1. 考虑下表中的数据集

(a)估计条件概率P(A|+)、P(B|+)、P(C|+)、P(A| −)、P(B|−)和 P(C|−)。

(b)根据(a)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)的类标签。

样本	A	B	C	类标号
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	+

(a) $P(A=1|+)=\frac{3}{5}, P(B=1|+)=\frac{1}{5}, P(C=1|+)=\frac{4}{5},$

$P(A=1|-)=\frac{2}{5}, P(B=1|-)=\frac{2}{5}, P(C=1|-)=\frac{5}{5}=1$

$P(A=0|+)=\frac{2}{5}, P(B=0|+)=\frac{4}{5}, P(C=0|+)=\frac{1}{5},$

$P(A=0|-)=\frac{3}{5}, P(B=0|-)=\frac{3}{5}, P(C=0|-)=0$

(b) 由朴素贝叶斯假设可知

$$P(+|A=0,B=1,C=0)=\frac{P(A=0,B=1,C=0|+)P(+)}{P(A=0,B=1,C=0)}$$

$$=\frac{P(A=0|+)\cdot P(B=1|+)\cdot P(C=0|+)\cdot P(+)}{P(A=0,B=1,C=0)}$$

$$=\frac{2}{5}\cdot\frac{1}{5}\cdot\frac{1}{5}\cdot\frac{5}{10}\cdot\frac{1}{P(A=0,B=1,C=0)}=\frac{1}{125}\cdot\frac{1}{P(A=0,B=1,C=0)}$$

$$P(-|A=0,B=1,C=0)=\frac{P(A=0,B=1,C=0|-)P(-)}{P(A=0,B=1,C=0)}$$

$$=\frac{P(A=0|-)\cdot P(B=1|-)\cdot P(C=0|-)\cdot P(-)}{P(A=0,B=1,C=0)}$$

$$=\frac{3}{5}\cdot\frac{1}{5}\cdot 0\cdot\frac{5}{10}\cdot\frac{1}{P(A=0,B=1,C=0)}=0$$

所以可知$A=0,B=1,C=0$的类标签应该是+

2. 考虑下表中显示的数据集。

(a)将每个事务 ID 视为一个购物篮，计算项集{e}，{b，d}和{b，d，e}的支持度

(b)使用(a)的计算结果，计算关联规则{b，d}→{e}和{e}→{b，d}的置信度。置信度是对称的度量吗?

(c)将每个顾客 ID 作为一个购物篮，重复(a)。应当将每个项看作一个二元变量(如果一个项在顾客的购买事务中至少出现了一次，则为 1;否则为 0)。

(d)使用(c)的计算结果，计算关联规则{b，d}→{e}和{e}→{b，d}的置信度。

顾客 ID	事务 ID	购买项
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) $\sigma(\{e\})=8,\sigma(\{b,d\})=2,\sigma(\{b,d,e\})=2$

$S(\{e\})=\frac{4}{5},S(\{b,d\})=\frac{1}{5},S(\{b,d,e\})=\frac{1}{5}$

(b) $c(\{b,d\}\rightarrow\{e\})=\frac{\sigma(\{b,d,e\})}{\sigma(\{b,d\})}=1,\ c(\{e\}\rightarrow\{b,d\})=\frac{\sigma(\{b,d,e\})}{\sigma(\{e\})}=\frac{1}{4}$

可以发现二者并不相等，因此不具有对称性，一般情况下对换后前件发生改变，导致分母大概率发生改变，此时分子不变，所以不具有对称性

(c) 按照顾客划分后表格变为

顾客 ID	购买项
1	{a，b，d，c，e}
2	{a，b，c，d，e}
3	{b，c，d，e}
4	{a，b，c，d}
5	{a，b，d，e}

$\sigma(\{e\})=4,\sigma(\{b,d\})=5,\sigma(\{b,d,e\})=4$

$S(\{e\})=\frac{4}{5},S(\{b,d\})=1,S(\{b,d,e\})=\frac{4}{5}$

(d)$c(\{b,d\}\rightarrow\{e\})=\frac{\sigma(\{b,d,e\})}{\sigma(\{b,d\})}=\frac{4}{5},\ c(\{e\}\rightarrow\{b,d\})=\frac{\sigma(\{b,d,e\})}{\sigma(\{e\})}=1$

3. 考虑下表中的一维数据集。

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

(b)根据距离权衡每个最近邻x_i的影响$w_i = \frac{1}{|x_i - x|}$，使用距离加权表决方法重复前面的分析。

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−

(a)

	1-最近邻	3-最近邻	5-最近邻	9-最近邻
类别	+	-	+	-

(b)

1-最近邻：取$x = 4.9$，故为+样本

3-最近邻：$w = \frac{1}{|5-4.9|} - \frac{1}{|5.2-5|} - \frac{1}{|5.3-5|} = 1.67 > 0$，所以预测为+

5-最近邻：$w = \frac{1}{|5-4.9|} - \frac{1}{|5.2-5|} - \frac{1}{|5.3-5|} + \frac{1}{|4.6-5|} + \frac{1}{|5-5.5|} = 6.17 > 0$，所以预测为+

