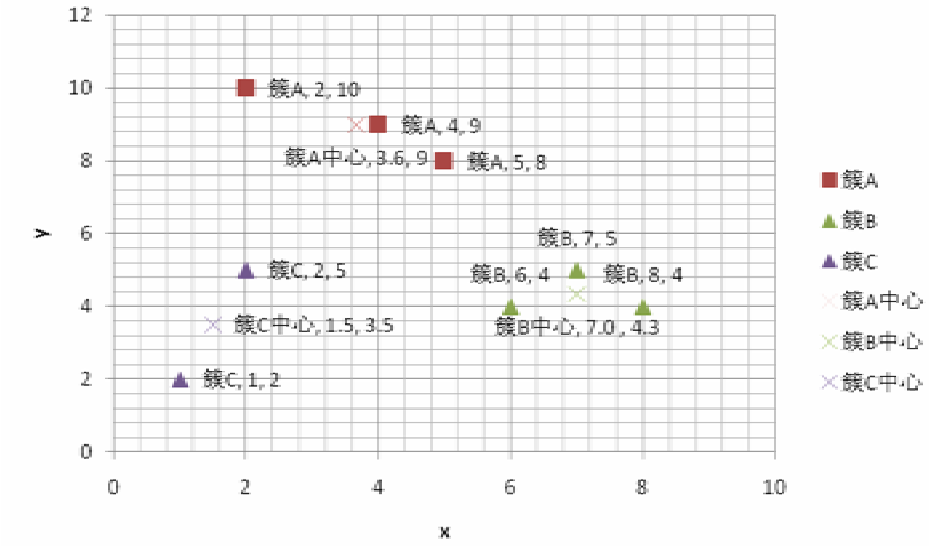
(a) 在第一次循环执行后的三个簇中心；

(b) 最后的三个簇中心及簇包含的对象。

答：(a)如图，



(b)如图，



4.3 聚类被广泛地认为是一种重要的数据挖掘方法，有着广泛的应用。对如下的每种情况给出一个应用例子：

(a) 采用聚类作为主要的数据挖掘方法的应用；

(b) 采用聚类作为预处理工具，为其它数据挖掘任务作数据准备的应用。

答：(a) 如电子商务网站中的客户群划分。根据客户的个人信息、消费习惯、浏览行为等信息，计算客户之间的相似度，然后采用合适的聚类算法对所有客户进行类划分；基于得到的客户群信息，相关的店主可以制定相应的营销策略，如交叉销售，根据某个客户群中的其中一个客户的购买商品推荐给另外一个未曾购买此商品的客户。

(b) 如电子商务网站中的推荐系统。电子商务网站可以根据得到的客户群，采用关联规则或者隐马尔科夫模型对每个客户群生成消费习惯规则，检测客户的消费模式，这些规则或模式可以用于商品推荐。其中客户群可以通过聚类算法来预先处理获取得到。

4.4 假设你将在一个给定的区域分配一些自动取款机以满足需求。住宅区或工作区可以被聚类以便每个簇被分配一个 ATM。但是，这个聚类可能被一些因素所约束，包括可能影响 ATM 可达性的桥梁，河流和公路的位置。其它的约束可能包括对形成一个区域的每个地域的 ATM 数目的限制。给定这些约束，怎样修改聚类算法来实现基于约束的聚类？

答：约束的存在会使得原来被定义在同一个簇的对象之间的距离发生变化。这时可以考虑将这些约束因素融入到距离的计算公式中，如在存在桥梁、河流和公路的区域中，可通过对象之间的连通性以及路径来计算距离；而地域 ATM 数目的限制问题则可在聚类的初始化阶段解决，如在 K-MEANS 的初始化时，可根据区域的个数来确定簇个数 K 的值，然后所选择的初始化种子尽可能分布在距离各个 ATM 附近的地方。

