1. Bài tập 1: Thuật toán Apriori và Luật Kết hợp (Association Rules)

Dữ liệu

Giao dịch	Các mặt hàng
t_1	Beef, Chicken, Milk
t_2	Beef, Cheese
t_3	Cheese, Boots
t_4	Beef, Chicken, Cheese
t_5	Beef, Chicken, Clothes, Cheese, Milk
t_6	Chicken, Clothes, Milk
t_7	Chicken, Milk, Clothes

Thông số

Tổng số giao dịch $N = 7$, minsup= 30% $\Rightarrow$ ngưỡng tối thiểu $\lceil 7 \times 0.3 \rceil = 3$, minconf= 80%.

Bước 1: Tập mục phổ biến 1-itemset (L_1)

Tập mục I = {Beef, Chicken, Milk, Cheese, Boots, Clothes}.

Itemset (C_1)	Count	Support	Frequent?
{Beef}	4	$\frac{4}{7} \approx 57.1\%$	Yes
{Chicken}	5	$\frac{5}{7} \approx 71.4\%$	Yes
{Milk}	4	$\frac{4}{7} \approx 57.1\%$	Yes
{Cheese}	4	$\frac{4}{7} \approx 57.1\%$	Yes
{Boots}	1	$\frac{1}{7} \approx 14.3\%$	No
{Clothes}	3	$\frac{3}{7} \approx 42.9\%$	Yes

$$L_1 = \{\{Beef\}, \{Chicken\}, \{Milk\}, \{Cheese\}, \{Clothes\}\}$$

Bước 2: Tập mục ứng viên 2-itemset (C_2) và phổ biến 2-itemset (L_2)

Các tập ứng viên:

$$\{Beef, Chicken\}, \{Beef, Milk\}, \{Beef, Cheese\}, \{Beef, Clothes\}, \{Chicken, Milk\}, \{Chicken, Cheese\}, \{Chicken, Clothes$$

Itemset (C_2)	Count	Support	Frequent?
{Beef, Chicken}	3	$\frac{3}{7} \approx 42.9\%$	Yes
{Beef, Milk}	2	$\frac{2}{7} \approx 28.6\%$	No
{Beef, Cheese}	3	$\frac{3}{7} \approx 42.9\%$	Yes
{Beef, Clothes}	1	$\frac{1}{7} \approx 14.3\%$	No
{Chicken, Milk}	4	$\frac{4}{7} \approx 57.1\%$	Yes
{Chicken, Cheese}	2	$\frac{2}{7} \approx 28.6\%$	No
{Chicken, Clothes}	3	$\frac{3}{7} \approx 42.9\%$	Yes
{Milk, Cheese}	1	$\frac{1}{7} \approx 14.3\%$	No
{Milk, Clothes}	3	$\frac{3}{7} \approx 42.9\%$	Yes
{Cheese, Clothes}	1	$\frac{1}{7} \approx 14.3\%$	No

$$L_2 = \{\{Beef, Chicken\}, \{Beef, Cheese\}, \{Chicken, Milk\}, \{Chicken, Clothes\}, \{Milk, Clothes\}\}$$

Bước 3: Tập mục ứng viên 3-itemset (C_3) và phổ biến 3-itemset (L_3)

Tạo C_3 từ L_2:

- Từ $\{Beef, Chicken\}$ và $\{Beef, Cheese\}$ tạo $\{Beef, Chicken, Cheese\}$, nhưng loại vì $\{Chicken, Cheese\} \notin L_2$.

- Từ $\{Chicken, Milk\}$ và $\{Chicken, Clothes\}$ tạo $\{Chicken, Milk, Clothes\}$, giữ lại vì tất cả các tập con 2-itemset đều phổ biến.

$$C_3 = \{\{Chicken, Milk, Clothes\}\}$$

Itemset (C_3)	Count	Support	Frequent?
{Chicken, Milk, Clothes}	3	$\frac{3}{7} \approx 42.9\%$	Yes

$$L_3 = \{\{Chicken, Milk, Clothes\}\}$$

Bước 4: Không có tập 4-itemset, thuật toán dừng

Bước 5: Tạo luật kết hợp từ L_3

Tập mục phổ biến $X = \{Chicken, Milk, Clothes\}$, $sup(X) = \frac{3}{7}$.

Luật có dạng $A \Rightarrow X \setminus A$ với confidence $\geq 80\%$.

