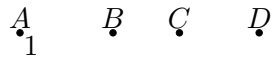
Data Mining et Data Warehousing, 29 mai 2012

Cahier fermé. Durée : 2 heures et 30 minutes

Nom et prénom
Année

Question 1



A, B, C et D sont quatre points colinéaires dans le plan. La distance entre A et B est de 7. La distance entre B et C est de 5. La distance entre C et D est de 6.

1. Détaillez l'exécution de l'algorithme de Dasgupta et Long ($\beta = 2$) sur ce jeu de points. Le point A est numéroté 1.
2. Dessinez le dendrogramme qui en résulte.
3. Pour $k \in \{2, 3\}$, comparez le k-clustering obtenu avec le k-clustering optimal.

.../20

.../SUITE

Situez chaque terme dans le cursus et expliquez de façon succincte mais précise.

Question 2 Leave-one-out.

.../5

Question 3 Closed itemset.

.../5

Question 4 Ward's method.

.../5

Question 5 Bisecting K-Means.

.../5

Question 6 À la page 356, les auteurs mentionnent la propriété suivante :

> If a rule $X \to Y \setminus X$ with $X \subseteq Y$ does not satisfy the confidence threshold, then any rule $X' \to Y \setminus X'$ with $X' \subseteq X$ must not satisfy the confidence threshold as well.

Donnez une preuve de cette propriété.

.../10

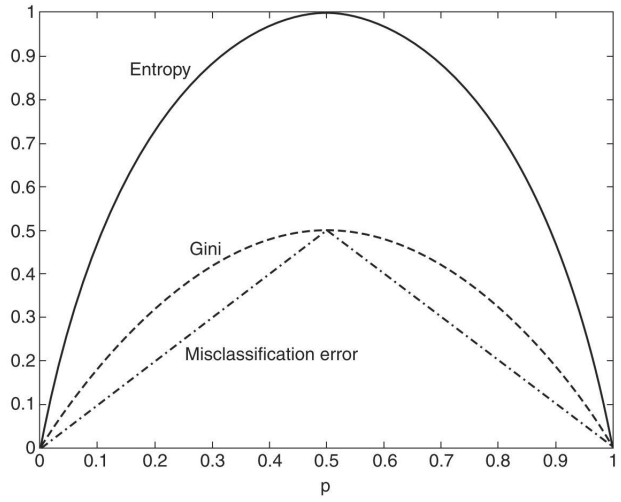


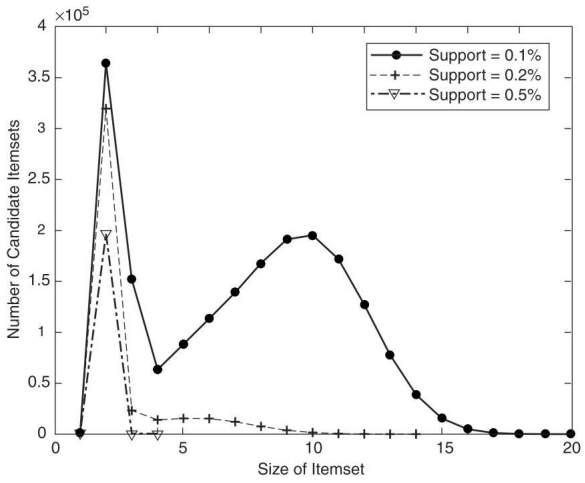
FIGURE 1 –

Question 7 Voir la figure 1.

1. Quelle est la valeur exacte de `Gini` pour $p = 0.3$?
2. Quelle est la valeur exacte de `Misclassification Error` pour $p = 0.373$?

Détaillez les calculs.

.../4



(a) Number of candidate itemsets.

FIGURE 2 –

Question 8 Expliquez la figure 2 en détail. Votre explication doit contenir (mais ne doit pas se limiter à) des réponses aux questions suivantes :

1. Pourquoi le "pic" de gauche ?
2. Comment est il possible que le nombre de candidats de taille 10 soit nettement supérieur au nombre de candidats de taille 4 ?

Question 9 Expliquez la figure 3 de façon détaillée. Votre explication doit contenir (mais ne doit pas se limiter à) des réponses aux questions suivantes :

1. Qu'est-ce que MDL ? Comment peut-on comparer la qualité de deux modèles de classification selon ce principe ?
2. Est-ce que ce principe est limité aux arbres de décision ?
3. Quel est le lien entre MDL et l'*overfitting* ?

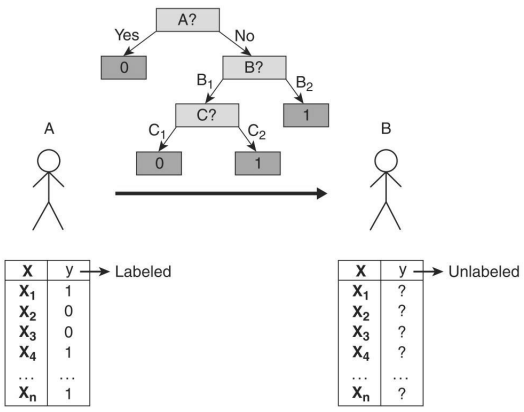


FIGURE 3 – The MDL principle.

.../10

Réponse à la question 8.

.../10

Réponse à la question 9.

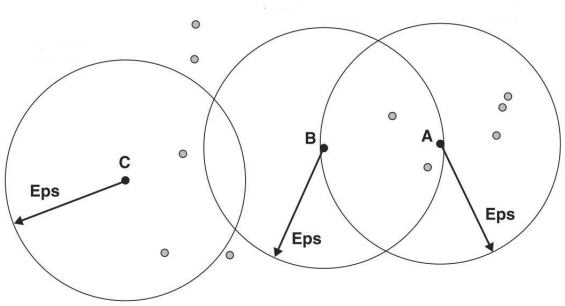


FIGURE 4 – Z.

Question 10 Voir la figure 4. L'objectif de cette figure est d'illustrer trois points (A, B et C) de type différent. Pour quelle(s) valeur(s) de *MinPts* cette-image est-elle correcte ?

.../2

Question 11 Expliquez la figure 4 de façon détaillée. Évitez des explications trop générales qui ne sont pas spécifiques pour la figure en question.

.../8