4.5 给出一个数据集的例子，它包含三个自然簇。对于该数据集，k-means(几乎总是)能够发现正确的簇，但二分 k-means 不能。

答：有三个完全一样的球型簇，三个簇的点的个数和分布密度以及位置均完全相同。其中两个簇关于第三个簇完全对称，则在 k-means 方法中可以很轻易找出这三个簇，但用二分 k-means 则会把处于对称中心的那个簇分成两半。

4.6 总 SSE 是每个属性的 SSE 之和。如果对于所有的簇，某变量的 SSE 都很低，这意味什么？如果只对一个簇很低呢？如果对所有的簇都很高？如果仅对一个簇高呢？如何使用每个变量的 SSE 信息改进聚类？

答：

(a) 如果对于所有的簇，某属性的 SSE 都很低，那么该属性值变化不大，本质上等于常量，对数据的分组没什么用处。

(b) 如果某属性的 SSE 只对一个簇很低，那么该属性有助于该簇的定义。

(c) 如果对于所有的簇，某属性的 SSE 都很高，那么意味着该属性是噪声属性。

(d) 如果某属性的 SSE 仅对一个簇很高，那么该属性与定义该簇的属性提供的信息不一致。在少数情况下，由该属性定义的簇不同于由其他属性定义的簇，但是在某些情况下，这也意味着该属性不利于簇的定义。

(e) 消除簇之间具有小分辨力的属性，比如对于所有簇都是低或高 SSE 的属性，因为他们对聚类没有帮助。对于所有簇的 SSE 都高且相对其他属性来说 SSE 也很高的属性特别麻烦，因为这些属性在 SSE 的总和计算中引入了很多的噪声。

4.7 使用基于中心、邻近性和密度的方法，识别图 4-19 中的簇。对于每种情况指出簇个数，并简要给出你的理由。注意，明暗度或点数指明密度。如果有帮助的话，假定基于中心即 K 均值，基于邻近性即单链，而基于密度为 DBSCAN。

图 4-19 题 4.7 图

答：

(a) 基于中心的方法有 2 个簇。矩形区域被分成两半，同时 2 个簇里都包含了噪声数据；
基于邻近性的方法有 1 个簇。因为两个圆圈区域受噪声数据影响而形成一个簇；
基于密度的方法有 2 个簇，每个圆圈区域代表一个簇，而噪声数据会被忽略。

(b) 基于中心的方法有 1 个簇，该簇包含图中的一个圆环和一个圆盘；
基于邻近性的方法有 2 个簇，外部圆环代表一个簇，内层圆盘代表一个簇；
基于密度的方法有 2 个簇，外部圆环代表一个簇，内层圆盘代表一个簇。

(c) 基于中心的方法有 3 个簇，每个三角形代表一个簇；
基于邻近性的方法有 1 个簇，三个三角形区域会联合起来因为彼此相互接触；
基于密度的方法有 3 个簇，每个三角形区域代表一个簇。即使三个三角形相互接触，但是所接触的区域的密度比三角形内的密度小。

(d) 基于中心的方法有 2 个簇。两组线被分到两个簇里；
基于邻近性的方法有 5 个簇。相互缠绕的线被分到一个簇中；
基于密度的方法有 2 个簇。这两组线定义了被低密度区域所分割的两个高密度的区域。

4.8 传统的凝聚层次聚类过程每步合并两个簇。这样的方法能够正确地捕获数据点集的(嵌套的)簇结构吗？如果不能，解释如何对结果进行后处理，以得到簇结构更正确的视图。

答:传统的凝聚层次聚类过程关键是每步合并两个最相似的簇，直至只剩下一个簇才停止聚类。该聚类方法并不能产生嵌套的簇结构。但我们可以采用树结构的方法来捕捉形成的层次结构，每次合并都记录父子簇之间的关系，最后形成一个清晰的树状层次簇结构。