Luật	$sup(A)$	Confidence	$\geq 80\%$	Kết quả
$\{Chicken, Milk\} \Rightarrow \{Clothes\}$	$\frac{4}{7}$	$\frac{3/7}{4/7} = 75\%$	No	Loại
$\{Chicken, Clothes\} \Rightarrow \{Milk\}$	$\frac{3}{7}$	$\frac{3/7}{3/7} = 100\%$	Yes	Chọn
$\{Milk, Clothes\} \Rightarrow \{Chicken\}$	$\frac{3}{7}$	100%	Yes	Chọn
$\{Chicken\} \Rightarrow \{Milk, Clothes\}$	$\frac{5}{7}$	60%	No	Loại
$\{Milk\} \Rightarrow \{Chicken, Clothes\}$	$\frac{4}{7}$	75%	No	Loại
$\{Clothes\} \Rightarrow \{Chicken, Milk\}$	$\frac{3}{7}$	100%	Yes	Chọn

Luật kết hợp được chọn:

$$\begin{cases} \{Chicken, Clothes\} \Rightarrow \{Milk\} \\ \{Milk, Clothes\} \Rightarrow \{Chicken\} \\ \{Clothes\} \Rightarrow \{Chicken, Milk\} \end{cases}$$

2. Bài tập 2: Thuật toán Apriori với Multiple Minimum Supports (MIS)

Dữ liệu

Giao dịch	Các mặt hàng
t_1	Beef, Bread
t_2	Bread, Clothes
t_3	Bread, Clothes, Milk
t_4	Cheese, Boots
t_5	Beef, Bread, Cheese, Shoes
t_6	Beef, Bread, Cheese, Milk
t_7	Bread, Milk, Clothes

Thông số

$$\begin{cases} MIS(Milk) = 50\% \Rightarrow \lceil 7 \times 0.5 \rceil = 4 \\ MIS(Bread) = 70\% \Rightarrow \lceil 7 \times 0.7 \rceil = 5 \\ MIS(\text{others}) = 25\% \Rightarrow \lceil 7 \times 0.25 \rceil = 2 \end{cases}$$

Bước 1: Sắp xếp các mục theo MIS tăng dần

Item	MIS (%)	Ngưỡng Count
Boots	25	2
Shoes	25	2
Beef	25	2
Clothes	25	2
Cheese	25	2
Milk	50	4
Bread	70	5

Thứ tự: $(Boots, Shoes, Beef, Clothes, Cheese, Milk, Bread)$.

Bước 2: Tập mục phổ biến 1-itemset (L_1)

Item	Count	MIS (Count)	Frequent?
Boots	1	2	No
Shoes	1	2	No
Beef	3	2	Yes
Clothes	3	2	Yes
Cheese	3	2	Yes
Milk	3	4	No
Bread	6	5	Yes

$$L_1 = \{\{Beef\}, \{Clothes\}, \{Cheese\}, \{Bread\}\}$$

Bước 3: Tập mục ứng viên 2-itemset (C_2) và phổ biến 2-itemset (L_2)

3.1 Tạo C_2 từ L_1

$$\{Beef, Clothes\}, \{Beef, Cheese\}, \{Beef, Bread\}, \{Clothes, Cheese\}, \{Clothes, Bread\}, \{Cheese, Bread\}$$

3.2 Kiểm tra tần suất

Itemset	Count	MIS (Count)	Frequent?
{Beef, Clothes}	0	2	No
{Beef, Cheese}	2	2	Yes
{Beef, Bread}	3	2	Yes
{Clothes, Cheese}	0	2	No
{Clothes, Bread}	3	2	Yes
{Cheese, Bread}	2	2	Yes

$$L_2 = \{\{Beef, Cheese\}, \{Beef, Bread\}, \{Clothes, Bread\}, \{Cheese, Bread\}\}$$

Bước 4: Tập mục ứng viên 3-itemset (C_3) và phổ biến 3-itemset (L_3)

Chỉ có thể tạo:

$$\{Beef, Cheese, Bread\}$$

Kiểm tra tần suất:

$$Count = 2, \quad MIS = \min(2, 2, 5) = 2 \Rightarrow \text{Frequent}$$

$$L_3 = \{\{Beef, Cheese, Bread\}\}$$

Bước 5: Không thể tạo 4-itemset, thuật toán dừng

—

3. Bài tập 3: Thuật toán C4.5 và Cây quyết định (Information Gain)

Dữ liệu

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

Bước 1: Tính Entropy của tập dữ liệu gốc

$$N = 15, \quad N_{Yes} = 9, \quad N_{No} = 6$$

$$Entropy(S) = -\frac{9}{15}\log_2\frac{9}{15} - \frac{6}{15}\log_2\frac{6}{15} \approx -0.6 \times (-0.737) - 0.4 \times (-1.322) = 0.971$$

