

Fouille de Données

Gilles Marcou et Nicolas Lachiche

Durée : 2 heures
Documents autorisés

Toutes les questions doivent trouver une réponse en 5 lignes max. Par ailleurs, ces questions n'ont pas nécessairement une solution standardisée unique et il sera fait attention à la cohérence de la réponse.

Fouille de données et Apprentissage de concepts

1. Expliquez les différences entre apprentissage supervisé et apprentissage guidé par l'utilisateur.

Apprentissage à base d'instances et apprentissage bayésien

1. Comment peut-on gérer les attributs non-pertinents dans les approches à base de distances ?
2. Dans quel cas peut-on dire qu'une approche telle que le plus proche voisin fournit une explication ?
3. Les descripteurs moléculaires fragmentaux composés de sous-graphes d'une molécule ($C=C-O$, $C-C-O$, $C=C(-O)-C$, $C-C$, etc.) constituent-ils des attributs indépendants ?
4. Pour quelles raisons un classifieur bayésien naïf fonctionne-t-il même quand l'hypothèse d'indépendance des attributs étant donnée la classe n'est pas satisfaite ?

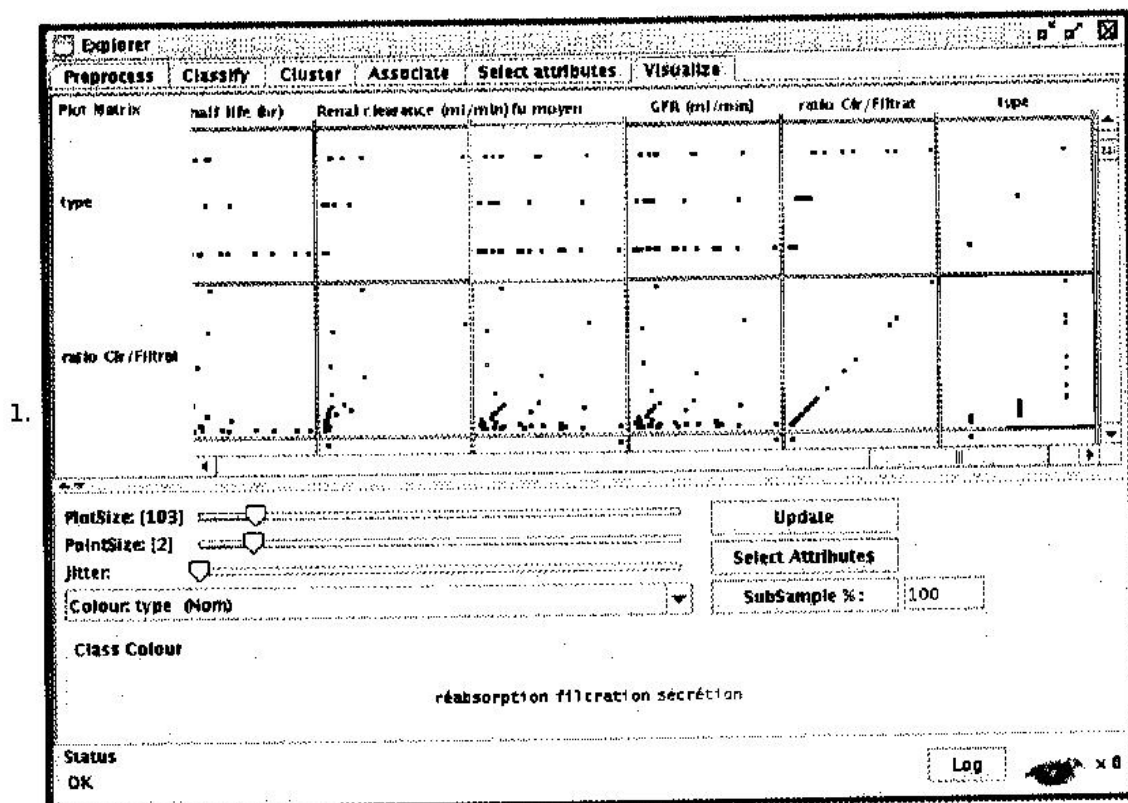
Arbres de décision

1. Pourquoi un arbre de décision serait-il plus lisible qu'un ensemble de règles ?
2. Pourquoi élague-t-on un arbre de décision ?
3. Quels sont les différents "coûts" dont on peut tenir compte dans un domaine médical lorsque l'on choisit les attributs les plus pertinents ?

