Faculté des Nouvelles Technologies de l'Information et de la Communication
Département : Informatique Fondamentale et ses Applications
Année Universitaire : 2023/2024
Module : DAMI (Master 1 SDIA)

TD N° 5 : Règles d'association – Algorithme Apriori
Mesures de performance de la classification

❖ Exercice 1 - Algorithme Apriori :

Supposons avoir enregistré dans un magasin informatique les achats faits par des clients (ensemble de transactions), selon le tableau suivant :

N°	Désignation
01	CD, DVD, Jeu, Antivirus
02	CD, Jeu, Souris, Caméra, Antivirus
03	Antivirus, FlashDisk, CD, Jeu, Caméra
04	Jeu, DVD, Tablette
05	FlashDisk, Antivirus, Tablette, Souris

1. Trouver tous les ensembles d'items fréquents vérifiant minsup ≥ 60%, en appliquant l'algorithme Apriori.
2. Déduire toutes les règles d'association.
3. Quelles sont les règles ayant une Confiance ≥ 80% ?

Exercice 2 – Mesures de performance de la classification:

Soit un problème de détection de spam dans les e-mails en utilisant un modèle de classification. Le modèle a été évalué sur un ensemble de 20 e-mails, numérotés de 1 à 20, avec les résultats de la prédiction par rapport aux étiquettes réelles présentées dans le tableau suivant :

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Classe réelle	Non	Oui	Non	Oui	Oui	Non	Oui	Non	Non	Oui	Oui	Non	Oui	Oui	Non	Oui	Oui	Non	Non	Oui
Classe prédite	Oui	Oui	Non	Oui	Oui	Non	Non	Non	Oui	Oui	Oui	Non	Oui	Non	Oui	Oui	Non	Non	Oui	Oui

1. Donnez la matrice de confusion.
2. Calculez la précision du modèle.
3. Calculez l'exactitude (accuracy).
4. Calculez le taux d'erreur.

Solution exercice 1 - Algorithme Apriori :

Les ensembles d'items fréquents vérifiant minsup ≥ 60% :

La Base de données Formelle :

A = CD, B = DVD, C = Jeux, D = Antivirus,

E = Souris, F = Cameras, G = FlashDisk,

H = Tablette.

	A	B	C	D	E	F	G	H
1	1	1	1	1	0	0	0	0
2	1	0	1	1	1	1	0	0
3	1	0	1	1	0	1	1	0
4	0	1	1	0	0	0	0	1
5	0	0	0	1	1	0	1	1

Etape i=1 :

C1 = {A, B, C, D, E, F, G, H}

C1	A	B	C	D	E	F	G	H
Sup	3/5	2/5	4/5	4/5	2/5	2/5	2/5	2/5

F1 = {A,C,D}

Etape i=2 : Combinaison à 2

C2 = {AC, AD, CD}

C2	AC	AD	CD
Sup	3/5	3/5	3/5

F2 = {AC, AD, CD}

Etape i=3 : Combinaison à 3

C3 = {ACD}

C3	ACD
Sup	3/5

F3 = {ACD}

Etape i=4 : Combinaison à 4

C4 = { } : L'ensemble vide. On s'arrete.

Donc : F = ∑ Fi = F1 ∪ F2 ∪ F3

= { A, C, D, AC, AD, CD, ACD }

Déduction de toutes les règles d'association, puis les règles ayant une Confiance ≥ 80% :

Au-delà de la 2ème itération

Motif Fréquent	Règle	Confiance	Qualité
AC	A ⇒ C	3/3 = 100%	Bonne Qualité
	C ⇒ A	3/4 = 75%	
AD	A ⇒ D	3/3 = 100%	Bonne Qualité
	A ⇒ D	3/4= 75%	
CD	C ⇒ D	3/4 = 75%	
	D ⇒ C	3/4 = 75%	
ACD	A ⇒ CD	3/3 = 100%	Bonne Qualité
	C ⇒ AD	3/4 = 75%	
	D ⇒ AC	3/4 = 75%	
	AC ⇒ D	3/3 = 100%	Bonne Qualité
	AD ⇒ C	3/3 = 100%	Bonne Qualité
	CD ⇒ A	3/3 = 100%	Bonne Qualité

Mesures de performance de la classification

Définitions :

Lorsque vous effectuez une classification à l'aide d'algorithmes de data mining, c'est important d'évaluer la performance du modèle. Les mesures de performance fournissent des indicateurs sur la qualité des prédictions du modèle. Voici quelques-unes des mesures les plus utilisées:

1. Matrice de Confusion :

La matrice de confusion est une table qui permet de résumer les performances d'un algorithme de classification. Elle compare les classes réelles aux classes prédites et comprend quatre éléments : Vrai Positif (VP), Vrai Négatif (VN), Faux Positif (FP) et Faux Négatif (FN).

	Prédit positif	Prédit négatif
Vrai positif	VP	FN
Vrai négatif	FP	VN

- Vrai Positif (VP) : le nombre d'observations positives correctement prédites.
- Vrai Négatif (VN) : le nombre d'observations négatives correctement prédites.
- Faux Positif (FP) : le nombre d'observations négatives incorrectement prédites comme positives.
- Faux Négatif (FN) : le nombre d'observations positives incorrectement prédites comme négatives.

2. Précision du Modèle (Precision) :

La précision mesure le nombre d'instances correctement prédites comme positives par rapport au nombre total d'instances prédites comme positives (Vrai Positif / (Vrai Positif + Faux Positif)). Elle indique la qualité des prédictions positives.

3. Exactitude (Accuracy) :

L'exactitude mesure la proportion totale de prédictions correctes parmi toutes les prédictions, qu'elles soient positives ou négatives. Sa formule est comme suit : ((Vrai Positif + Vrai Négatif) / Total).

4. Taux d'Erreur du Modèle :

Le taux d'erreur représente le pourcentage d'erreurs de classification dans l'ensemble de test ((Faux Positif + Faux Négatif) / Total). Il donne une idée globale de la performance du modèle.

NB : Taux d'erreur = 1 - Accuracy

5. Rappel (Sensibilité) du Modèle (Recall) :

Le rappel mesure le nombre d'instances correctement prédites comme positives par rapport au nombre total d'instances réellement positives (Vrai Positif / (Vrai Positif + Faux Négatif)). Il est également appelé sensibilité ou taux de vrais positifs.

Solution exercice 2 – Mesures de performance de la classification:

Matrice de confusion :

Test	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Classe réelle	Non	Oui	Non	Oui	Oui	Non	Oui	Non	Non	Oui	Oui	Non	Oui	Oui	Non	Oui	Oui	Non	Non	Oui
Classe prédite	Oui	Oui	Non	Oui	Oui	Non	Non	Non	Oui	Oui	Oui	Non	Oui	Non	Oui	Oui	Non	Non	Oui	Oui

	Prédit positif	Prédit négatif
Vrai positif	VP = 8	FN = 3
Vrai négatif	FP = 4	VN = 5

Précision du modèle :

$$\frac{VP}{VP + FP} = \frac{8}{8 + 4} = \frac{8}{12} = 0{,}666\ldots$$

Exactitude (accuracy) :

$$\frac{VP + VN}{Total} = \frac{8 + 5}{20} = \frac{13}{20} = \ 0{,}65$$

Taux d'erreur :

$$\frac{FP + FN}{Total} = \frac{4 + 3}{20} = \frac{7}{20} = 0{,}35$$

Ou : Taux d'erreur = 1 – Exactitude = 1 – 0,68 = 0,35