4.9 我们可以将一个数据集表示成对象节点的集合和属性节点的集合，其中每个对象与每个属性之间有一条边，该边的权值是对象在该属性上的值。对于稀疏数据，如果权值为 0，则忽略该边。双划分聚类(Bipartite)试图将该图划分成不相交的簇，其中每个簇由一个对象节点集和一个属性节点集组成。目标是最大化簇中对象节点和属性节点之间的边的权值，并且最小化不同簇的对象节点和属性节点之间的边的权值。这种聚类称作协同聚类(co-clustering)，因为对象和属性之间同时聚类。

(a) 双划分聚类(协同聚类)与对象和属性集分别聚类有何不同？
(b) 是否存在某些情况，这些方法产生相同的结果？
(c) 与一般聚类相比，协同聚类的优点和缺点是什么？

答:

(a) 对于普通聚类，只有一组约束条件被运用，要么与对象相关，要么与属性相关。对于协同聚类，两组约束条件同时被运用。因此，单独地划分对象和属性无法产生相同的结果。

(b) 是的。例如，如果一组属性只与一个特定的簇对象相关，如在所有其他簇中的对象权值为 0；相反地，这组对象在一个簇中对所有其他属性的权值为 0，那么由协同聚类发现的簇会与分别通过对象和属性聚类的结果一样。以文本聚类作为例子，有一组文本组成的簇仅包含了某部分的词或短语，相反地，某些词或短语也仅出现在部分文本里。

(c) 协同聚类的优点是能自动产生由属性描述的簇，这种描述信息带有更强的表达能力，在

现实应用中，较由对象描述的簇更有用(如在客户细分中，有属性描述的簇能为营销战略决策提供更为直接的辅助作用)。但具有强区分能力的属性有时候出现严重的重叠现象，这种情况下协同聚类产生的结果并不理想，如出现重叠率很高的聚类结果。

4.10 下表中列出了 4 个点的两个最近邻。使用 SNN 相似度定义，计算每对点之间的 SNN 相似度。

点	第一个近邻	第二个近邻
1	4	3
2	3	4
3	4	2
4	3	1

答：SNN 即共享最近邻个数为其相似度。

点 1 和点 2 的 SNN 相似度：0（没有共享最近邻）

点 1 和点 3 的 SNN 相似度：1（共享点 4 这个最近邻）

点 1 和点 4 的 SNN 相似度：1（共享点 3 这个最近邻）

点 2 和点 3 的 SNN 相似度：1（共享点 4 这个最近邻）

点 2 和点 4 的 SNN 相似度：1（共享点 3 这个最近邻）

点 3 和点 4 的 SNN 相似度：0（没有共享最近邻）

4.11 对于 SNN 相似度定义，SNN 距离的计算没有考虑两个最近邻表中共享近邻的位置。换言之，可能希望基于以相同或粗略相同的次序共享最近邻的两个点以更高的相似度。

(a) 描述如何修改 SNN 相似度定义，基于以粗略相同的次序共享近邻的两个点以更高的相似度；

(b) 讨论这种修改的优点和缺点。

答：(a)

i) 把两个最近邻表中共享近邻按照距离从小达大排列，然后以一个最近邻表为基准，调整另一个最近邻表中近邻的位置，总共调整步数为 r，其中共享的最近邻个数为 n，则相似度可以简单定义为 $\frac{n}{r+\varepsilon}$ 或 $\log\left(\frac{n}{r+\varepsilon}\right)$，其中 ε 为平滑参数，可以简单设置为 $\varepsilon=1$。

