
Examen

Questions (5 points)

1. Pourquoi le prétraitement des données est-il important dans les algorithmes de machine learning ?
2. Que représente la confiance dans une règle d'association ?
3. Quelle est la différence entre k-means et DBSCAN ?
4. Qu'est-ce que la marge dans un SVM, et pourquoi cherche-t-on à la maximiser ?

Exercice 1. (10 points)

La Table 1 présente une base transactionnelle de produits achetés par des clients. Chaque ligne correspond à une transaction, et chaque produit est représenté par un 1 (acheté) ou un 0 (non acheté).

TABLE 1 – Base de transactions.

Transactions	A	B	C	D	E
1	0	0	1	1	1
2	1	1	0	1	1
3	1	1	1	1	1
4	0	0	0	0	1

1. Appliquez l'algorithme FP-Growth avec un support minimal de 50% pour identifier les motifs fréquents dans cette base transactionnelle.
2. Est-ce que la règle $C \rightarrow D$ est intéressante ? Justifiez votre réponse.

Nous allons utiliser les mêmes transactions mais en classifiant chaque ligne selon un critère spécifique. Par exemple, la classe peut être déterminée par l'achat du produit **A** (si le produit A est acheté, la classe = 1, sinon la classe = -1).

3. Dessinez les points dans un plan bidimensionnel en prenant en compte les **produits B et C comme caractéristiques d'entrée**, et identifiez visuellement une hyperplan qui sépare les deux classes.
4. Calculez le vecteur $\vec{w}$ pour un hyperplan de séparation optimale. Les multiplicateurs de Lagrange et le biais b sont :

$$\alpha_1 = 0, \quad \alpha_2 = 1, \quad \alpha_3 = 1, \quad \alpha_4 = 0, \quad b = -0.5$$

5. Déduire la classe de la transaction (1, 1, 0, 1, 1) (Tous les produits sont achetés sauf C) en utilisant l'hyperplan trouvé.

Exercice 2. (5 points)

La Table 2 présente une base d'entraînement utilisée pour prédire si un client effectuera un achat en ligne. Les colonnes représentent des attributs tels que le temps passé sur le site, la catégorie d'intérêt, etc.

TABLE 2 – Base d'entraînement.

ID	Temps	Catégorie d'Intérêt	Métier	Achètera (Oui/Non)
1	5-10	Sports	Etudiant	Oui
2	10-15	Voyage	Ingénieur	Non
3	15-20	Maison	Enseignant	Oui
4	5-10	Maison	Etudiant	Non
5	5-10	Sports	Cadre	Oui
6	5-10	Voyage	Cadre	Non
7	15-20	Sports	Ingénieur	Oui
8	15-20	Voyage	Enseignant	Oui
9	10-15	Maison	Cadre	Non
10	5-10	Sports	Cadre	Oui

Étant donné que la racine de l'arbre est **Temps**, on utilise l'index de Gini.

1. Dédurre l'arbre de décision (**sans calculs**).
2. Donner l'Accuracy du modèle.
3. Donner la matrice de confusion en l'expliquant.

Bon courage.

Corrigé type - Examen

Questions (5 points)

- Pourquoi le prétraitement des données est-il important ? Il améliore la qualité des données en gérant les valeurs manquantes, le bruit et les caractéristiques disproportionnées, ce qui conduit à de meilleures performances des modèles. (1 pts)

- Que représente la confiance dans une règle d'association ?

C'est la probabilité qu'une conséquence B soit vraie lorsque la condition A est satisfaite (1 pts) :

$$\text{Confidence}(A \Rightarrow B) = \frac{\text{Support}(A \cup B)}{\text{Support}(A)}.$$

- Différence entre k-means et DBSCAN :

- k-means : le nombre de classes k est fixé à l'avance, sensible au bruit, clusters sphériques. (0.5 pts)
- DBSCAN : pas besoin de k , détecte des clusters denses et arbitraires, gère le bruit. (0.5 pts)

- Qu'est-ce que la marge dans un SVM et pourquoi la maximiser ? La marge est la distance entre l'hyperplan et les points les plus proches (Support Vecteurs) (1 pts). Maximiser la marge améliore la généralisation et la robustesse du modèle. (1 pts)

Exercice 1. (10 points)

1. Motifs fréquents (FP-Growth)

On calcule les motifs de taille 1 (Table 1).