SVM et Réseaux neuronaux

1. Dans le cas linéaire, et en deux dimensions, quel type de fonction apprend une SVM ? Représentez le graphiquement.
2. En une dimension, quelle est la sortie d'un réseau de fonctions à bases radiales ? Représentez graphiquement l'approximation d'une fonction continue par un ensemble de gaussiennes.

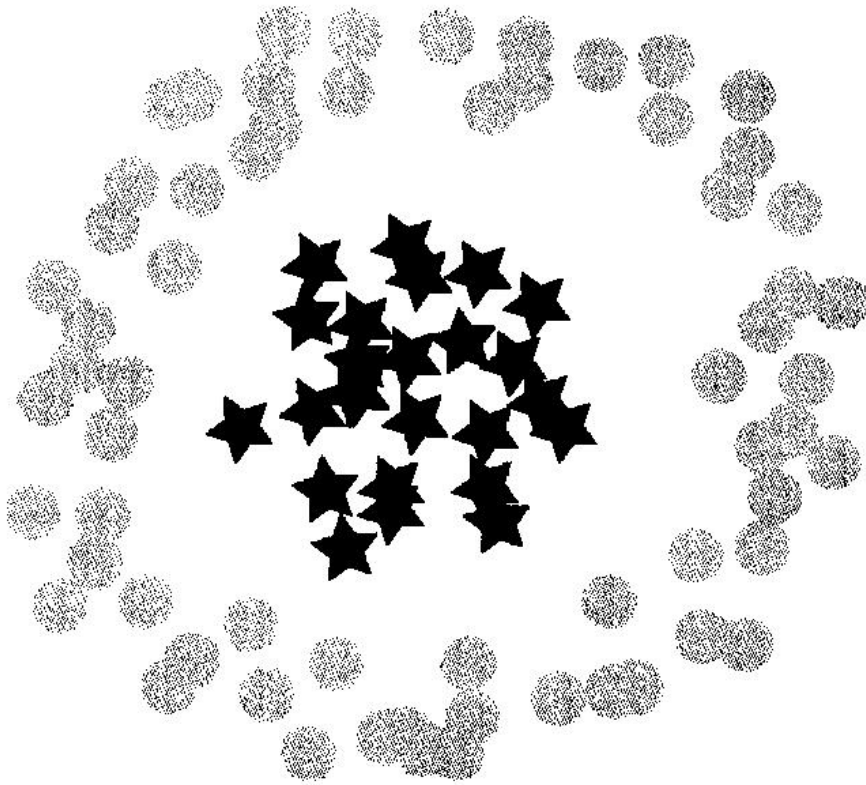
Pratique



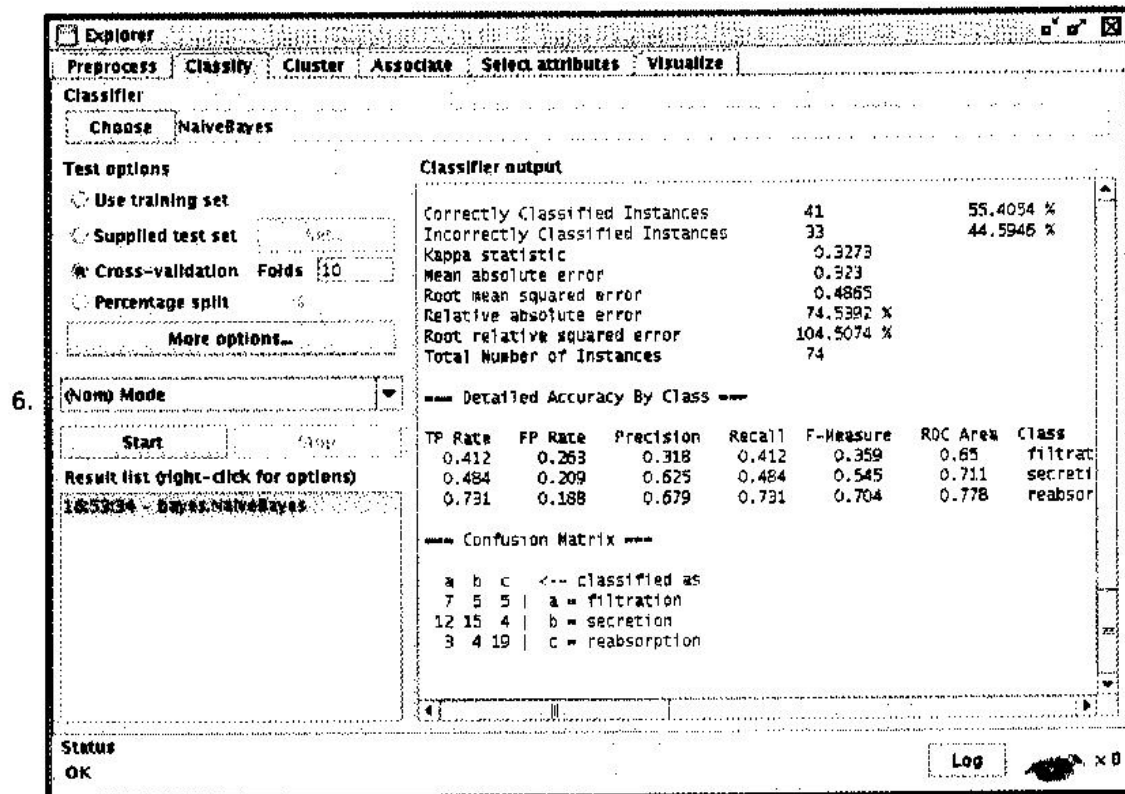
Les données concernant l'élimination rénale de médicament sont analysées dans Weka. Que pouvez-vous dire des attributs représentés dans cette capture d'écran de Weka?

2. Que se passe-t-il si vous utilisez le filtre **RemoveFolds** avec les paramètres *fold* et *numfold* valant respectivement 1 et 10, le paramètre *invertSelection* étant faux?
3. Weka fait fréquemment usage d'un paramètre nommé *Seed*. De quoi s'agit-il?
4. Citez une méthode de clustering non-hiérarchique et une hiérarchique disponibles dans Weka.

5.



Un jeu de donnée est structuré en deux familles la première constituant un anneau autour de la seconde. Cette situation est illustrée ici. Quel comportement attendez-vous de l'algorithme k-mean?



Après avoir réalisé une classification du mode d'élimination rénale de médicaments à l'aide de descripteurs chemoinformatiques, Weka présente les résultats reproduits ici. Comment interprétez vous la matrice de confusion?

7. Que signifie la colonne ROC Area dans la sortie de Weka reproduite plus haut?
8. Dans la méthode Weka MSP, que commande le paramètre *BuildRegressionTree*?
9. La méthode MultiLayerPerceptron de Weka permet de construire de façon automatisé des réseau de neurones. Quelle valeur donneriez-vous au paramètre *hiddenLayers* pour construire un réseau contenant deux couches cachées, l'une de 5 neurones et l'autre 3?
10. Pour quelle raison, la méthode MultiLayerPerceptron de Weka inclut-elle des paramètres concernant un jeu de données de validation?
11. Ayant créé un problème linéairement séparable sur l'applet de *libsvm*, vous échouez à parvenir à séparer les données, quelque soit le noyau choisi et les paramètres de ce dernier. Avez-vous une explication?