ii) 找出两个最近邻表中的最长子序列，最长子序列长度为 L, 共享的最近邻个数为 N，则相似度为 $N\times L$。

(b) 以上相似度计算方法更好地体现了共享近邻在最近邻表中位置的因素对相似度评价的影响，增大了略同次序共享近邻的相似度；缺点是计算复杂度增大。

4.12 一种稀疏化邻近度矩阵的方法如下：对于每个对象，除对应于对象的 k-最近邻的项之外，所有的项都设置为 0。然而，稀疏化之后的邻近度矩阵一般是不对称的。

(a) 如果对象 a 在对象 b 的 k-最近邻中，为什么不能保证 b 在对象 a 的 k-最近邻中？

(b) 至少建议两种方法，可以用来使稀疏化的矩阵是对称的。

答：(a) 可能出现对象 a 的密度较大而对象 b 的密度小的情况。另一种典型情况是：在一个只包含 k+1 个正常对象和一个异常对象的数据区域中，异常对象 b 的 k 个最近邻为 k+1 个正常对象中距离其最近的 k 个对象，但正常对象 a 的 k 个最近邻不可能包含异常对象（这是因为异常对象 b 本来就是距离正常对象最远的点，因此，它与正常对象 a 的距离会大于任何其它正常对象与 a 之间的距离）

(b)

i) 在两个对象 a 和 b 中，只要其中一个对象在另一个对象的最近另列表中，我们就设置 $M_{ba} = M_{ab} = 1$（M_{ab} 是指邻近度矩阵中 a 行和 b 列交叉项的值）；

ii) 当某个对象 a 不在另一对象 b 的 k 最近邻中时，不管另一对象 b 是否在该对象 a 的最近邻中，我们设置 $M_{ba} = M_{ab} = 0$。

4.13 给出一个簇集合的例子，其中基于簇的接近性的合并得到的簇集合比基于簇的连接强度(互连性)的合并得到的簇集合更自然。

答：簇集合例子如下图所示：

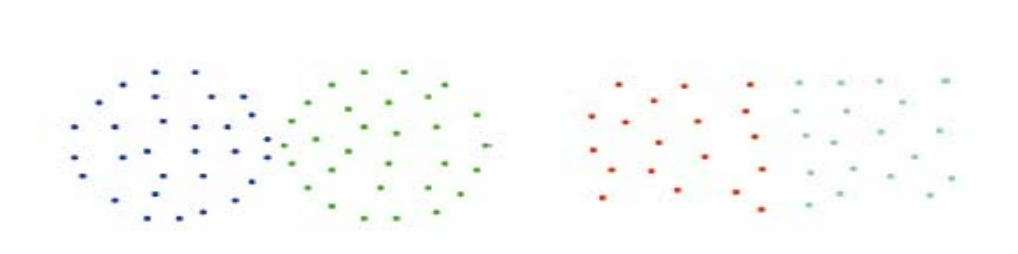


图 (a)　　图 (b)

图(a)两个簇合并后的接近性显然较图(b)要差，因此，如果是基于簇的接近性来合并图中的簇，则会合并图(b)的两个簇，合并得到的簇结构几乎与原来的每个簇一样。

第 5 章关联分析

5.1 列举关联规则在不同领域中应用的实例。

答：在医学领域：发现某些症状与某种疾病之间的关联，为医生进行疾病诊断和治疗提供线索；

在商业领域：发现商品间的联系，为商场进行商品促销及摆放货架提供辅助决策信息；

在地球科学领域：揭示海洋、陆地和大气过程之间的关系。

5.2 给出如下几种类型的关联规则的例子，并说明它们是否是有价值的。

(a)高支持度和高置信度的规则；

(b)高支持度和低置信度的规则；

(c)低支持度和低置信度的规则；

(d)低支持度和高置信度的规则。

答：(a)如牛奶 -> 面包，由于这个规则很明显，所以不具有价值。

(b)如牛奶->大米，由于牛奶、大米销售量都比较高，所以有高支持度。但是很多事务不同时包括牛奶和大米，所以置信度很低，不具有价值。

(c)如可乐->洗衣粉，由于置信度低，所以不具有价值。

(d)如尿布->啤酒，虽然支持度低，不过置信度高，具有价值。

5.3 数据集如表 5-14 所示：

表 5-14 习题 5.3 数据集

Customer ID	Transaction ID	Items Bought
1	0001	{a, d, e}
1	0024	{a, b, c, e}
2	0012	{a, b, d, e}