TABLE 1 – Supports des motifs de taille 1. (0.5 pts)

Item	support
A	2
B	2
C	2
D	3
E	4

En suite les items sont ordonnés selon leur support (Table 2 (0.5 pts)).

TABLE 2 – Support par ordre.

Item	support
E	4
D	3
C	2
B	2
A	2

TABLE 3 – Items ordonnés selon leur support. (0.5 pts)

Transactions	A	B	C	D	E	Ordre des items
T1	0	0	1	1	1	<i>E, D, C</i>
T2	1	1	0	1	1	<i>E, D, B, A</i>
T3	1	1	1	1	1	<i>E, D, C, B, A</i>
T4	0	0	0	0	1	<i>E</i>

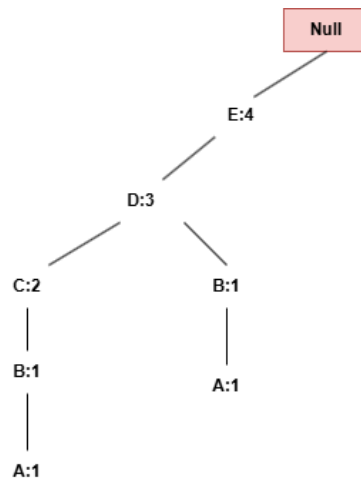


FIGURE 1 – L'arbre FP-tree. (2 pts)

Pour construire l'arbre FP-tree, on liste les item sets dans la liste des transactions selon leur ordre d'apparition (3).

L'arbre FP-Tree est donnée dans la figure 1.

À partir de l'arbre FP-Tree, on construit la table 4.

TABLE 4 – Motifs fréquents. (2.5 pts)

Item	Base de motifs conditionnels	FP-tree	Motifs fréquents trouvés
A	$\{EDCB : 1\}, \{EDB : 1\}$	$\{E, D, B : 2\}, \{C : 1\}$	$\{EA : 2\}, \{DA : 2\}, \{BA : 2\}, \{DBA : 2\}, \{EDA : 2\}, \{EBA : 2\}, \{EDBA : 2\}$
B	$\{ED : 1\}, \{EDC : 1\}$	$\{E : 2, D : 2, C : 1\}$	$\{EB : 2\}, \{DB : 2\}, \{EDB : 2\}$
C	$\{ED : 2\}$	$\{E, D : 2\}$	$\{EC : 2\}, \{DC : 2\}, \{EDC : 2\}$
D	$\{E : 3\}$	$\{E : 3\}$	$\{ED : 3\}$

L'ensemble des motifs fréquents sont : $\{A\}, \{B\}, \{C\}, \{D\}, \{E\}, \{EA\}, \{DA\}, \{BA\}, \{DBA\}, \{EDA\}, \{EBA\}, \{EB\}, \{DB\}, \{EDB\}, \{EDBA\}, \{EC\}, \{DC\}, \{EDC\}, \{ED\}$.

Si vous choisissez un autre ordre dans la table 2, cela reste aussi une solution correcte.

2. Est-ce que la règle $C \rightarrow D$ est intéressante ? Justifiez votre réponse.

La règle $C \rightarrow D$ signifie "Si le produit C est acheté, alors le produit D est aussi acheté". Calculons le support et la confiance de cette règle (0.5 pts) :

- Support de $C \cap D$: les transactions 1 et 3, soit $2/4 = 50\%$.
- Support de C : $2/4 = 50\%$.
- Confiance de $C \rightarrow D$: $\frac{50\%}{50\%} = 1$ (100%).

La règle a une confiance de 100%, ce qui signifie qu'à chaque fois que C est acheté, D l'est aussi. Cette règle est donc intéressante. (0.5 pts)

3. Dessiner les points dans un plan bidimensionnel en prenant en compte les produits B et C comme caractéristiques d'entrée.

La représentation graphique des points de coordonnées B,C est donnée par la figure 2. On voit bien à travers cette figure que la droite optimale qui permet de séparer correctement les points de coordonnées (B,C) est la droite $B=0.5$.

