1.1 考虑下表中的事务性数据集：

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	$\{a, b, e\}$

1) 每个事务 ID 对应一条事务，计算{e}，{b, c}，{b, c, e}的支持度。
2) 使用（1）的计算结果，计算关联规则{b, c}→{e}和{e}→{b, c}的置信度。
3) 从（2）的结果看，置信度是对称的吗？请根据计算公式分析其对称性。

1) $s(\{e\}) = \frac{8}{10} = \frac{4}{5}$

$s(\{b,c\}) = \frac{3}{10}$

$s(\{b,c,e\}) = \frac{2}{10} = \frac{1}{5}$

2) $s(\{b,c\} \to \{e\}) = \frac{2}{3}$

$s(\{e\} \to \{b,c\}) = \frac{2}{8} = \frac{1}{4}$

3) 置信度不是对称的.

置信度的计算公式为

$c(X \to Y) = \frac{\sigma(X \cup Y)}{\sigma(X)}$　　$c(Y \to X) = \frac{\sigma(X \cup Y)}{\sigma(Y)}$

由于 $\sigma(X)$ 和 $\sigma(Y)$ 可能不同，故不对称.

?0等是对称

1.2 考虑如下二元分类的数据集：（设为 D）

User interest	User occupation	Click
Tech	Professional	1
Fashion	Student	0
Fashion	Professional	0
Sports	Student	0
Tech	Student	1
Tech	Retired	0
Sports	Professional	1

1) 计算分别以属性 User interest 和 User occupation 划分时的信息增益。构建决策树将会选择哪个属性？
2) 计算分别以属性 User interest 和 User occupation 划分时的 Gini 指数。构建决策树将会选择哪个属性？

1) 整体熵：（根据是否点击）

$$Ent(D) = -\sum_{k=1}^{2} p_k \log_2 p_k = -\left(\frac{3}{7}\log_2\frac{3}{7} + \frac{4}{7}\log_2\frac{4}{7}\right) \simeq 0.985$$

以 User interest 为特征进行划分

D^1(User interest = Tech)　D^2(User interest = Fashion)

D^3(User interest) = Sports)

$Ent(D^1) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \simeq 0.918$

$Ent(D^2) = -(1\times\log_2 1) = 0$

$Ent(D^3) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$

$$\therefore Gain(D, \text{User interest}) = Ent(D) - \sum_{v=1}^{3}\frac{|D^v|}{|D|}Ent(D^v)$$

$$= 0.985 - \left(\frac{3}{7}\times 0.918 + \frac{2}{7}\times 1\right)$$

$$\simeq 0.306$$

同理 以 User occupation（简写为 U_o）为特征进行划分.

D^1(U_o = Professional)　D^2(U_o = Student)　D^3(U_o = Retired)

$Ent(D^1) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \simeq 0.918$

$Ent(D^2) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \simeq 0.918$

$Ent(D^3) = -(1\log_2 1) = 0$

$\therefore Gain(D, U\_o) = 0.985 - \left(\frac{3}{7}\times 0.918 + \frac{3}{7}\times 0.918\right) \simeq 0.198$

前者的信息增益更大，故应选 User interest.

2) $Gini(D) = 1 - \sum_{k=1}^{K} p_k^2$

属性 a 的基尼指数定义为 $Gini\_index(D, a) = \sum_{v=1}^{V}\frac{|D^v|}{|D|}Gini(D^v)$.

以 User interest 为特征进行划分

$Gini(D^1) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = \frac{4}{9}$

$Gini(D^2) = 1 - 1^2 = 0$

$Gini(D^3) = 1 - \left[\left(\frac{1}{2}\right)^2 + \left(\frac{1}{2}\right)^2\right] = \frac{1}{2}$

故 $Gini\_index(D, \text{User interest}) = \frac{3}{7}\times\frac{4}{9} + \frac{2}{7}\times\frac{1}{2} = \frac{4}{21} + \frac{1}{7} = \frac{7}{21} = \frac{1}{3}$

以 User occupation 为特征进行划分

$Gini(D^1) = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{2}{3}\right)^2\right] = \frac{4}{9}$

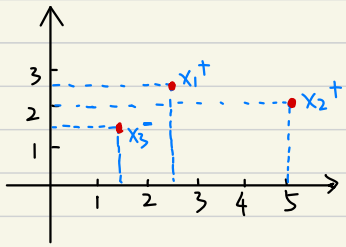
$Gini(D^2) = 1 - \left[\left(\frac{2}{3}\right)^2 + \left(\frac{1}{3}\right)^2\right] = \frac{4}{9}$

$Gini(D^3) = 1 - 1^2 = 0$

故 $Gini\_index(D, \text{User occupation}) = \frac{3}{7}\times\frac{4}{9} + \frac{3}{7}\times\frac{4}{9} = \frac{8}{21}$

前者的 Gini 指数更小，故应选 User interest.