2	0031	{a, c, d, e}
3	0015	{b, c, e}
3	0022	{b, d, e}
4	0029	{c, d}
4	0040	{a, b, c}
5	0033	{a, d, e}
5	0038	{a, b, e}

(a) 把每一个事务作为一个购物篮，计算项集{e}, {b, d}和{b, d, e}的支持度。

(b) 利用(a)中结果计算关联规则{b, d}→{e} 和 {e}→{b, d}的置信度。置信度是一个对称的度量吗？

(c) 把每一个用户购买的所有商品作为一个购物篮，计算项集{e}, {b, d}和{b, d, e}的支持度。

(d) 利用(b)中结果计算关联规则{b, d}→{e} 和 {e}→{b, d}的置信度。置信度是一个对称的度量吗？

答：(a) s({e}) = 8/10 =0.8 ;

s({b,d}) = 2/10 = 0.2 ;

s({b,d,e}) = 2/10 = 0.2.

(b) c({b,d}->{e}) = s({b,d,e})/s({b,d}) = 0.2/0.2 = 1;

c({e}->{b,d}) =s({b,d,e})/s({e}) = 0.2/0.8 = 0.25.

由于 c({b,d}->{e})≠c({e}->{b,d})，所以置信度不是一个对称的度量。

(c) 如果把每一个用户购买所有的所有商品作为一个购物篮，则

s({e}) = 4/5 =0.8 ;

s({b,d}) = 5/5 = 1 ;

s({b,d,e}) = 4/5 = 0.8.

(d) 利用 c 中结果计算关联规则{b, d}→{e} 和 {e}→{b, d}的置信度,则

c({b,d}->{e}) = 0.8/1 = 0.8

c({e}->{b,d}) = 0.8/0.8 = 1

置信度不是一个对称的度量

5.4 关联规则是否满足传递性和对称性的性质？举例说明。

答：关联规则不满足传递性和对称性！

例如：s(A,B) = 50%　　s(A) = 70%

s(A,C) = 20%　　s(B) = 90%

s(B,C) = 70%　　s(C) = 60%

设最小置信度 minconf = 60%，则:

c(A → B) = s(A,B) / s(A) =71% > minconf

c(B → C) = s(B,C) / s(B)=66% > minconf

但是 c(A → C) = s(A,C) / S(A)=28% < minconf，不满足传递性

c(B → A)= s(A,B) / s(B)=55% < minconf， 不满足对称性

5.5 Apriori 算法使用先验性质剪枝，试讨论如下类似的性质

(a) 证明频繁项集的所有非空子集也是频繁的

(b) 证明项集 s 的任何非空子集 s ’的支持度不小于 s 的支持度

(c) 给定频繁项集 l 和它的子集 s，证明规则“ s’→(l – s’)”的置信度不高于 s→(l – s)的置信度，其中 s’是 s 的子集

(d) Apriori 算法的一个变形是采用划分方法将数据集 D 中的事务分为 n 个不相交的子数据集。证明 D 中的任何一个频繁项集至少在 D 的某一个子数据集中是频繁的。

证明：(a)设 s 为频繁项集，s’为 s 的子集，min_supp_count 为最小支持度计数。由于包含 s 的事务也一定包含 s’，所以 support_ count(s’) ≥support_count(s)≥ min_support_count， s’也是频繁的。

(b)设数据集为 D，|D|为数据集中的事务数。由于 support_ count(s’) ≥support_count(s)，所以 support_count(s’)/|D| ≥support_count(s)/|D|，即 support (s’) ≥support (s)。

(c) 规则“ s→(l – s’)”的置信度 confidence(s→(l – s)) = support(l)/support(s)，规则“ s’→(l – s’)”的置信度 confidence(s’→(l – s’)) = support(l)/support(s’)。