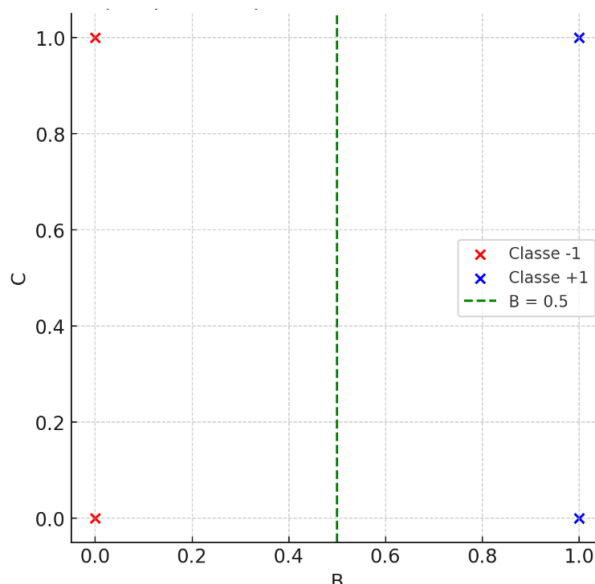


FIGURE 2 – La représentation graphique des points de coordonnées (B,C). (1 pts)

4. Calculez le vecteur $\vec{w}$ pour un hyperplan de séparation optimale. (1 pts)

Les multiplicateurs de Lagrange sont donnés par :

$$\alpha_1 = 0, \quad \alpha_2 = 1, \quad \alpha_3 = 1, \quad \alpha_4 = 0, \quad b = -0.5$$

Le vecteur $\vec{w}$ est calculé comme suit :

$$\vec{w} = \sum_{i=1}^4 \alpha_i y_i \vec{x}_i$$

Où y_i est la classe associée à chaque point et $\vec{x}_i$ est le vecteur des caractéristiques (B, C) .

En utilisant les valeurs de α données :

$$\vec{w} = (w_1, w_2) = 1 \times (1, 0) + 1 \times (1, 1) = (2, 1)$$

L'équation du séparateur (droite) est donnée donc par :

$$w_1 \times B + w_2 \times C + b = 0 \quad (1)$$

ce qui est équivalent à

$$2 \times B + C - 0.5 = 0 \quad (2)$$

On peut bien remarquer que contrairement à la droite tracée dans la question 1, cette nouvelle droite, elle ne correspond pas à la droite optimale (voir la figure 3).

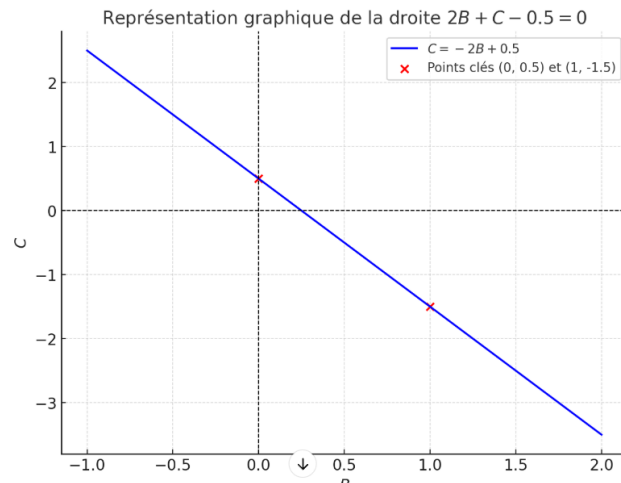


FIGURE 3 – Séparateur trouvé.

Pour trouver la droite optimale, les multiplicateurs de Lagrange devraient être par exemple :

$$\alpha_1 = 0, \quad \alpha_2 = 1, \quad \alpha_3 = 0, \quad \alpha_4 = 1, \quad b = -0.5$$

5. Dédurre la classe de la transaction (1, 1, 0, 1, 1) (1 pts)

En utilisant la droite de décision $2B + C - 0.5 = 0$, on peut déduire la classe de la transaction en remplaçant ces coordonnées $B = 1$, et $C = 0$.

$$2(1) + 0 - 0.5 = 1.5 \quad (3)$$

Le signe est positive, donc sa classe est +1.

Exercice 2.

1. Arbre de décision (sans calculs)

En considérant que la racine de l'arbre est le **Temps**, on peut construire l'arbre de décision comme suit : Nous remarquons que, pour les cas où le **Temps** est égal à 10-15 et 15-20, la classe est toujours la même, respectivement **Non** et **Oui**. En revanche, pour le cas où le **Temps** est égal à 5-10, nous avons trois exemples dans la classe **Oui** et deux exemples dans la classe **Non**. Nous devons donc diviser ce groupe selon un autre attribut.

