Introduction à WEKA

3 avril 2015

Ces exercices introduisent WEKA et vous permettejnt d'essayer plusieurs machines d'apprentissage et de data mining, ainsi que des méthodes de pré-traitement et de visualisation de données.

1 WEKA Explorer

— Téléchargez *weka.zip* sur le site de l'UE
— Exécutez weka par la commande *java -jar weka.jar*
— Téléchargez les jeux de données disponibles
— Ouvrez le dataset *weather.nominal* afin de comprendre le format ARFF

Preprocessing :
— Chargez le dataset *weather.nominal*
— Appliquez un filtre afin de supprimer des attributs

Visualisation :
— Chargez le fichier *iris*
— Cliquez sur *Visualize All* afin de comprendre la distribution des données
— Allez dans l'onglet *Visualize* afin de voir la matrice de distribution
— Créez un ensemble de données ne contenant plus que des instances de deux catégories en sélectionnant l'un des éléments de la matrice, puis en faisant une sélection par rectangle

Classification :
— Prenez le dataset *weather*
— Classifiez ce dataset à l'aide de l'arbre de décision (J48) en utilisant les même données en train et en test
— Examinez l'arbre produit
— Visualisez l'arbre (boutton droit)
— Inteprétez la *classification accuracy* et la *confusion matrix*
— Testez le classifier sur l'ensemble de test

2 Classification

Nous allons travailler sur les datasets *glass*

,

k-plus proches voisins
— Chargez *glass* et faites une classification (k=1) par cross-validation
— Repetez en utilisatn k=10 puis k=20

— Faire la même chose sur les datasets *glass-minusatt* et *glass-withnoise*
— Interprétez les résultats

J48

— Faites tourner J48 sur *glass*
— Visualisez l'arbre produit et simulez sa prédiction sur une instance
— Quelles es l'erreur produite ?
— Faites la même chose sur les autres jeux de données
— Qu'en pensez vous ?
— Comparez les performances entre le kNN et J48

3 Selection de features

Discrétisation :
— Chargez le dataset *sick*
— Classifiez avec J48 et notez la performance
— Dans 'Preprocess', utilisez le 'unsupervised → Discretize' pour discretiser en 10 bins
— Ré-entrainze le modèle et notez la performance
— Refaire la même chose pour 3 bins
— Quelle performance obtenez vous ?

Features Selection :
— Chargez le dataset *mushroom* et appliquez J48 et IBk avec cross-validation
— Sélectionnez les attributs en utilisant CfsSubsetEval et GreedyStepwise
— Quels résultats obtenez vous ?
— Utilisez AttributeSelectedClassifier (avec CfsSubsetEval et GreedyStepwise search) pour les classifiers J48 et IBk et évaluez par corss-validation
— Qu'en déduisez vous

4 Règles d'association

— Chargez le dataset *vote*
— Appliquez *APriori*
— Que signifient les règles extraites ?
— Changez le nombre de règles extraites
— Trouvez des règles 'qui font sens'
— Faire la même chose sur *supermarket*