TD ORANGE DATA MINING

SISE 2016/2017

Les données utilisées pour le TD sont tirées du dataset Zoo proposé par le site UCI Machine Learning repository [1].

Pour une meilleure compréhension nous avons modifié le jeu de données. Celui-ci est téléchargeable à l'adresse : https://we.tl/OOtwGMrmHn

Vous pouvez vous servir de la documentation du logiciel pour réaliser les exercices :
http://docs.orange.biolab.si/3/visual-programming/index.html#widgets

Exercice 1 - Importation et visualisation des données

1. Lancez Orange Canvas et cliquez sur **New** pour créer un nouveau workflow.
2. Téléchargez le fichier zoo_type.csv à l'adresse indiquée ci-dessus. Importez les données, paramétrez la variable "*type*" en *nominal* et placez-la en *target.* (widget File).
3. Visualisez le contenu du fichier. Combien y a-t-il de variables quantitatives, qualitatives, d'individus ? (Voir Data Info, Data Table).
4. Combien y a-t-il d'animaux venimeux (venimous) ? (cf. Visualize).

Exercice 2 - Méthodes non supervisées

1. Analyse en Composantes Principales
    a. Dans l'espace de travail, cliquez sur le widget d'importation des données pour re-définir la nature des variables. Définissez toutes les variables en *numeric, à* l'exception du *type*.
    b. Ajoutez un widget de sélection des colonnes en excluant la variable *legs* étant donné que cette variable est numérique contrairement aux autres qui sont binaires.
    c. Effectuez une ACP et sélectionnez le nombre d'axes (catégorie Unsupervized).
    d. Calculez les coordonnées des individus sur les axes factoriels, puis les corrélations des variables avec les axes (Data Table). Quelle interprétation pouvez-vous donnez aux deux premières composantes ?
    e. Affichez le graphique des individus et colorez les points selon leur type. (Scatter Plot).
    f. Faites le parallèle entre les variables et les individus, et interprétez les axes factoriels.
2. Classification ascendante hiérarchique
    a. Calculez la distance entre les individus en utilisant la distance de Jaccard (appropriée dans le cadre de variables qualitatives binaires) (Catégorie Unsupervized). Reliez ce widget à Select Columns.
    b. Visualisez les distances entre les individus (Distance Matrix).

c. A partir des distances calculées précédemment, construisez une classification ascendante hiérarchique en utilisant le critère de Ward. Pour une meilleure lecture des résultats, vous pouvez afficher le nom des animaux sur le dendrogramme.
d. Choisissez la coupure optimale (Sélection : Height Ratio).
e. Pour visualiser la composition des groupes, jouez sur la sélection des classes sur le dendrogramme et utilisez le widget Image Viewer.
f. Projetez les individus dans le premier plan factoriel et colorez les points selon les clusters. Comparez avec le graphique des individus obtenu en 1.e. La classification obtenue semble-t-elle meilleure que la typologie déjà existante ?

Exercice 3 - Méthodes supervisées

1. Ajoutez un Data sampler, et reliez-le au widget Select Columns. Fixez la proportion de l'échantillon d'apprentissage à 70%.
2. Ajoutez un arbre de classification et reliez-le au Data Sampler. Utilisez l'indicateur "entropy".
3. Affichez l'arbre (Classification Tree Viewer). Donnez au moins une règle de décision.
4. Jouez avec la sélection des classes pour visualiser les individus (Image Viewer).
5. Ajoutez autant de méthodes supervisées que vous le souhaitez (Random Forest, SVM, Règles bayésiennes…)
6. Evaluez la qualité de vos modèles avec un Test & Scores *(cf.* [3], section Example) (Indice : reliez le Data Sample à Data et Test Data)
7. Ajoutez une courbe ROC et faites varier les classes dans l'option Widget Class
8. On va maintenant essayer de mettre en évidence les individus qui ont été prédits à tort par les modèles. Pour cela, créez autant de matrices de confusion que de méthodes supervisées et reliez-les au Test & Score.
9. Ajoutez un diagramme de Venn relié à chacune des matrices de confusion, puis sélectionnez dans les matrices les observations mal prédites. Avez-vous des animaux prédits à tort par plusieurs modèles à la fois ? Lesquels ?

[1] https://archive.ics.uci.edu/ml/datasets/Zoo
[2]https://www.xlstat.com/fr/solutions/fonctionnalites/classification-ascendante-hierarchique-cah
[3]
http://docs.orange.biolab.si/3/visual-programming/widgets/evaluation/testlearners.html