《数据仓库与数据挖掘》期末练习

一、考虑表1中的数据集

表1 样本数据集

Instance	A	B	C	Class
1	0	0	1	−
2	1	0	1	+
3	0	1	0	−
4	1	0	0	−
5	1	0	1	+
6	0	0	1	+
7	1	1	0	−
8	0	0	0	−
9	0	1	0	+
10	1	1	1	+

(a) 估计条件概率P(A=1|+)，P(B=1|+)，P(C=1|+)，P(A=1|-)，P(B=1|-)和P(C=1|-)。
(b) 根据上述条件概率，使用朴素贝叶斯方法预测样本(A=1, B=1,C=1)的类标号。
(c) 比较P(A=1)，P(B=1)和 P(A=1,B=1)，描述A、B之间的关系。
(d) 使用m估计方法（p=1/3且m=3）估计条件概率，与(a)相比哪种方法更健壮，为什么?

二、给定如图1所示的格结构和事务表，用如下标记直接在图上标识出每一个结点，
假定初始支持度阈值为30%，当候选项长度达到3时，支持度阈值递减为20%，后续依次按10%递减。

M：如果结点是极大频繁项集。

C：如果结点是闭频繁项集。

N：如果结点是频繁的，但既不是极大的也不是闭的。

I：如果结点是非频繁的。

Transaction ID	Items Bought
1	$\{a, b, d, e\}$
2	$\{b, c, d\}$
3	$\{a, b, d, e\}$
4	$\{a, c, d, e\}$
5	$\{b, c, d, e\}$
6	$\{b, d, e\}$
7	$\{c, d\}$
8	$\{a, b, c\}$
9	$\{a, d, e\}$
10	$\{b, d\}$

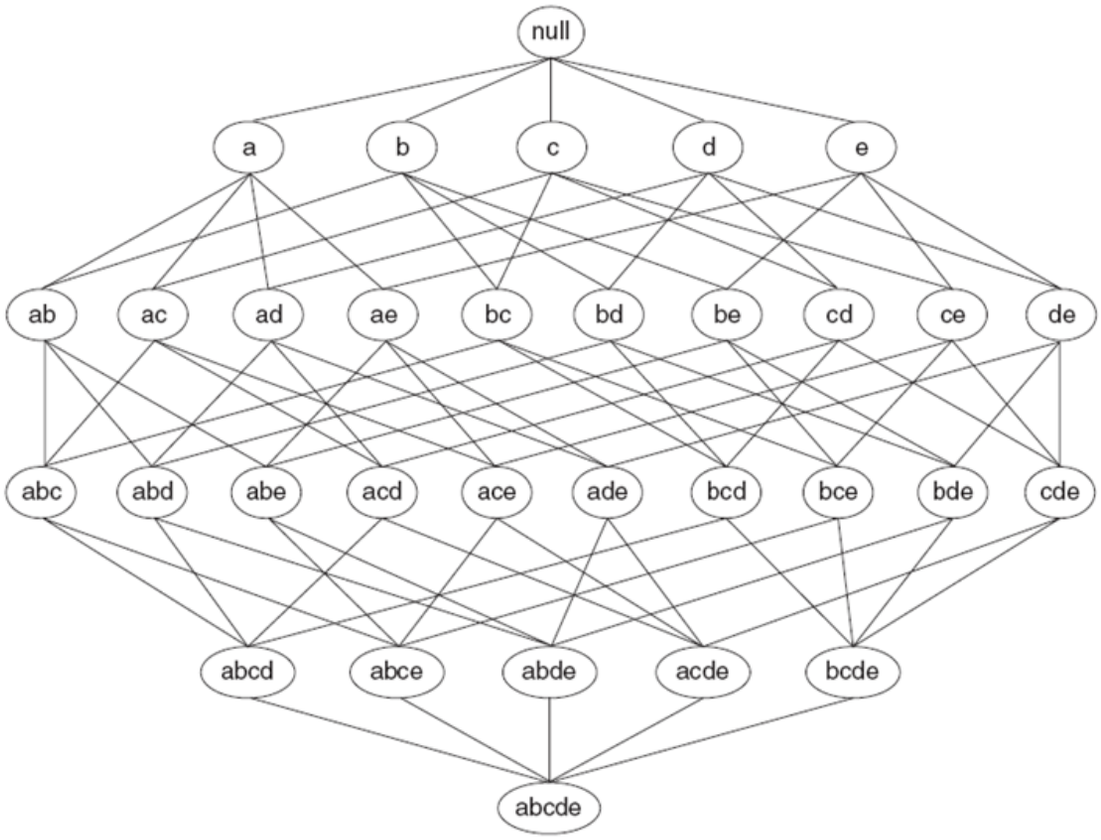

图1格结构和事务表

三、说明支持度－置信度框架存在的问题，试给出解决的方法。

四、请用k-均值算法把表2中S的8个样本数据聚为3个簇，并给出每个簇的平均值点，假设初始迭代时选择X_1、X_4和X_7作为初始簇中心点。

表2 样本数据集S

数据点	属性1	属性2	数据点	属性1	属性2
X_1	2	10	X_5	7	5
X_2	2	5	X_6	6	4
X_3	8	4	X_7	1	2
X_4	5	8	X_8	4	9

五、决策树判断

节点的选择有哪些策略（即节点的选择）？根据下表数据，使用ID3算法构建决策树预测模型（需有完整的过程描述）。

OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play

sunny	72	95	false	Don’t Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don’t Play

六、分别描述K-means与K-medoids算法，并针对两种算法的异同点以及优劣进行分析比较。