9-最近邻：$w = \frac{1}{|5-4.9|} - \frac{1}{|5.2-5|} - \frac{1}{|5.3-5|} + \frac{1}{|4.6-5|} + \frac{1}{|5-5.5|} - \frac{1}{|5-0.5|} - \frac{1}{|5-3|} + \frac{1}{|5-4.5|} - \frac{1}{|5-7|} = 6.94 > 0$，所以预测为+

4. 传统的关联规则挖掘方法使用支持度和置信度度量来剪枝没有兴趣的规则

(a)使用下表中的事务数据，绘制出下面每个规则对应的列联表。规则:{b} → {c}，{a} → {d}，{b} → {d}，{e} → {c}，{c} → {a}。

(b)利用(a)的列联表，按照下面的度量计算。要给出具体的计算过程。

i.支持度

ii.置信度

iii 提升度

事务 ID	购买项
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}

(a) 列联表如下所示

$\{b\} \rightarrow \{d\}$

	d	$\bar{d}$	
b	6	1	7
b	3	0	3
	9	1	10

$\{b\} \rightarrow \{c\}$

	c	$\bar{c}$	
b	3	4	7
$\bar{b}$	2	1	3
	5	5	10

$\{e\} \rightarrow \{c\}$

	c	$\bar{c}$	
e	2	4	6
$\bar{e}$	3	1	4
	5	5	10

$\{c\} \rightarrow \{a\}$

	a	$\bar{a}$	
c	2	3	5
$\bar{c}$	3	2	5
	5	5	10

$\{a\} \rightarrow \{d\}$

	d	$\bar{d}$	
a	4	1	5
$\bar{a}$	5	0	1
	9	1	10

(b)

i:支持度

$$S(\{b\} \rightarrow \{c\}) = \frac{\sigma(\{b,c\})}{N} = \frac{3}{10},\ S(\{a\} \rightarrow \{d\}) = \frac{\sigma(\{a,d\})}{N} = \frac{2}{5}, S(\{b\} \rightarrow \{d\}) = \frac{\sigma(\{b,d\})}{N} = \frac{3}{5},$$

$$S(\{e\} \rightarrow \{c\}) = \frac{\sigma(\{e,c\})}{N} = \frac{1}{5}, S(\{c\} \rightarrow \{a\}) = \frac{\sigma(\{a,c\})}{N} = \frac{1}{5}$$

ii:置信度

$$\mathrm{C}(\{b\} \rightarrow \{c\}) = \frac{\sigma(\{b,c\})}{\sigma(\{b\})} = \frac{3}{7},\ \mathrm{C}(\{a\} \rightarrow \{d\}) = \frac{\sigma(\{a,d\})}{\sigma(\{a\})} = \frac{4}{5}, \mathrm{C}(\{b\} \rightarrow \{d\}) = \frac{\sigma(\{b,d\})}{\sigma(\{b\})} = \frac{6}{7},$$

$$\mathrm{C}(\{e\} \rightarrow \{c\}) = \frac{\sigma(\{e,c\})}{\sigma(\{e\})} = \frac{1}{3}, \mathrm{C}(\{c\} \rightarrow \{a\}) = \frac{\sigma(\{a,c\})}{\sigma(\{c\})} = \frac{2}{5}$$

iii:提升度

$$I(A,B) = \frac{S(A,B)}{S(A) \times S(B)} = \frac{N \times f_{11}}{f_{1+} \times f_{+1}}$$

$$I(\{b\},\{c\}) = \frac{S(\{b\},\{c\})}{S(\{b\}) \times S(\{c\})} = \frac{6}{7}, I(\{a\},\{d\}) = \frac{S(\{a\},\{d\})}{S(\{a\}) \times S(\{d\})} = \frac{8}{9},$$

$$I(\{b\},\{d\}) = \frac{S(\{b\},\{d\})}{S(\{b\}) \times S(\{d\})} = \frac{20}{21}, I(\{e\},\{c\}) = \frac{S(\{e\},\{c\})}{S(\{e\}) \times S(\{c\})} = \frac{2}{3}$$

$$I(\{a\},\{c\}) = \frac{S(\{a\},\{c\})}{S(\{a\}) \times S(\{c\})} = \frac{4}{5}$$

5. 一维点的集合是:{6，12，18，24，30，42，48}，执行 K 均值算法

(a)对于下列每组初始质心，将每个点指派到最近的质心，创建两个簇，然后对两个簇的每组质心分别计算总平方误差。对每组质心，给出这两个簇和总平方误差

i.(18，45)

ii.(15，40)

(b)两组质心代表稳定解吗，即如果在该数据集上，使用给定的质心作为初始质心运行 K 均值，所产生的簇会有改变吗?

(a)

i. 簇 1：6，12，18，24，30　　簇 2：42，48

总平方误差$SSE = (18-6)^2 + (18-12)^2 + (18-18)^2 + (18-24)^2 + (30-18)^2 + (46-42)^2 + (48-42)^2 = 378$

ii. 簇 1：6，12，18，24　　簇 2：30，42，48

总平方误差$SSE = (15-6)^2 + (15-12)^2 + (15-18)^2 + (15-24)^2 + (30-40)^2 + (40-42)^2 + (48-40)^2 = 348$

(b)

对于 i，簇 1 的质心为 18，簇 2 的质心为 45

对于 ii，簇 1 的质心为 15，簇 2 的质心为 40

可以发现对于这两种情况质心都没有发生改变，所以此后经过重新分配也不会再发生改变，故两组质心代表稳定解