Module Fouille de données

2h. Seuls les transparents du cours (slides) sont autorisés.

NB : La notation tient compte de la présentation. Elle est donnée à titre indicatif.

Exercice 1 : Classification supervisée (7 pts)

classe prédite / classe réelle	A	B	C	D	Total
A	40	1	2	2	**45**
B	1	13	1	0	**15**
C	2	2	17	0	**21**
D	1	0	0	18	**19**
Total	**44**	**16**	**20**	**20**	**100**

La *précision* pour une classe donnée mesure le taux d'exemples corrects parmi les exemples prédits dans cette classe. Le *rappel* mesure le taux d'exemples corrects parmi les exemples de la classe. Le taux de *faux positifs* d'une classe mesure le nombre d'objets positifs parmi ceux n'appartenant pas à la classe. Le taux de *vrais positifs* d'une classe mesure le nombre d'objets positifs parmi les vrais objets de la classe. Soit la matrice ci-dessus obtenue par application du modèle M1. Pour les questions 1) à 6), vous devez **indiquer aussi la formule de calcul**.

1- Calculer le taux d'erreur en généralisation ; 0,5 pt
2- Calculer le taux de généralisation (accuracy rate) ; 0,5 pt
3- Calculer le taux de faux positifs (FP rate) pour la classe A ; 0,5 pt
4- Calculer le taux de vrais positifs (TP rate) pour la classe B ; 0,5 pt
5- Calculer la précision pour la classe C ; 0,5 pt
6- Calculer le rappel pour la classe D ; 0,5 pt
7- On souhaite calculer la précision totale du modèle M1 en tenant compte de toutes les classes. Donner la formule correspondante. Il n'est pas nécessaire d'effectuer le calcul. 1 pt
8- Ecrire un algorithme (pseudo-code) qui, à partir de 2 matrices T1 et T2 identiques à celle ci-dessus, obtenues sur un même jeu de données (15324 exemples, 21 attributs) par application respective des modèles M1 et M2, donne comme résultat le modèle à choisir par l'utilisateur pour la résolution de son problème. 2 pts
9- Comment peut t-on s'assurer que ce mode de comparaison de modèles est robuste ? 1 pt

Exercice 2 : Motifs et Règles d'association (10 pts)

Let {a, b, c, d, e} be a set of items ; $\{t_1, t_2, t_3, t_4, t_5, t_6\}$ be a set of transactions described as :

t_1 = b, c, d t_2 = a, b, c, d, e t_3 = a, b, c, e t_4 = a, b, d, e t_5 = b, c, e t_6 = a, b, d, e

The minimum support threshold is to set to 3/6 i.e. 50%.

Let U is a frequent itemset ; V is a proper subset of U ; and V is not empty.

Theorem : If a rule V → U–V does not satisfy the confidence threshold, then any rule V' → U-V', where V' is a subset of V, must not satisfy the confidence threshold as well.

1. Prove the theorem above . 1 pt
2. Find all frequent itemsets using Apriori algorithm. Provide details. 3 pts
3. Draw the lattice of frequent itemsets. 1 pt
4. Find all frequent closed itemsets. Explain. 1 pt
5. Find one frequent closed itemset that has two minimal generators. Provide details. 1 pt
6. Find all frequent maximal itemsets. 1 pt
7. Give the negative border of frequent itemsets. 1 pt
8. Let the minimal confidence equal to 66%, list three valid rules. Provide details. 1 pt

Exercice 3 : classification (3 pts)

Une entreprise dispose d'une base de données de 3400 clients et d'une vingtaine d'attributs parmi lesquels 10 attributs sont jugés pertinents pour un objectif de classification non supervisée. Parmi ces 10 attributs, cinq sont à valeur booléenne, trois sont à valeur nominale ou catégorielle et deux ont des valeurs numériques (dans $\Re$).

Question : Comment allez-vous procéder pour élaborer une classification de ces clients, qui doit permettre de définir une politique relationnelle avec ces clients ?