M1 MABS – Examen de Fouille de Données – Janvier 2012

durée : 2h

Documents et téléphone portable interdits

Clustering /4.5

1. Qu'est ce que le clustering ? Donnez le principe général des méthodes, c'est-à-dire ce qu'elles cherchent à accomplir ?

2. Pourquoi normaliser les données avant de réaliser un partitionnement ?

3. Citez les différents types de données que l'on peut rencontrer et les mesures de dissimilarité que l'on peut utiliser pour chacun d'eux.

4. Citez les méthodes et leur principe pour le calcul de dissimilarité entre 2 ensembles d'objets (2 clusters).

5. Comment déterminer k pour les méthodes k-means ou k-médoïdes ? A partir de quelle mesure ?

Règles d'association /6

6. Donner les formules du support et de la confiance d'une règle d'association.

7. Que peut-on reprocher à ces mesures ?

8. Quelles autres mesures connaissez-vous ?

9. Citez les méthodes que vous connaissez pour la recherche d'itemsets fréquents.

10. Laquelle est la plus efficace ? Pourquoi ?

11. Appliquez la première vue en cours à la base de transactions suivante en détaillant chaque itération. Quels sont les itemsets fréquents pour un support minimal de 20% ?

A, B, E
B, D
B, C
A, B, D
A, C
B, C
A, C
A, B, C, E
A, B, C
A, B, C, F

12. Quelles sont les règles d'associations que l'on peut extraire de la forme item1, item2 → item3 pour une confiance minimale de 50% ?

Performances et mesures /4.5

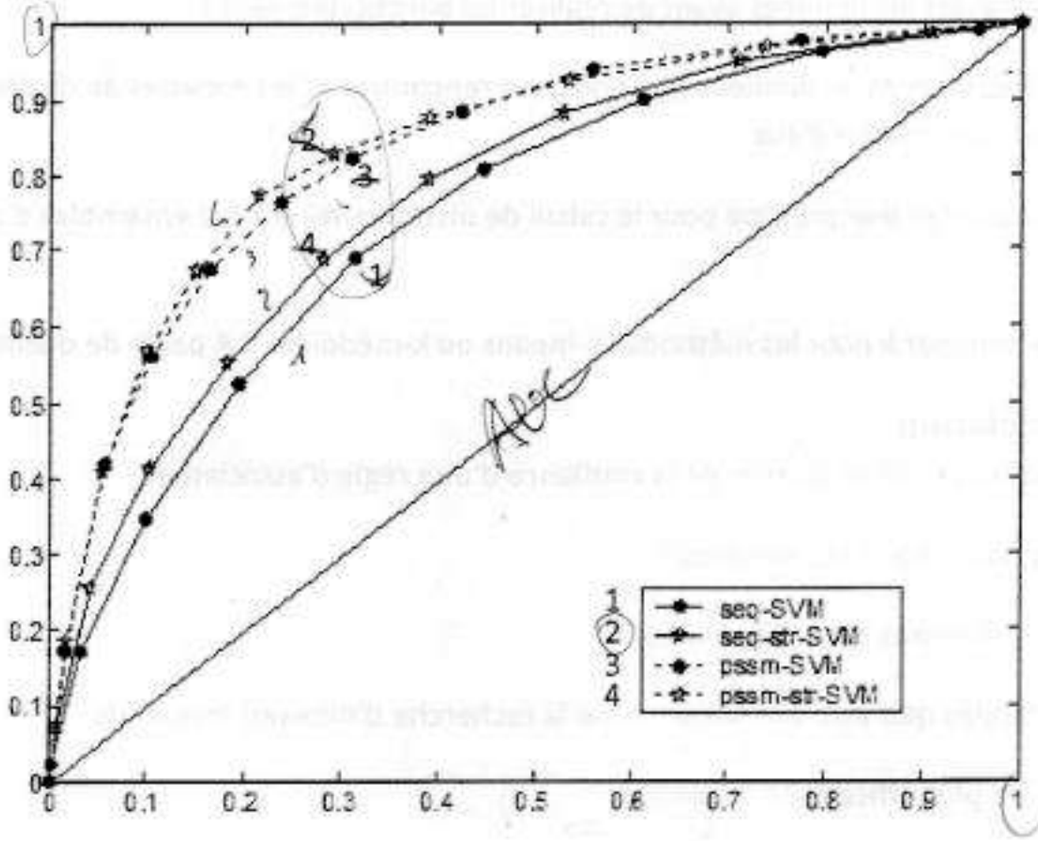
13. Comment peut-on évaluer les performances d'un classificateur, d'une méthode de classification (éventuellement par rapport à d'autres), et de leurs paramètres ?

14. Donnez la définition de la sensibilité et de la spécificité (en français et éventuellement à l'aide de formules).

15. Qu'est-ce que la robustesse ?

16. Vous avez ci-dessous une courbe ROC. Expliquez le principe de son tracé et notamment à quoi correspondent les abscisses et les ordonnées ainsi que la diagonale qui n'est pas dans la légende ? A quoi peut donc servir ce type de graphique ?

17. Les 4 courbes correspondent à différentes méthodes et de jeux de paramètres pour la prédiction de sites de liaison sur une séquence ADN. Laquelle choisiriez-vous pour effectuer une prédiction et pourquoi ?



Méthodes bayésiennes /3

18. Quelle est l'hypothèse faite pour un classificateur bayésien naïf ? Pourquoi est-elle souvent nécessaire ?

19. A partir du jeu de données d'apprentissage suivant, donnez le modèle (table de probabilités) construit par un classificateur bayésien naïf.

Peau	Couleur	Taille	Chair	Classe
poilue	brune	grande	dure	comestible
poilue	verte	grande	dure	comestible
imberbe	rouge	grande	molle	dangereuse
poilue	verte	grande	molle	comestible
poilue	rouge	petite	dure	comestible
imberbe	rouge	petite	dure	comestible
imberbe	brune	petite	dure	comestible
poilue	verte	petite	molle	dangereuse
imberbe	verte	petite	dure	dangereuse
poilue	rouge	grande	dure	comestible
imberbe	brune	grande	molle	comestible
imberbe	verte	petite	molle	dangereuse
poilue	rouge	petite	molle	comestible
imberbe	rouge	grande	dure	dangereuse
imberbe	rouge	petite	dure	comestible
poilue	verte	petite	dure	dangereuse

20. Est-ce qu'il est dangereux de manger un animal à la peau poilue, de couleur rouge, de grande taille, à la chair molle ? Bien entendu, vous argumenterez votre réponse.