由于 support (s’) ≥support (s)，故“ s’→(l – s’)”的置信度不高于 s→(l – s)的置信度。

(d) 反证法证明。

设 min_support 为最小支持度。D 划分为 $d_1\ d_2 \ldots d_n$ 个子数据集，包含的事务数分别为 $a_1\ a_2 \ldots a_n$。如果 D 中的某一个频繁项集 s 在 D 的所有子数据集中是非频繁的，在每个子数据集中包含 s 的事务数为 $c_1\ c_2 \ldots c_n$，则

$c_1 \le a_1 * min\_support$，$c_2 \le a_2 * min\_support$，…，

$c_n / \le a_n * min\_support$。$(c_1+c_2+\ldots+c_n) \le (a_1+a_2+\cdots a_n) * min\_support$。

由于$(c_1+c_2+\ldots+c_n)$为数据集 D 中包含 s 的事务数，$a_1+a_2+\cdots a_n$为数据集 D 的事务数，所以 s 是非频繁的，与 s 在 D 中是频繁的矛盾。命题得证。

5.6 考虑如下的频繁 3-项集：{1, 2, 3}，{1, 2, 4}，{1, 2, 5}，{1, 3, 4}，{1, 3, 5}，{2, 3, 4}，{2, 3, 5}，{3, 4, 5}。

(a)根据 Apriori 算法的候选项集生成方法，写出利用频繁 3-项集生成的所有候选 4-项集。

(b)写出经过剪枝后的所有候选 4-项集

答：(a)利用频繁 3-项集生成的所有候选 4-项集：

{1,2,3,4}　{1,2,3,5}　{1,2,4,5}　{1,3,4,5}　{2,3,4,5}

(b)经过剪枝后的所有候选 4-项集：

{1,2,3,4}　{1,2,3,5}

5.7 一个数据库有 5 个事务，如表 5-15 所示。设 min_sup=60%，min_conf = 80%。

表 5-15 习题 5.7 数据集

事务 ID	购买的商品
T100	{M, O, N, K, E, Y}
T200	{D, O, N, K, E, Y}
T300	{M, A, K, E}
T400	{M, U, C, K, Y}
T500	{C, O, O, K, I ,E}

(a) 分别用 Apriori 算法和 FP-growth 算法找出所有频繁项集。比较两种挖掘方法的效率。

(b) 比较穷举法和 Apriori 算法生成的候选项集的数量。

(c) 利用(1)所找出的频繁项集，生成所有的强关联规则和对应的支持度和置信度。

答：(1) 频繁 1-项集：M,O,K,E,Y

频繁 2-项集：{M,O}， {O,K}， {O,E}，{K,Y}，{K,E}

频繁 3-项集：{O,K,E}

(2)穷举法：

$M=2^k-1=2^{11}-1=2047$

Apriori 算法：23

(3) {O,K}—>{E}，支持度 0.6，置信度 1

{O,E}—>{k}，支持度 0.6，置信度 1

5.8 购物篮分析只针对所有属性为二元布尔类型的数据集。如果数据集中的某个属性为连续型变量时，说明如何利用离散化的方法将连续属性转换为二元布尔属性。比较不同的离散方法对购物篮分析的影响。

答： 首先利用等频、等宽等方法将连续属性离散化，然后将离散化后的每个区间映射为一个二元属性。

离散化时，如果区间太宽，可能因为缺乏置信度而失去某些模式；如果区间太窄，则可能因为缺乏支持度而失去某些模式。

5.9 分别说明利用支持度、置信度和提升度评价关联规则的优缺点。

答：支持度

优点：支持度高说明这条规则可能适用于数据集中的大部分事务。

缺点：若支持度阈值过高，则许多潜在的有意义的模式由于包含支持度小的项而被删去；

若支持度阈值过低，则计算代价很高而且产生大量的关联模式。

置信度

优点：置信度高说明如果满足了关联规则的前件，同时满足后件的可能性也非常大。

缺点：找到负相关的关联规则。

提升度：

优点：提升度可以评估项集 A 的出现是否能够促进项集 B 的出现

缺点：会产生出现伪相互独立的规则。

5.10 表 5-16 所示的相依表汇总了超级市场的事务数据。其中 hot dogs 指包含热狗的事务，$\overline{\text{hot dogs}}$ 指不包含热狗的事务。hamburgers 指包含汉堡的事务，$\overline{\text{hamburgers}}$ 指不包含汉堡的事务。

