DOI – EXAMEN

Les exercices sont indépendants. Tout document de cours et de TD est autorisé. Le barème est donné à titre indicatif.

1 Optimisation (6 points)

Une entreprise, disposant de 10 000 m^2 de carton en réserve, fabrique et commercialise 2 types de boites en carton. La fabrication d'une boite en carton de type 1 ou 2 requiert, respectivement, 1 et 2 m^2 de carton ainsi que 2 et 3 minutes de temps d'assemblage.

Seules 200 heures de travail sont disponibles pendant la semaine à venir. Les boites sont agrafées et, pour une boite du second type, il faut quatre fois le nombre d'agrafes nécessaires à une boite du premier type. Le stock d'agrafes disponible permet d'assembler au maximum 15 000 boites du premier type. Les boites sont vendues respectivement 3 et 5 euros.

Q1.1 Formuler le problème de la recherche d'un plan de production maximisant le chiffre d'affaires de l'entreprise sous forme d'un programme linéaire (Indication : la valeur numérique du nombre d'agrafes pour fabriquer une boite de type 1 n'est pas nécessaire pour établir la contrainte sur le nombre d'agrafes).

On désire trouver le plan de production optimal à l'aide du solveur cplex (en mode interactif).

Q1.2 Rédiger en langage OPL le programme linaire établi en question Q1.1.

La résolution avec cplex conduit au plan de production optimal suivant :

— nombre de boites de type 1 à fabriquer : 600
— nombre de boites de type 2 à fabriquer : 3 600

Le chiffre d'affaires optimal est de 19 800 euros.

Q1.3 Donner le programme dual du programme primal établi en Q1.1 et le résoudre. Vérifier la validité de la solution du dual.

Q1.4 En téléphonant à son fournisseur, l'entreprise apprend qu'il est possible de se faire livrer immédiatement du carton au prix de 2 centimes le m^2. Que conseillez-vous au responsable des réapprovisionnements et pourquoi?

2 Partitionnement de données (4 points)

Considérons quatre relations candidates pour les processus de Machine Learning :

— R(A, B) fragmentée horizontalement en R1, R2 et R3 en utilisant l'attribut B tels que :
  — $R1 = \sigma_{B<30}(R)$
  — $R2 = \sigma_{30\leq B<60}(R)$
  — $R3 = \sigma_{B\geq 30}(R)$
— S(A, C, D) partitionnée en trois fragments S1, S2, et S3 en fonction du schéma de fragmentation de la table R.
— T(C, E, F, G) fragmentée verticalement en T1(C, E), T2(C, F) et T3(C, G).
— U(G, H) fragmentée horizontalement en deux fragments U1 et U2 tels que :
  — $U1 = \sigma_{H<500}(U)$
  — $U2 = \sigma_{H\geq 500}(U)$

Considérions la requête suivante pour extraire des instances :

$$\Pi_{B,F,H}(\sigma_{B<20 \wedge H<800}(R \text{ join } S \text{ join } T \text{ join } U))$$

Q2.1 Donner une réécriture de cette requête en fonction des fragments.

3 Itemsets fréquents et règles d'association (5,5 points)

La table 1 indique la présence (1) ou l'absence (0) de personnages dans huit romans de Terry Pratchett. Pour simplifier l'écriture, chaque personnage est représenté par une initiale (A, C, G, H, R, S ou T).

	Angua von Überwald	**C**arrot Ironfoundersson	**G**ranny Weatherwax	**H**avelock Vetinari	**R**incewind	**S**amuel Vimes	**T**woflower
Livre	**A**	**C**	**G**	**H**	**R**	**S**	**T**
The Colour of Magic	0	0	0	0	1	0	1
Wyrd Sisters	0	0	1	0	0	0	0
Guards! Guards!	0	1	0	1	0	1	0
Witches Abroad	0	0	1	0	0	0	0
Men at Arms	1	1	0	1	0	1	0
Interesting Times	0	0	0	1	1	0	1
Night Watch	1	1	0	0	0	1	0
Thud!	1	1	0	1	0	1	0

TABLE 1 – Apparitions de personnages dans huit romans

Pour les questions de cet exercice, lorsqu'une valeur de support est demandée, vous pouvez choisir de calculer, au choix, le support relatif ou le support absolu.

Q3.1 Calculer le support des itemsets suivants :

— {H}
— {A, C}
— {C, G}
— {A, C, H}

Q3.2 Donner une interprétation textuelle de la valeur de support de l'itemset {C, G}.

Q3.3 Lister les itemsets fréquents pour un support minimum de 3, avec leur valeur de support.

Sur cette même table 1, les itemsets fréquents maximaux calculés pour un support minimum de 1 sont les suivants :

— {G} (support = 2)
— {H, R, T} (support = 1)
— {A, C, H, S} (support = 2)

Q3.4 À partir de ces fréquents maximaux *uniquement*, donner, en justifiant votre réponse, une borne inférieure ou supérieure de la valeur de support des itemsets suivants :

— {H}
— {G, T}
— {A, C}

Q3.5 En vous basant sur votre réponse à la question Q3.3, lister les itemsets fréquents clos pour un support minimum de 3.

