<数据挖掘>2016-2017 真题

（注意：卷面成绩总共 80 分，另有作业 20 分）

一. 简答题(每个 4 分，共 20 分)

1. 为什么说朴素贝叶斯分类法是朴素的？

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。

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，例如，广泛使用的评分函数有：似然、误差平方和、准确率等。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。

4. 请对决策树分类法中建树的过程进行简单描述。

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。

5. DBSCAN 算法是一种非常重要的基于密度的聚类方法。请指出它的优缺点。

答：DBSCAN 方法的优点是：

(1) 不需要事先确定族的个数；

(2) 聚类速度快，使用索引（例如 R*tree）时，DBSCAN 的时间复杂度为 *O(nlogn)*，n 为数据库中数据对象的个数，否则，DBSCAN 的时间复杂度为 O（n^2）；

(3) 对噪声数据不敏感；

(4) 能发现任意形状的族，例如，DBSCAN 可以找出如图 10.4 所示的族。

DBSCAN 的缺点是：

(1) 输入参数ε和 MinPts 的值较难确定。

(2) 当数据库中数据对象的密度分布不均匀时，用相同的参数值可能得不到好的聚类结果。

(3) 可能会产生”链条”现象。如图 10.10 所示，左边的上、下两个本应独立的族连接在了一起，产生了类似”链条”的现象。

(4) 使用 R*-tree 索引时，由于 R*-tree 在高维空间中不够有效，导致 DBSCAN 算法在处理高维数据时性能下降。

二. 现有如下事物数据库，设 min sup=70%, min conf=75%

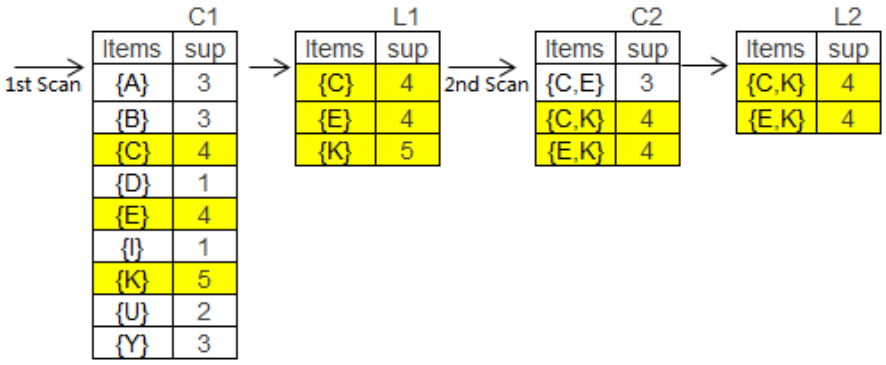
（1）请用 Apriori 算法找出所有的频繁项目集。（12 分）

（2）请写出所有的关联规则（8 分）

TID	items bought
T100	{A, B, C, K, E, Y}
T200	{D, B, C, K, E, Y}
T300	{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=70%，频度=3.5



所有频繁项集：1-项集（L1）：{C},{E},{K}

2-项集（L2）：{C,K},{E,K}

2)求关联规则

min_conf=75%=0.75，则

L1	没关联规则，因为左右没东西。	
L2	C=>K	CK/C=4/4=1
	K=>C	CK/K=4/5=0.8
	E=>K	EK/E=4/4=1
	K=>E	EK/K=4/5=0.8

输出所有 confidence>0.75 的规则有：C=>K，K=>C，E=>K，K=>E

三. 假设数据挖掘的任务是将如下的十个点（用(x, y)代表位置）聚类为三个类。

A1(1,9), A2(3,8), A3(5,4), A4(9,2), B1(6,8), B2(10,5), B3(3,8), C1(10,2), C2(4,8), C3(2,6) , C4(7,3)

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和聚类结果。（20 分）

答：1、第一次迭代：1）选择三个初始中心：A1, B1 和 C1

2）计算 11*3 个距离

		A1(1,9)	A2(3,8)	A3(5,4)	A4(9,2)	B1(6,8)	B2(10,5)	B3(3,8)	C1(10,2)	C2(4,8)	C3(2,6)	C4(7,3)
C1	**A1(1,9)**	0	3	9	15	6	13	3	16	4	4	12
C2	**B1(6,8)**	6	3	5	9	0	7	3	10	2	6	6
C3	**C1(10,2)**	16	13	7	1	10	3	13	0	12	12	4

注：找曼哈顿距离最近的那个点就是新中心点，d(i,j)=|X1-X2|+|Y1-Y2|。**如 A1(1,9)到 A2(3,8)的** d=|1-3|+|9-8|=3 ,红色为距中心点最近的

