<数据挖掘>2015-2016 真题

一. 简答题(每个 10 分，共 40 分)

1. 什么是关联规则的支持度？什么是关联规则的可信度？

答：(1)给定关联规则 X=>Y，支持度指项集 X 和 Y 在数据库 D 中同时出现的概率，即 Pr(XUY)。

(2)给定关联规则 X=>Y，可信度指项集 X 出现的情况下，项集 Y 在数据库 D 中同时出现的条件概率，即 Pr(X|Y)= Pr(XUY)/Pr(X)。

2. 数据挖掘的特点有哪些？

答：(1) **数据挖掘的数据必须是真实的**。数据挖掘所处理的数据通常是已经存在的真实数据，而不是为了进行数据分析而专门收集的数据。因此，数据收集本身不属于数据挖掘所关注的焦点，这是数据挖掘区别于大多数统计任务的特征之一。

(2) **数据挖掘所处理的数据必须是海量的**。如果数据集很小的话，采用单纯的统计分析方 法就可以了。但是，当数据集很大时，会面临许多新的问题，诸如数据的有效存储、快速访问、合理表示等。

(3) **查询一般是决策制定者（用户）提出的随机查询**。查询要求灵活，往往不能形成精确的查询要求，要靠数据挖掘技术来寻找可能的查询结果。

(4) **挖掘出来的知识一般是不能预知的，数据挖掘发现的是潜在的、新颖的知识**。这些知识在特定环境下是可以接受、可以理解、可以运用的，但不是放之四海皆准的。

3. 数据挖掘组件化思想包含哪几部分？

答：**数据挖掘算法都是由 5 个"标准组件"构成的，即模型或模式结构、数据挖掘任务、评分函致、搜索和优化方法、数据管理策略**。

每一种组件都蕴含着一些非常通用的系统原理，掌握了每一种组件的基本原理之后，再来理解由不同组件“装配”起来的算法就变得相对轻松一些。而且，不同算法之间的比较也变得更加容易，因为能从组件这个层面看出算法之间的异同。

1)通过数据挖掘过程所得到的知识通常被称为模型（model）或模式(pattern)。模型是全局的，模式是局部的。

2)根据数据分析者的目标，可以将数据挖掘任务分为：模式挖掘，模型挖掘(描述建模,预测建模)

3)评分函数用来对数据集与模型（模式）的拟合程度进行评估。常用的评分函数有：似然（likelihood）函数、误差平方和、准确率等。

4)搜索和优化的目标是确定模型（模式）的结构及其参数值，以使评分函数达到最小值（或最大值）。

5)数据管理策略应该设计有效的数据组织和索引技术，或者通过采样、近似等手段，来减少数据的扫描次数，从而提高数据挖掘算法的效率。

确定模型（模式）结构和评分函数的过程通常由人来完成,而优化评分函数的过程通常需要计算机辅助来实现。

4. 为什么说 naive Bayesian 分类法是 naive 的？

答：朴素贝叶斯分类之所以称之为“朴素”的,是因为在分类的计算过程中做了一个朴素的假设,假定属性值之间是相互独立的。

该假设称作类条件独立，做此假设的目的是为了简化计算。

二. 现有如下事物数据库，设 min_sup=60%, min_conf=80%

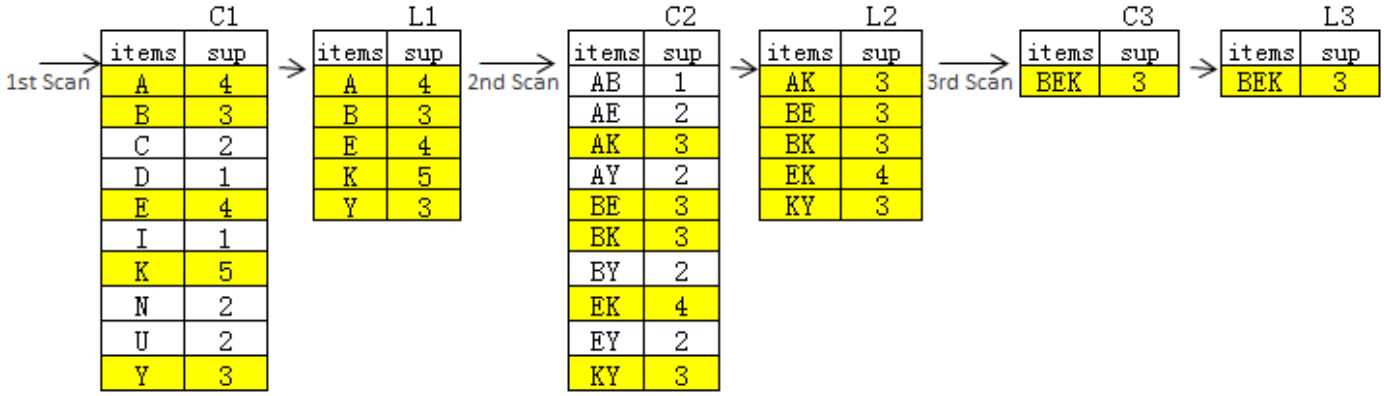
（1）请用 Apriori 算法找出所有的频繁项目集。（12 分）

（2）请写出所有的关联规则（8 分）

TID	items bought
T100	{A, B, N, K, E, Y}
T200	{D, B, N, K, E, Y}
T300	{A, A, K, E}
T400	{A, U, C, K, Y}
T500	{C, B, U, K, I, E}