Bước 2: Tính Information Gain cho từng thuộc tính

A. Gain(S, Age)

Age	$\lvert S_v\rvert$	N_{Yes}	N_{No}	Entropy
young	5	3	2	$-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \approx 0.971$
middle	5	4	1	≈ 0.722
old	5	3	2	≈ 0.971

$$Gain(S, Age) = 0.971 - \frac{5}{15} \times 0.971 - \frac{5}{15} \times 0.722 - \frac{5}{15} \times 0.971 = 0.971 - 0.888 = 0.083$$

B. Gain(S, Has_job)

Has_job	$\|S_v\|$	N_{Yes}	N_{No}		Entropy
true	4	4	0	0	(tinh khiết)
false	11	5	6		≈ 0.994

$$Gain(S, Has\_job) = 0.971 - \frac{4}{15} \times 0 - \frac{11}{15} \times 0.994 = 0.971 - 0.729 = 0.242$$

C. Gain(S, Own_house)

Own_house	$\|S_v\|$	N_{Yes}	N_{No}	Entropy
true	7	6	1	≈ 0.592
false	8	3	5	≈ 0.954

$$Gain(S, Own\_house) = 0.971 - \frac{7}{15} \times 0.592 - \frac{8}{15} \times 0.954 = 0.971 - 0.785 = 0.186$$

D. Gain(S, Credit_rating)

Credit_rating	$\|S_v\|$	N_{Yes}	N_{No}	Entropy
fair	5	4	1	≈ 0.722
good	5	4	1	≈ 0.722
excellent	5	4	1	≈ 0.722

$$Gain(S, Credit\_rating) = 0.971 - 3 \times \frac{5}{15} \times 0.722 = 0.971 - 0.722 = 0.249$$

Bước 3: Chọn thuộc tính gốc

Thuộc tính **Credit_rating** có Gain cao nhất (0.249), chọn làm nút gốc.

Bước 4: Xây dựng nhánh con

- Credit_rating = excellent: $N = 4$, $N_{Yes} = 4$, $Entropy = 0 \Rightarrow$ Lá Yes.
- Credit_rating = good: $N = 6$, $N_{Yes} = 4$, $N_{No} = 2$, $Entropy \approx 0.918$. Chọn thuộc tính tiếp theo (Age hoặc Own_house).
- Credit_rating = fair: $N = 5$, $N_{Yes} = 1$, $N_{No} = 4$, $Entropy \approx 0.722$. Tiếp tục phân chia.

Bước 5: Luật dựa trên cây

$$\begin{cases} \text{IF } Credit\_rating = excellent \Rightarrow Class = Yes \\ \text{IF } Credit\_rating = good \text{ AND } Age = old \Rightarrow Class = Yes \\ \text{IF } Credit\_rating = fair \text{ AND } Own\_house = true \Rightarrow Class = Yes \\ \text{IF } Credit\_rating = fair \text{ AND } Own\_house = false \text{ AND } Has\_job = false \Rightarrow Class = No \end{cases}$$

—

4. Bài tập 4: Ma trận nhầm lẫn và các chỉ số đánh giá

	Predicted NO	Predicted YES	Total
Actual NO	TN=50	FP=10	60
Actual YES	FN=5	TP=100	105
Total	55	110	165

$$\text{Accuracy} = \frac{TP+TN}{N} = \frac{100+50}{165} \approx 0.909$$
$$\text{Misclassification Rate} = \frac{FP+FN}{N} = \frac{10+5}{165} \approx 0.091$$
$$\text{True Positive Rate (Recall, Sensitivity)} = \frac{TP}{TP+FN} = \frac{100}{105} \approx 0.952$$
$$\text{False Positive Rate} = \frac{FP}{FP+TN} = \frac{10}{60} \approx 0.167$$
$$\text{True Negative Rate (Specificity)} = \frac{TN}{TN+FP} = \frac{50}{60} \approx 0.833$$
$$\text{Precision} = \frac{TP}{TP+FP} = \frac{100}{110} \approx 0.909$$
$$\text{Prevalence} = \frac{TP+FN}{N} = \frac{105}{165} \approx 0.636$$