3）将每一个对象分配给离自己最近的 cluster（中心），第一次循环结果：

Cluster1：**A1,A2,B3,C3**

Cluster2：**A3,B1,C2**

Cluster3：**A4,B2,C1,C4**

2、第二次迭代：1）算出每个 cluster 的新中心

Cluster1 的中心：$\lfloor(1+3+3+2)/4，(9+8+8+6)/4\rfloor$ =（2，7）

Cluster2 的中心：$\lfloor(5+6+4)/3，(4+8+8)/3\rfloor$ =（5，6）

Cluster3 的中心：$\lfloor(9+10+10+7)/4，(2+5+2+3)/4\rfloor$ =（9，3）

（注：采用向下取整的方法）

2）计算 11*3 个距离

		A1(1,9)	A2(3,8)	A3(5,4)	A4(9,2)	B1(6,8)	B2(10,5)	B3(3,8)	C1(10,2)	C2(4,8)	C3(2,6)	C4(7,3)
C1	**A1(2,7)**	2	1	7	13	4	11	1	14	2	2	10
C2	**B1(5,6)**	7	4	2	8	3	6	4	9	3	3	5
C3	**C1(9,3)**	14	11	5	1	8	3	11	2	10	10	2

3）重新分组，第二次循环结果：

Cluster1：**A1,A2,B3, C2,C3**

Cluster2：**A3,B1**

Cluster3：**A4,B2,C1,C4**

2、第三次迭代：1）算出每个 cluster 的新中心

Cluster1 的中心：$\lfloor(1+3+3+4+2)/5，(9+8+8+8+6)/5\rfloor$ =（2，7）

Cluster2 的中心：$\lfloor(5+6)/2，(4+8)/2\rfloor$ =（5，6）

Cluster3 的中心：$\lfloor(9+10+10+7)/4，(2+5+2+3)/4\rfloor$ =（9，3）

因为第三次和第二次新中心一样，故分组结果也一样，直接就收敛了，结束了。

四. 在下面的雇员数据表中，数据已经被离散化，如“31. . .35”代表年龄在 31 到 35 之间的人。Count表示该行在数据集重复出现的次数。假设Status是类标签属性，给定新元组“Systems, 31. . .35, 41k. . .45k”，请问如果用朴素贝叶斯分类法的话，该元组的类标签应该是什么？请写出计算步骤（20 分）

department	*status*	*age*	*salary*	*count*
sales	senior	31. . .35	46k. . .50k	30
sales	junior	26. . .30	26k. . .30k	40
sales	junior	31. . .35	31k. . .35k	40
systems	junior	21. . .25	46k. . .50k	20
systems	senior	31. . .35	66k. . .70k	5
systems	junior	26. . .30	46k. . .50k	3
systems	senior	41. . .45	66k. . .70k	3
marketing	senior	36. . .40	46k. . .50k	10
marketing	junior	31. . .35	41k. . .45k	4
secretary	senior	46. . .50	36k. . .40k	4
secretary	junior	26. . .30	26k. . .30k	6

答：

Class: C1: status=senior

C2: status =junior

输入向量 X =(**Systems, 31. . .35, 41k. . .45k**)

每个类别的 P(Ci) ，i=1,2 。 根据训练样本计算 P(Senior)= $\frac{52}{165}$ P(Junior)= $\frac{113}{165}$

为了计算 P(X|Ci)，i=1,2 需要先计算下列条件概率：

P(Systems|Senior) = $\frac{8}{52}$ P(31…35|Senior) = $\frac{35}{52}$ P(41k…45k|Senior)= $\frac{0}{52}$

P(Systems|Junior) = $\frac{23}{113}$ P(31…35|Junior) = $\frac{44}{113}$ P(41k…45k|Junior)= $\frac{4}{113}$

根据贝叶斯公式：（1）假设 status=senior 时，P（senior|X）= $\frac{P(X|\text{senior})P(\text{senior})}{P(X)}$

（2）假设 status =junior 时，P（junior|X）= $\frac{P(X|\text{junior})P(\text{junior})}{P(X)}$

利用上表作为训练数据，使用朴素贝叶斯分类法（属性直接的条件独立假设），为每一个类别算一个概率

P（Senior|X∝ P(X|Senior)P(Senior)= P(Systems|Senior) P(31…35|Senior) P(41k…45k|Senior) P(Senior)

$=\frac{8}{52}*\frac{35}{52}*\frac{0}{52}*\frac{52}{165}=0$

P（Junior|X∝ P(X|Junior)P(Junior)= P(Systems|Junior) P(31…35|Junior) P(41k…45k|Junior) P(Junior)

$=\frac{23}{113}*\frac{44}{113}*\frac{4}{113}*\frac{113}{165}=0.002$

由于 salary 在训练集中没有出现结果为 0，所以采用平滑方法：

P（xj|ci）$\frac{(\text{count(xj,ci))+mp}}{(\text{count(ci)+m}}=\frac{(\text{count(xj,ci))+1}}{(\text{count(ci)+2}}$ 其中 m=|C|= 2 , p=1/|C|=1/2

那么，

P（Senior|X∝ P(X|Senior)P(Senior)= P(Systems|Senior) P(31…35|Senior) P(41k…45k|Senior) P(Senior)

$=\frac{8+1}{52+2}*\frac{35+1}{52+2}*\frac{0+1}{52+2}*\frac{52}{165}=0.0006$

P（Junior|X∝ P(X|Junior)P(Junior)= P(Systems|Junior) P(31…35|Junior) P(41k…45k|Junior) P(Junior)

$=\frac{23+1}{113+2}*\frac{44+1}{113+2}*\frac{4+1}{113+2}*\frac{113}{165}=0.0024$

P（Senior|X）< P（Junior|X），故该新元组类别属于 C2：status= Junior