Q3.6 Calculer la confiance et la mesure de Kulczynski des règles d'association suivantes :

— A, C → H
— H → A, C
— C → S
— S → C

La table 2 donne les thèmes principaux associés à chaque roman.

Livre	**Thème**
The Colour of Magic	Rincewind
Wyrd Sisters	Witches
Guards! Guards!	City Watch
Witches Abroad	Witches
Men at Arms	City Watch
Interesting Times	Rincewind
Night Watch	City Watch
Thud!	City Watch

TABLE 2 – Thèmes principaux des romans

Q3.7 Comment feriez-vous pour étudier les règles d'association qui existent entre les personnages et les thèmes des romans ?

4 Denial constraints (1,5 points)

La table 3 (notée E dans la suite de l'exercice) liste les examens de la période B.

Matière	**Date**	**Option**	**Durée**	**Salle**	**Coefficient**
LOS	2021-03-22 9:30	IA	120	A106	1
SAV	2021-03-22 13:45	IA	90	B140	0.5
APM	2021-03-22 16:30	IA	90	B140	0.5
DOI	2021-03-23 9:30	IA-D	120	B374	1
VAT	2021-03-23 9:30	IA-S	120	A102	0.5
SSE	2021-03-23 16:45	IA-S	120	B140	0.5

TABLE 3 – Examens de la période B

Q4.1 Donner, pour chacune des *denial constraints* suivantes : une interprétation textuelle de la règle, sa validité par rapport aux données (règle vraie ou fausse) et un contre-exemple si la règle est fausse.

— $\forall t_1, t_2 \in E, \neg(t_1.\text{date} = t_2.\text{date} \wedge t_1.\text{option} = t_2.\text{option})$
— $\forall t_1, t_2 \in E, \neg(t_1.\text{durée} < t_2.\text{durée} \wedge t_1.\text{coefficient} > t_2.\text{coefficient})$
— $\forall t \in E, \neg(t.\text{option} = \text{IA} \wedge t.\text{salle} \neq \text{B140})$
— $\forall t_1, t_2 \in E, \neg(t_1.\text{option} = t_2.\text{option} \wedge t_1.\text{coefficient} < t_2.\text{coefficient} \wedge t_1.\text{date} > t_2.\text{date})$

[L'exercice 5 est sur la page suivante]

5 Classification supervisée (3 points)

Une entreprise cherche à améliorer ses critères de présélection des moteurs à étudier dans le cadre de son programme de maintenance prédictive. Le responsable du programme a estimé qu'une opération de maintenance sur un moteur sain représente un surcoût d'environ 20 k€, alors que l'absence de maintenance sur un moteur défaillant représente en moyenne un surcoût de 800 k€.

Lors d'une étude préliminaire de classification avec un modèle linéaire, vous séparez les 15 000 cas répertoriés en deux ensembles : 13 500 cas servant à entraîner le modèle et 1 500 cas pour le test. Vous obtenez alors comme résultat la matrice de confusion suivante sur le jeu de données de test :

	Classe réelle	
Classe prédite	**Défaillant**	**Sain**
Défaillant	32	58
Sain	28	1382

Q5.1 Calculer le rappel et le taux de faux positifs [1] de ce modèle.

En étudiant l'impact du seuil de classification du modèle utilisé, vous obtenez la courbe ROC de la figure 1.

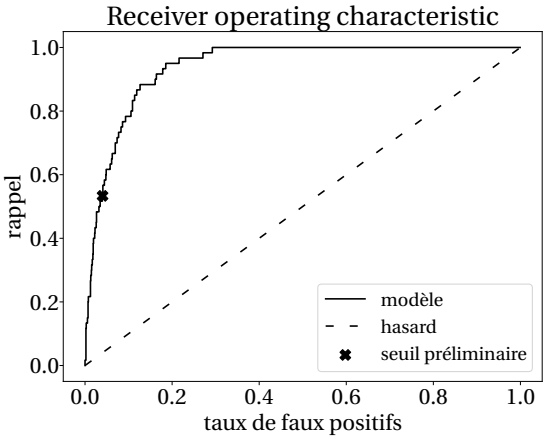

FIGURE 1 – Courbe ROC du modèle prédictif linéaire

Q5.2 Par rapport aux besoins de l'entreprise exprimés en début d'exercice, que suggérez-vous comme ajustement du seuil de détection ?

Q5.3 Vous souhaitez ensuite étudier l'influence du découpage réalisé entre jeu d'entraînement et jeu de test sur la qualité de la prédiction. Comment procédez-vous ?

1. False Positive Rate : $\mathrm{FPR} = \dfrac{\mathrm{FP}}{\mathrm{FP}+\mathrm{TN}}$