Si nous choisissons l'attribut **Catégorie d'intérêt**, nous constatons que :

- Tous les exemples appartenant à la catégorie **Sports** (3 exemples) sont dans la classe **Oui**.
- Tous les exemples appartenant à la catégorie **Voyage** (1 exemple) sont dans la classe **Non**.
- Tous les exemples appartenant à la catégorie **Maison** (1 exemple) sont dans la classe **Non**.

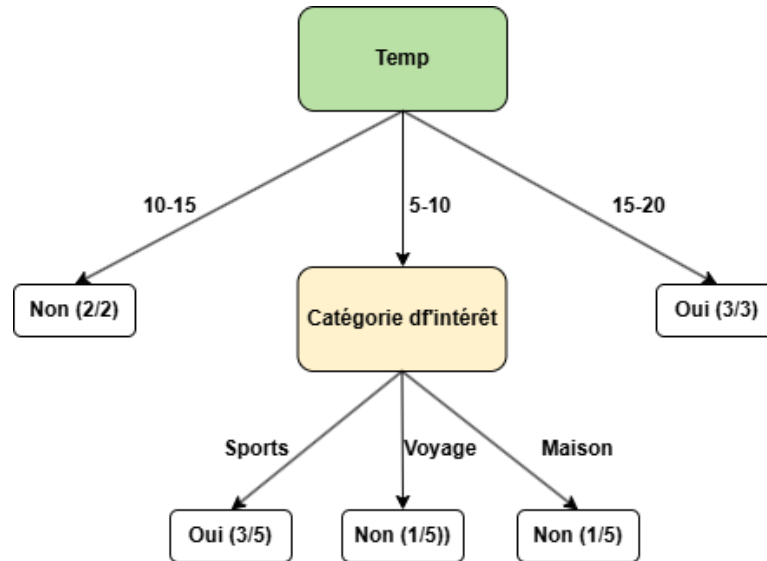


FIGURE 4 – L’arbre de décision correspondant à la table 2. (2.5 pts)

En revanche, si nous prenons l’attribut **Métier**, les exemples ne sont pas bien répartis. Par exemple, nous trouvons un étudiant dans la classe **Oui** et un autre dans la classe **Non**. Ainsi, le meilleur attribut pour diviser les données après le chemin **Temps = 5-10** est **Catégorie d’intérêt**.

L’arbre final est donné dans la figure 4.

2. Accuracy du modèle

TABLE 5 – Base d’entraînement.

ID	Temps	Catégorie d’Intérêt	Métier	Y	$\hat{Y}$
1	5-10	Sports	Etudiant	Oui	Oui
2	10-15	Voyage	Ingénieur	Non	Non
3	15-20	Maison	Enseignant	Oui	Oui
4	5-10	Maison	Etudiant	Non	Non
5	5-10	Sports	Cadre	Oui	Oui
6	5-10	Voyage	Cadre	Non	Non
7	15-20	Sports	Ingénieur	Oui	Oui
8	15-20	Voyage	Enseignant	Oui	Oui
9	10-15	Maison	Cadre	Non	Non
10	5-10	Sports	Cadre	Oui	Oui

L’accuracy est définie comme le ratio des prédictions correctes sur le total des prédictions.

$$\text{Accuracy} = \frac{\text{Nombre de prédictions correctes}}{\text{Nombre total de données}}$$

Dans ce cas, en appliquant l'arbre de décision, toutes les prédictions sont correctes. Il y a 10 exemples au total, donc :

$$\text{Accuracy} = \frac{10}{10} = 1 \text{ (ou 100\%).}$$

(1 pts)

3. Matrice de confusion

La matrice de confusion compare les prédictions du modèle aux résultats réels. Pour ce problème binaire (**Oui/Non**), la matrice est **(1 pts)** :

Classe réelle	Oui (Prédit)	Non (Prédit)
Oui (Réal)	6	0
Non (Réal)	0	4

Explication (0.5 pts) :

- 6 clients ont été correctement prédits comme **Achètera = Oui**.
- 4 clients ont été correctement prédits comme **Achètera = Non**.
- Il n'y a aucune erreur dans les prédictions (**Faux Positifs** ou **Faux Négatifs** = 0).