1.3 已知正例点 $x_1 = (2.5, 2.5)^T$（+1），$x_2 = (5, 2)^T$（+1），和负例点 $x_3 = (1.5, 1.5)^T$（-1），试用 SVM 对其进行分类，求最大间隔分离超平面，并指出所有的支持向量。



设超平面方程：$\vec{w}^T\vec{x} + b = 0$

引入 Lagrange 乘子 α_i，$i=1,2,3$，得到 Lagrange 函数

$L(\vec{w}, b, \vec{\alpha}) = \frac{1}{2}\|\vec{w}\|^2 - \sum_{i=1}^{3}\alpha_i(y_i(\vec{w}^T\vec{x_i} + b) - 1)$

令 $L(\vec{w}, b, \vec{\alpha})$ 关于 $\vec{w}$ 和 b 的偏导为 0，得：

$\vec{w} = \sum_{i=1}^{3}\alpha_i y_i x_i \qquad \sum_{i=1}^{3}\alpha_i y_i = 0$

回代，即求解

$\min \frac{1}{2}\sum_{i=1}^{3}\sum_{j=1}^{3}\alpha_i\alpha_j y_i y_j (\vec{x_i}^T \cdot \vec{x_j}) - \sum_{i=1}^{3}\alpha_i$

约束条件 $\begin{cases} \alpha_1 + \alpha_2 - \alpha_3 = 0 \\ \alpha_i \geq 0 \quad i=1,2,3 \end{cases}$

对上式展开

$\frac{1}{2}(12.5\alpha_1^2 + 29\alpha_2^2 + 4.5\alpha_3^2 + 17.5\alpha_1\alpha_2 - 7.5\alpha_1\alpha_3 - 10.5\alpha_2\alpha_3) - \alpha_1 - \alpha_2 - \alpha_3$

将 $\alpha_3 = \alpha_1 + \alpha_2$ 代入上式，得

$\frac{1}{2}(12.5\alpha_1^2 + 29\alpha_2^2 + 4.5(\alpha_1+\alpha_2)^2 + 17.5\alpha_1\alpha_2 - 7.5\alpha_1(\alpha_1+\alpha_2) - 10.5\alpha_2(\alpha_1+\alpha_2))$
$-2\alpha_1 - 2\alpha_2$

$= \frac{1}{2}(12.5\alpha_1^2 + 29\alpha_2^2 + 4.5\alpha_1^2 + 9\alpha_1\alpha_2 + 4.5\alpha_2^2 + 17.5\alpha_1\alpha_2 - 7.5\alpha_1^2 - 7.5\alpha_1\alpha_2$
$-10.5\alpha_1\alpha_2 - 10.5\alpha_2^2) - 2\alpha_1 - 2\alpha_2$

$= \frac{1}{2}(9.5\alpha_1^2 + 23\alpha_2^2 + 9.5\alpha_1\alpha_2) - 2\alpha_1 - 2\alpha_2$

$= 4.75\alpha_1^2 + 11.5\alpha_2^2 + 4.25\alpha_1\alpha_2 - 2\alpha_1 - 2\alpha_2 \triangleq A$

分别对 α_1，α_2 求偏导，并使其为 0

$\frac{\partial A}{\partial \alpha_1} = 9.5\alpha_1 + 4.25\alpha_2 - 2 = 0$

$\frac{\partial A}{\partial \alpha_2} = 23\alpha_2 + 4.25\alpha_1 - 2 = 0$

$\Rightarrow \begin{cases} \alpha_1 = \frac{200}{1069} \approx 0.187 \\ \alpha_2 = \frac{56}{1069} \approx 0.052 \end{cases}$

$\alpha_3 = \alpha_1 + \alpha_2 = 0.239 = \frac{256}{1069}$

（若不满足则令其中一个为 0）

$\therefore \vec{w} = \frac{200}{1069} \times 1 \times (2.5, 2.5)^T + \frac{56}{1069} \times 1 \times (5, 2)^T + \frac{256}{1069} \times (-1) \times (1.5, 1.5)^T$

$= \left(\frac{396}{1069}, \frac{228}{1069}\right) \approx (0.370, 0.213)^T$

$b = 1 - (2_1 x$

x_1、x_2、x_3 对应的 λ 均不为0，故这3个都为支持向量
（$\lambda = 0$，对最终结果无影响）