PRESENTATION
DU LOGICIEL DE DATA-MINING

SEANCE DE TRAVAUX DIRIGES

26 octobre 2016

Auteurs :

- ADECHINA Charlemagne
- HOLZWARTH Marc
- MERAD Manel
- YABAS Eric

1. Présentation de l'exercice

L'exercice est basé sur un jeu de données IBM « HR Employee Attrition and Performance » https://www.ibm.com/communities/analytics/watson-analytics-blog/hr-employee-attrition/ visant à prédire l'attrition des employés en fonction de différentes caractéristiques :

Age	JobInvolvement	PerformanceRating
BusinessTravel	JobLevel	RelationshipSatisfaction
DailyRate	JobRole	StandardHours
Department	JobSatisfaction	StockOptionLevel
DistanceFromHome	MaritalStatus	TotalWorkingYears
Education	MonthlyIncome	TrainingTimesLastYear
EducationField	MonthlyRate	WorkLifeBalance
EmployeeCount	NumCompaniesWorked	YearsAtCompany
EmployeeNumber	Over18	YearsInCurrentRole
EnvironmentSatisfaction	OverTime	YearsSinceLastPromotion
Gender	PercentSalaryHike	YearsWithCurrManager
HourlyRate		

L'objectif de l'exercice est de réaliser différents modèles de prédiction de l'attrition (problème de classification) et d'estimer leur performance.

Définition attrition : https://fr.wikipedia.org/wiki/Attrition

2. Réalisation

Etape 1 – Préparation

1. Télécharger le fichier Excel « WA_Fn-UseC_-HR-Employee-Attrition.xlsx » depuis le lien ci-dessus, puis lancer Weka.

2. Le fichier téléchargé est au format Excel, qui n'est pas un format géré nativement par Weka. L'installation d'un package supplémentaire est requis. Ouvrez le gestionnaire de packages (Menu « Tools »), chercher « Excel », sélectionner le package « WekaExcel » puis lancer l'installation. Une fois l'installation terminée, fermez le gestionnaire de package.

Etape 2 – Découverte et classification via l'interface Explorer

1. Choisir le module « Explorer », puis charger le fichier « WA_Fn-UseC_-HR-Employee-Attrition.xlsx ».

2. Quel est la moyenne d'âge et l'écart type des individus observés ?

3. On souhaite maintenant visualiser la répartition des effectifs dans la classe âge en fonction de l'attrition (Liste de sélection au-dessus du graphique). Quelle constatation peut-on faire ?

4. Pour réaliser un premier modèle de prédiction, la variable « Attrition » doit être la dernière. Choisissez et appliquez le filtre « Reorder » (unsupervised → attribute, paramètre d'indices = « 1,3-last,2 »). Une fois la variable déplacée, enregistrer les données dans un nouveau fichier au format Weka «.arff » (Nous utiliserons le jeu de données avec l'extension .arff pour la suite du TD.

5. Aller dans l'onglet « Classify » puis sélectionner l'algorithme « NaiveBayes » et lancer l'apprentissage sur 66% des données. Combien d'individu sont mal classés ? Quel est le pourcentage d'individus correctement classés ?

Etape 3 – Classification via l'interface graphique KnowledgeFlow

1. On souhaite maintenant créer des modèles de classification via l'interface graphique, choisissez le module « KnwoledgeFlow » de la fenêtre principale Weka.
   NB : A chaque ajout d'un composant/lien dans l'interface, il sera nécessaire de réexécuter l'ensemble du modèle via le bouton « Run » en haut à gauche de la fenêtre.

2. Pour charger le fichier Arff de données, ajouter un composant ArffLoader, puis indiquer le chemin d'accès au fichier dans la configuration.

3. Pour pouvoir générer le modèle de prédiction, Weka doit connaître la classe sur laquelle faire la prédiction. Ajoutez le composant ClassAssigner (groupe Evaluation), puis configurer le en indiquant la classe « Attrition ».

4. Ensuite, séparer le fichier en échantillon d'apprentissage et de test (composant TrainTestSplitMaker à relier avec le dataSet). Weka décompose automatiquement le fichier avec 66% en apprentissage et 34% en test.

5. Générer le modèle de classification via l'algorithme d'arbre de décision J48 (reliez le trainingSet et le testSet successivement sur le composant), avec un minimum de 15 individus par feuille (minNumObj). Visualiser ensuite l'arbre de décision (composant GraphViewer). Quelle est la variable la plus corrélée avec l'Attrition ?

6. Maintenant, on souhaite visualiser la performance du modèle créé (connecter le composant J48 via le composant ClassifierPerformanceEvaluator avec la connection BatchCassifier et relié ce dernier à un TextViewer). Quel est le pourcentage d'individus correctement classés ? Quel est le taux d'erreur du modèle ?

7. On souhaite maintenant comparer ce modèle avec un modèle basé sur la régression logistique. Ajoutez le composant Logistic (à relier avec le composant TrainTestSplitMaker) et visualiser les résultats (TextViewer).
   - Quelle est la variable qui influence le plus négativement sur le modèle ? Positivement ?
   - Analyser l'odd-ratio entre l'âge et l'attrition. Peut-on dire que l'âge est indépendant de la non-attrition ?

8. Répéter l'étape 6. Quel est le pourcentage d'individus correctement classés ? Quel est le meilleur des trois modèles ?

Etape 4 – Comparaison des performances de modèles de classification via l'interface Experimenter

1. On voudrait maintenant comparer les performances de ces modèles de classification à d'autres modèles mais de manière plus rapide. Pour cela, on va utiliser le module « Experimenter » de Weka (accessible depuis la fenêtre principale).

2. Dans l'onglet « Setup », cliquez sur new puis ajouter le dataset des données RH au format arff issue de l'étape 2.

3. Ajouter les algorithmes d'apprentissage à tester suivants :
   - ZeroR, paramètres par défaut
   - NaiveBayes, paramètres par défaut
   - J48, avec un minimum de 15 individus par feuille (minNumObj)
   - Logistic, paramètres par défaut
   - SGD, paramètres par défaut
   - RandomForest, paramètres par défaut

4. Allez dans l'onglet « Run » et lancer l'exécution. Une fois le traitement terminé, allez dans l'onglet « Analyze ».
   NB : Chaque algorithme est exécuté 10 fois (paramètre NumFolds précédent) sur le jeu de données, avec un principe de « cross validation » qui découpe le jeu de données en 10 parties utilisées successivement comme training/test set (paramètre NumFolds précédent)

5. Nous allons maintenant procéder à l'analyse comparative des algorithmes. Cliquez sur « Experiment », puis sur « Perform test ». Weka affiche alors les pourcentages de classification correcte pour chaque algorithme (avec un niveau de confiance à 5%). Quel algorithme présente les meilleures performances ? Les moins bonnes ?

3. Récapitulatif

Lors de ce TD, vous avez appris à utiliser les fonctionnalités suivantes de WEKA :

Module	Fonctionnalités
Package manager	▪ Installer un nouveau package
Explorer	▪ Reformater le jeu de données ▪ Consulter des données de statistique descriptive ▪ Générer un modèle de classification
KnowledgeFlow	▪ Générer des modèles de classification via la conception graphique d'un enchainement de plusieurs étapes
Experimenter	▪ Comparer rapidement les performances de plusieurs algorithmes de classification sur un même jeu de données

Pour approfondir l'utilisation de WEKA, vous pouvez consulter les vidéos du MOOC officiel (très bien fait) disponible en ligne sur la chaîne Youtube https://www.youtube.com/user/WekaMOOC