—

5. Bài tập 5: Decision List và Đánh giá

Dữ liệu

No.	outlook	temperature	humidity	windy	play
1	sunny	85.0	85.0	FALSE	no
2	sunny	80.0	90.0	TRUE	no
3	overcast	83.0	86.0	FALSE	yes
4	rainy	70.0	96.0	FALSE	yes
5	rainy	68.0	80.0	FALSE	yes
6	rainy	65.0	70.0	TRUE	no
7	overcast	64.0	65.0	TRUE	yes
8	sunny	72.0	95.0	FALSE	no
9	sunny	69.0	70.0	FALSE	yes
10	rainy	75.0	80.0	FALSE	yes
11	sunny	75.0	70.0	TRUE	yes
12	overcast	72.0	90.0	TRUE	yes
13	overcast	81.0	75.0	FALSE	yes
14	rainy	71.0	91.0	TRUE	no

Decision List

$$\begin{cases} \text{outlook} = \text{overcast} \Rightarrow \text{yes} \\ \text{windy} = \text{TRUE} \Rightarrow \text{no} \\ \text{outlook} = \text{sunny} \Rightarrow \text{no} \\ \text{otherwise} \Rightarrow \text{yes} \end{cases}$$

Dự đoán và kết quả

No. Kết quả	outlook	windy	Actual	Rule Applied	Predicted
1 TN	sunny	FALSE	no	3	no
2 TN	sunny	TRUE	no	2	no
3 TP	overcast	FALSE	yes	1	yes
4 TP	rainy	FALSE	yes	4	yes
5 TP	rainy	FALSE	yes	4	yes
6 TN	rainy	TRUE	no	2	no
7 TP	overcast	TRUE	yes	1	yes
8 TN	sunny	FALSE	no	3	no
9 FN	sunny	FALSE	yes	3	no
10 TP	rainy	FALSE	yes	4	yes
11 FN	sunny	TRUE	yes	2	no
12 TP	overcast	TRUE	yes	1	yes
13 TP	overcast	FALSE	yes	1	yes
14 TN	rainy	TRUE	no	2	no

Ma trận nhầm lẫn

	Predicted NO	Predicted YES	Total
Actual NO	TN=5	FP=0	5
Actual YES	FN=2	TP=7	9
Total	7	7	14

Các chỉ số đánh giá

$$\text{Accuracy} = \frac{TP + TN}{N} = \frac{7 + 5}{14} = 0.857$$
$$\text{Precision} = \frac{TP}{TP + FP} = \frac{7}{7} = 1.0$$
$$\text{Recall} = \frac{TP}{TP + FN} = \frac{7}{9} \approx 0.778$$
$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{5}{5} = 1.0$$
$$F_1 = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 2 \times \frac{1.0 \times 0.778}{1.0 + 0.778} = 0.875$$

—

6. Bài tập 6: Thuật toán K-Means

Dữ liệu

Subject	A	B
1	1.0	1.0
2	1.5	2.0
3	3.0	4.0
4	5.0	7.0
5	3.5	5.0
6	4.5	5.0
7	3.5	4.5

Khoảng cách Euclidean

Khoảng cách giữa điểm $x = (x_A, x_B)$ và tâm cụm $m = (m_A, m_B)$:

$$d(x, m) = \sqrt{(x_A - m_A)^2 + (x_B - m_B)^2}$$

Vòng lặp 1

Tâm cụm ban đầu:

$$m_1 = (1.0, 1.0), \quad m_2 = (5.0, 7.0)$$

Phân cụm: Gán mỗi điểm vào cụm gần nhất (tính khoảng cách Euclidean).

Subject	$d(x, m_1)$	$d(x, m_2)$	Cluster
1 (1.0,1.0)	0.00	7.21	C_1
2 (1.5,2.0)	1.12	6.10	C_1
3 (3.0,4.0)	3.61	3.61	C_1 (chọn theo quy ước)
4 (5.0,7.0)	7.21	0.00	C_2
5 (3.5,5.0)	4.72	2.50	C_2
6 (4.5,5.0)	5.31	2.06	C_2
7 (3.5,4.5)	4.30	2.92	C_2