表 5-16 习题 5.10 相依表

	hot dogs	$\overline{\text{hot dogs}}$	Σrow
Hamburgers	2,000	500	2,500
$\overline{\text{hamburgers}}$	1,000	1,500	2,500
Σcol	3,000	2,000	5,000

假设挖掘出的关联规则是“ hot dogs ⇒ hamburgers”。给定最小支持度阈值 25%和最小置信度阈值 50%，这个关联规则是强规则吗？

计算关联规则“ hot dogs ⇒ hamburgers”的提升度，能够说明什么问题？购买热狗和购买汉堡是独立的吗？如果不是，两者间存在哪种相关关系？

答：s({hot dogs})=3000/5000=60%; s({hot dogs, hamburgers})=2000/5000=40%

C({hot dogs}→{hamburgers})=40%/60%=66.7%

故这个关联规则是强规则。

S({hamburgers})=2500/5000=50%

提升度 lift({hot dogs}→{hamburgers})= C({hot dogs}→{hamburgers})/ S({hamburgers}) =1.334 提升度大于 1，表明 hot dogs 和 hamburgers 不是互相独立的，二者之间存在正相关关系。

5.11 对于表 5-17 所示序列数据集，设最小支持度计数为 2，请找出所有的频繁模式。

表 5-17 习题 5.11 数据集

Sequence ID	Sequence ID
1	<a(abc)(ac)d(c f)>
2	<(ad)c(bc)(ae)>
3	<(e f)(ab)(d f)cb>
4	<eg(a f)cbc>

答：频繁 1-序列：

<a>、<b>、<c>、<d>、<e>、<f>、<(ab)>、<(bc)>

频繁 2-序列：

<aa>、<ab>、<ac>、<ad>、<af>、<a(bc)>

<ba>、<bc>、<bc>、<bf>

<ca>、<cb>、<cc>

<db>、<dc>

<ea>、<eb>、<ec>、<ef>

<fb>、<fc>

<(bc)a>

频繁 3-序列：

<aba>、<abc>、<aca>、<acb>、<acc>、<adc>、<a(bc)a>

<bdc>

<dcb>

<eab>、<eac>

<fbc>、<fcb>

频繁 4-序列：

<eacb>

第 6 章离群点挖掘

6.1 为什么离群点挖掘是重要的？

答:离群点是指与大部分其它对象不同的对象，在数据的散布图中，它们远离其它数据点，其属性值显著地偏离期望的或常见的属性值。(1) 因为离群点可能是度量或执行错误所导致的，例如相对少的离群点可能扭曲一组值的均值和标准差，或者改变聚类算法产生的簇的集合。(2) 因为离群点本身可能是非常重要的，隐藏着重要的信息，在欺诈检测，入侵检测等方面有着广泛的应用。所以离群点挖掘是非常重要的。

6.2 讨论基于如下方法的离群点检测方法潜在的时间复杂度：使用基于聚类的、基于距离的和基于密度的方法。不需要专门技术知识，而是关注每种方法的基本计算需求，如计算每个对象的密度的时间需求。

答：如果使用 K-means 算法，它的时间复杂度就是 $O(n)$，一般基于邻近度和基于密度的算法的时间复杂度都是 $O(n^2)$，但是对于低维数据，使用专门的数据结构，如树或者 k-d

树，可以把基于邻近度的算法的时间复杂度降低到$O(n\log n)$，而对基于密度的算法来说，如果使用基于网格的算法，则可以把时间复杂度降低到$O(n)$，但这种方法不太精确而且也是用于低维数据。

