Module Fouille de données

2h. Documents non autorisés sauf **1 feuille A4 recto-verso MANUSCRITE**

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

Exercice 1 : Classification supervisée (10 pts)

1) Pourquoi l'élagage est utile pour l'induction à partir d'arbres de décision ? (1 pt)
2) Quelles sont les 2 approches courantes pour l'élagage d'arbres de décision ? Expliquer leur principe. (1 pt)
3) Comparer les avantages et inconvénients de la classification de type « eager » (ex : arbre de décision, règles, réseau bayésien) par rapport à la classification de type « lazy » (ex : k-plus proches voisins, raisonnement à partir de cas). (1 pt)
4) Etant donnée la relation suivante « weather.symbolic » où l'attribut de classe est « Play». Quels sont les attributs pertinents pour la classification ? Expliquez pourquoi. (1pt)
5) En considérant que l'ensemble d'apprentissage contient les 8 premières instances {O1 à O8}. Utilisez la technique de classification bayésienne naïve avec la formule de Laplace et les attributs pertinents pour prédire la classe de chacune des 6 dernières instances {O9 à O14}. (3 pts)
6) Construire la matrice de confusion. (0,5 pt)
7) Déterminer le taux d'erreur, la précision et le rappel. (1,5 pt)
8) Quel est le type de la méthode d'évaluation utilisée ici ? (1 pt)

```
@relation weather.symbolic
@attribute name          nominal
@attribute outlook       {sunny, overcast, rainy}
@attribute temperature   {hot, mild, cool}
@attribute humidity      {high, normal}
@attribute windy         {TRUE, FALSE}
@attribute play          {yes, no}
@data
O1,  overcast, hot,      high,     FALSE,    yes
O2,  rainy,    mild,     high,     FALSE,    yes
O3,  rainy,    cool,     normal,   FALSE,    yes
O4,  overcast, cool,     normal,   TRUE,     yes
O5,  sunny,    cool,     normal,   FALSE,    yes
O6,  sunny,    hot,      high,     FALSE,    no
O7,  sunny,    hot,      high,     TRUE,     no
O8,  rainy,    cool,     normal,   TRUE,     no
O9,  rainy,    mild,     normal,   FALSE,    yes
O10, sunny,    mild,     normal,   TRUE,     yes
O11, sunny,    mild,     high,     FALSE,    no
O12, overcast, mild,     high,     TRUE,     yes
O13, overcast, hot,      normal,   FALSE,    yes
O14, rainy,    mild,     high,     TRUE,     no
```

Exercice 2 : Recherche de motifs ensemblistes (8 pts)

Soient O un ensemble fini d'exemples ou de transactions ; A un ensemble fini d'attributs ou d'items, X et Y sont des sous-ensembles de A.

- Un ensemble d'items X est dit fréquent si le nombre de transactions contenant les items de X est supérieur au seuil de support minimum (minsup).
- Une règle associative (ou d'association) est une implication de la forme « si X alors Y » (notée X → Y, X est la prémisse, et Y est la conclusion) et l'intersection de X et Y est vide, traduisant le fait que si les items X sont présents dans une transaction, alors les items Y le sont avec une certaine probabilité (confiance).

- Le support d'une règle est la mesure indiquant le pourcentage de transactions qui vérifient une règle associative.
- La confiance d'une règle est la mesure indiquant le pourcentage de transactions qui vérifient la conclusion d'une règle associative parmi celles qui vérifient la prémisse.

Soit l'ensemble des transactions suivantes : {P, L}, {P, C, B, O}, {L, C, B, K}, {P, L, C, B}, {P, L, C, K}.

Travail demandé.

1- Représenter ces transactions dans une table transactionnelle horizontale, puis verticale, et enfin dans une table relationnelle. (1 pt)
2- Calculer les itemsets candidats et fréquents pour un support minimum égal à 2. Dessiner le demi-treillis correspondant, en indiquant le support de chaque nœud. (3 pts)
3- Parmi les itemsets générés, lister 2 itemsets fréquents maximaux, et 2 itemsets fréquents fermés qui ne sont pas maximaux. (1 pt)
4- Générer au moins deux règles associatives valides (minconf = 60%), à partir d'un 3-itemset. (1 pt)
5- Quelle est la taille de l'espace de recherche des itemsets fréquents d'un ensemble A ayant m items ? Quel est le nombre de règles que l'on peut générer à partir d'un k-itemset ? (1 pt)
6- Montrer que la confiance de X → Y n'est pas toujours égale à celle de Y → X. (1 pt)

Exercice 3 : Motifs séquentiels (2 pts)

1- Expliquer le problème de fouille de motifs séquentiels. Quelle en est l'utilité ? (1 pt)
2- Décrivez le principe de résolution. (1 pt)