Cụm mới:

$$C_1 = \{1, 2, 3\}, \quad C_2 = \{4, 5, 6, 7\}$$

Cập nhật tâm cụm:

$$m_1' = \left(\frac{1.0 + 1.5 + 3.0}{3}, \frac{1.0 + 2.0 + 4.0}{3}\right) = (1.83, 2.33)$$

$$m_2' = \left(\frac{5.0 + 3.5 + 4.5 + 3.5}{4}, \frac{7.0 + 5.0 + 5.0 + 4.5}{4}\right) = (4.125, 5.375)$$

Vòng lặp 2

Phân cụm lại với tâm mới m_1', m_2':

Subject	$d(x, m_1')$	$d(x, m_2')$	Cluster
1	1.57	5.37	C_1
2	0.47	4.27	C_1
3	4.16	1.78	C_2
4	5.64	1.85	C_2
5	2.92	0.73	C_2
6	3.20	0.53	C_2
7	2.45	1.07	C_2

Cụm mới:

$$C_1 = \{1, 2\}, \quad C_2 = \{3, 4, 5, 6, 7\}$$

Cập nhật tâm cụm:

$$m_1'' = \left(\frac{1.0 + 1.5}{2}, \frac{1.0 + 2.0}{2}\right) = (1.25, 1.5)$$

$$m_2'' = \left(\frac{3.0 + 5.0 + 3.5 + 4.5 + 3.5}{5}, \frac{4.0 + 7.0 + 5.0 + 5.0 + 4.5}{5}\right) = (3.9, 5.1)$$

Vòng lặp 3

Phân cụm lại:

Subject	$d(x, m_1'')$	$d(x, m_2'')$	Cluster
1	0.56	4.86	C_1
2	0.56	3.40	C_1
3	2.96	1.49	C_2
4	6.09	2.16	C_2
5	3.51	0.41	C_2
6	3.78	0.61	C_2
7	3.19	0.78	C_2

Cụm không đổi, thuật toán hội tụ.

$$\boxed{\begin{cases} C_1 = \{1, 2\}, \quad m_1 = (1.25, 1.5) \\ C_2 = \{3, 4, 5, 6, 7\}, \quad m_2 = (3.9, 5.1) \end{cases}}$$

—

7. Bài tập 7: Phân cụm Phân cấp (Hierarchical Clustering)

Dữ liệu

ID	X_1	X_2
A	1	1
B	1.5	1.5
C	5	5
D	3	4
E	4	4
F	3	3.5

Bước 0: Ma trận khoảng cách Euclidean

Khoảng cách giữa $X = (x_1, x_2)$ và $Y = (y_1, y_2)$:

$$d(X, Y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

	A	B	C	D	E	F
A	0	0.707	5.657	3.606	4.243	3.041
B	0.707	0	4.950	3.041	3.536	2.404
C	5.657	4.950	0	1.414	1.0	2.062
D	3.606	3.041	1.414	0	1.0	0.5
E	4.243	3.536	1.0	1.0	0	1.118
F	3.041	2.404	2.062	0.5	1.118	0

Bước 1: Ghép cụm gần nhất

Khoảng cách nhỏ nhất là $d(D, F) = 0.5$. Ghép (D, F).

Bước 2: Cập nhật ma trận (Single linkage)

Khoảng cách giữa cụm mới (D, F) và điểm khác là min khoảng cách từng điểm:

$$d(A, (D, F)) = \min(d(A, D), d(A, F)) = \min(3.606, 3.041) = 3.041$$

Tương tự cho các điểm khác.

Bước 3: Ghép cụm tiếp theo

Khoảng cách nhỏ nhất tiếp theo là $d(A, B) = 0.707$. Ghép (A, B).

Bước 4: Cập nhật ma trận

Tính khoảng cách giữa các cụm mới theo single linkage.

Bước 5: Ghép cụm tiếp theo

Khoảng cách nhỏ nhất là $d(C, E) = 1.0$. Ghép (C, E).

Bước 6: Ghép cụm (D, F) và (C, E)

Khoảng cách $d((D, F), (C, E)) = 1.0$.

Bước 7: Ghép cụm cuối cùng

Ghép (A, B) với (D, F, C, E) tại khoảng cách 2.404.

Dendrogram

- (D, F) hợp nhất ở mức 0.5
- (A, B) hợp nhất ở mức 0.707
- (C, E) hợp nhất ở mức 1.0
- (D, F) và (C, E) hợp nhất ở mức 1.0
- (A, B) hợp nhất với cụm lớn còn lại ở mức 2.404