答：1)求频繁项集

Min_sup=60%，频度=3



所有频繁项集：1-项集（L1）：{A},{B},{E},{K},{Y}

2-项集（L2）：{A,K},{B,E},{B,K},{E,K},{K,Y}

3-项集（L3）：{B,E,K}

2)求关联规则

min_conf=80%=0.8，则

L1	没关联规则，因为左右没东西。	
L2	A=>K	AK/A=3/4=0.75
	K=>A	AK/K=3/5=0.6
	B=>E	BE/B=3/3=1
	E=>B	BE/E=3/4=0.75
	B=>K	BK/B=3/3=1
	K=>B	BK/K=3/5=0.6
	E=>K	EK/E=4/4=1
	K=>E	EK/K=4/5=0.8
	K=>Y	KY/K=3/5=0.6
	Y=>K	KY/Y=3/3=1
L3	B=>EK	BEK/B=3/3=1
	E=>BK	BEK/E=3/4=0.75
	K=>BE	BEK/K=3/5=0.6
	BE=>K	BEK/BE=3/3=1
	BK=>E	BEK/BK=3/3=1
	EK=>B	BEK/EK=3/4=0.75

输出所有 confidence>0.8 的规则有：B=>E,B=>K,E=>K,Y=>K,B=>EK,BE=>K,BK=>E

三. 假设数据挖掘的任务是将如下的 8 个点（用(x, y)代表位置）聚类为三个类。

A1(1,9), A2(2,8), A3(5,4), B1(6,8), B2(10,5), B3(6,8), C1(10,2), C2(4,8), C3(2,6)

距离函数是曼哈顿函数。假设初始我们选择 A1, B1 和 C1 为每个聚类的中心，

请用 k-means 算法给出在第一次循环执行后的三个聚类中心。（20 分）

答：曼哈顿距离： d=|x1-x2|+|y1-y2|

族			C1	C2	C3
族中心点			1,9	6,8	10,2
A1	1	9	0	6	16
A2	2	8	2	4	14
A3	5	4	9	5	7
B1	6	8	6	0	10
B2	10	5	13	7	3
B3	6	8	6	0	10
C1	10	2	16	10	0
C2	4	8	4	2	12
C3	2	6	4	6	12

第一族中心	(1+2+2)/3	(9+8+6)/3
	2	8
第二族中心	(5+6+6+4)/4	(4+8+8+8)/4
	5	7
第三族中心	(10+10)/2	(5+2)/2
	10	4

族			C1	C2	C3
族中心点			2,8	5,7	10,4
A1	1	9	2	6	14
A2	2	8	0	4	12
A3	5	4	7	3	5
B1	6	8	4	2	8
B2	10	5	11	7	1
B3	6	8	4	2	8
C1	10	2	14	10	2
C2	4	8	2	2	10
C3	2	6	2	4	10

第一族中心	(1+2+4+2)/4	(9+8+8+6)/4
	2	8
第二族中心	(5+6+6)/3	(4+8+8)/3
	6	7
第三族中心	(10+10)/2	(5+2)/2
	10	4

族			C1	C2	C3
族中心点			2,8	6,7	10,4
A1	1	9	2	7	14
A2	2	8	0	5	12
A3	5	4	7	4	5
B1	6	8	4	1	8
B2	10	5	11	6	1
B3	6	8	4	1	8
C1	10	2	14	9	2
C2	4	8	2	3	10
C3	2	6	2	5	10

第一次循环后三个族聚类中心分别为（2，8）、（5，7）、（10、4）

第二次循环后三个族聚类中心分别为（2，8）、（6，7）、（10、4）（注：这部分以下题目没要求可以不写）

第三次循环后三个族聚类中心分别为（2，8）、（6，7）、（10、4）

因为第二次和第三次结果一样，直接就收敛了，结束了。

四. 请给出决策树分类法的方法和步骤。（20 分）

答：决策树的生成是一个从根节点开始，从上到下的递归过程。一般采用分而治之的方法，通过不断地将训练样本划分成子集来构造决策树。

假设给定的训练集 T 总共有 m 个类别。则针对 T 构造决策树时，会出现以下三种情况：

(1) 如果 T 中所有样本的类别相同，那么决策树只有一个叶子结点。

(2) 如果 T 中没有可用于继续分裂的变量，则将 T 中出现频率最高的类别作为当前结点的类别。

(3) 如果 T 包含的样本属于不同的类别，根据变量选择策略，选择最佳的变量和划分方式将 T 分为几个子集 T1, T2, …, Tk，每个数据子集构成一个内部结点。

对于某个内部结点继续进行判断，重复上述操作，直到满足决策树的终止条件为止。终止条件就是，结点对应的所有样本属于同一个类别，或者 T 中没有可用于进一步分裂的变量。

决策树构建算法 Generate_decision_tree。

输入：训练集 T，输入变量集 A，目标（类别）变量 Y

输出：决策树 Tree

Generate_decision_tree(T,A,Y)

1:如果 T 为空，返回出错信息；

2:如果 T 的所有样本都属于同一个类别 C，则用 C 标识当前节点并返回；

3:如果没有可分的变量，则用 T 中出现频率最高的类别标识当前结点并返回；

4:根据变量选择策略选择最佳变量 X 将 T 分为 k 个子集（T1，T2，…Tk）；

5:用 X 标识当前结点；

6:对 T 的每一个子集 Ti

7:NewNode= Generate_decision_tree(Ti,A-X,Y)； //递归操作

8:生成一个分枝，该分枝由结点 X 指向 NewNode；

9:返回当前结点。

在上述算法中，结点分裂（第 4 步）是生成决策树的重要步骤。只有根据不同的变量将单个结点分裂成多个结点，方能形成多个类别，因此整个问题的核心就是如何选择分裂变量。