6.3 许多用于离群点检测的统计检验方法是在这样一种环境下开发的：数百个观测就是一个大数据集。我们考虑这种方法的局限性：

(a) 如果一个值与平均值的距离超过标准差的三倍，则检测称它为离群点。对于 1000000 个值的集合，根据该检验，有离群点的可能性有多大？(假定正态分布)；

(b) 一种方法称离群点是具有不寻常低概率的对象。处理大型数据集时，该方法需要调整吗？如果需要，如何调整？

答：(a)如果指的是单面的点的距离超过标准差的 3 倍，那么概率就是 0.00135，则有 1350 个离群点；如果指的是两面的点的距离超过标准差的 3 倍，那么概率就是 0.0027，则有 2700 个离群点。

(b)具有百万个对象的数据集中，有成千上万个离群点，我们可以接受它们作为离群点或者降低临界值用以减少离群点。

6.4 假定正常对象被分类为离群点的概率是 0.01，而离群点被分类为离群点概率为 0.99，如果 99%的对象都是正常的，那么假警告率或误报率和检测率各为多少？(使用下面的定义)

$$\text{检测率} = \frac{\text{检测出的离群点个数}}{\text{离群点的总数}}$$

$$\text{假警告率} = \frac{\text{假离群点的个数}}{\text{被分类为离群点的个数}}$$

答： 假警告率=(99%*1%)/(99%*1%+1%*99%)=50%

检测率=(1%*99%)/(1%)=99%

6.5 从包含大量不同文档的集合中选择一组文档，使得它们尽可能彼此相异。如果我们认为相互之间不高度相关(相连接、相似)的文档是离群点，那么我们选择的所有文档可能都被分类为离群点。一个数据集仅由离群对象组成可能吗？或者，这是误用术语吗？

答：离群点暗含的意思是稀有的、不常见的，有很多离群点的定义在一定的程度上融合了这个概念。然而,在一些情况下，离群点通常不会普遍发生，举一个相关例子：网络故障，但有一个具体的定义。这就使得它能够区分这两种情况：纯粹检测一个异常和所要处理的对象大多数都是异常。同时，如果异常的概念是由数学或由算法定义的，这些定义可能会导致这样的一种情况：所研究的数据集中大部分或所有的对象都被归类为异常。另一种观点则可能认为如果不能够定义一种有意义的正常的情形，那么所有的对象都是异常。(“独特“这一术语通常也是用于这种情况。)总的来说，这可以被看作是哲学问题或语义问题。一个好的定义(尽管不可能是没有争议的)是能够分辨出当所收集的对象大多数或全部都是异常这一种情况。

6.6 考虑一个点集，其中大部分点在低密度区域，少量点在高密度区域。如果我们定义离群点为低密度区域的点，则大部分点被划分为离群点。这是对基于密度的离群点定义的适当使用吗？是否需要用某种方式修改该定义？

答：如果密度有一个绝对意义，比如被指定到某一定义域内，那么它可能会非常合理的考虑

把大部分的点作为异常。然而，在很多情况下，为了能够准确使用异常检测技术，通常会考虑使用相对密度这一概念。

6.7 一个数据分析者使用一种离群点检测算法发现了一个离群子集。出于好奇，该分析者对这个离群子集使用离群点检测算法。

(a) 讨论本章介绍的每种离群点检测技术的行为。(如果可能，使用实际数据和算法来做)；

(b) 当用于离群点对象的集合时，你认为离群点检测算法将做何反应？

答：(a)在某些情况下，以统计学为基础的异常检测技术，在离群子集上使用这将是无效的使用技术，因为这种检测方法的假设将不再成立。对于那些依赖于模型的方法也是如此。以邻近点为基础或者以密度为基础的方法主要取决于特定的技术。如果保留原来的参数，使用距离或密度的绝对阈值的方法会将异常归类为一个异常对象的集合。其他相关方法会将大部分异常归类为普通点或者将一部分归类为异常。

(b)一个对象是否异常取决于整个对象的集合。因此，期望一种异常检测技术能够辨别一个异常集合，就像原始集合中并不存在这样一个异常集合，这是不合理的。