Algorithmes Avancés

Projet Noté

Ce projet a pour but de faire pratiquer une méthodologie caractéristique de l'approche « Data Science » pour la classification de données. Il se déroulera en trois parties, la première étant focalisée sur le « Clustering » et les deux autres étant liées à la classification supervisée avec des données imposées et libres. Voici la subdivision des différentes parties :

1. Clustering avec les données du *Titanic*.
2. Classification avec des données imposées : *Breast Cancer Wisconsin* (Diagnostic) ; *Wine*. Trois modèles d'apprentissage supervisé seront obligatoirement utilisés.
3. Classification supervisée avec un jeu de données de votre choix, ainsi que trois modèles d'apprentissage.

Des détails supplémentaires sont donnés dans les paragraphes suivants. La liste des classes de *Scikit* se trouve à la page : http://scikit-learn.org/stable/modules/classes.html.

Partie 1.

- Utiliser la méthode des *K-Means* (`cluster.KMeans`) en faisant varier le paramètre K (et éventuellement d'autres paramètres).
- Les données à utiliser sont celles du *Titanic* qui sont dans *CyberLearn* (répertoire *tp*). Vous pouvez modifier le format du fichier s'il ne vous convient pas.
- Les données sont décrites par trois variables (classe billet, âge, sexe) et la classe d'appartenance (rescapé/non-rescapé).
- Le but est de trouver la valeur de K qui minimise la moyenne des distances moyennes entre les barycentres des groupes et les instances des groupes.
- Il faudra afficher un graphique qui présentera en ordonnée cette valeur moyenne et en abscisse la valeur de K. Utiliser *matplotlib* (https://matplotlib.org/).
- Pour le meilleur K il faudra en plus afficher en trois dimensions les exemples des deux classes avec deux couleurs différentes. Il sera aussi possible de réaliser trois visualisations 2D pour chaque couple de variables. Pour chaque affichage il faudra aussi mettre en évidence les barycentres.
- Déterminer si l'on voit quelque chose d'intéressant sur ces graphes 2D-3D (régions dans lesquelles une majorité de points appartient à une seule classe).

Partie 2.

- Utilisation des données *Breast Cancer* et *Wine*. On peut les charger respectivement avec `datasets.load_breast_cancer` et `datasets.load_wine`.

- Certains modèles d'apprentissage « souffrent » fortement des facteurs d'échelle ; il est donc conseillé de normaliser les données. Spécifiquement, regarder dans `sklearn.preprocessing`.
- Le but est de déterminer par validation croisée à 5 segments le meilleur modèle d'apprentissage parmi la méthode des K-plus proches voisins (`neighbors.KNeighborsClassifier`), les arbres de décision (`tree.DecisionTreeClassifier`) et le Perceptron multi-couche (`neural_network.MLPClassifier`). La procédure de validation croisée doit être répétée 10 fois pour chaque sélection de paramètres.
- Pour la validation croisée regarder dans le document 06 (*Cyberlearn*) les diapos 42 et 43, puis dans `sklearn.model_selection`.
- Pour les arbres, regarder http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. Il est demandé de faire varier le paramètre `min_samples_leaf`.
- Pour le Perceptron multi-couches : http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html. C'est un peu plus délicat. Fixer la fonction d'activation (de votre choix), puis définir quelques architectures avec au maximum deux couches cachées. Pour éviter le sur-apprentissage utiliser l'option `early_stopping` avec la valeur `True`.

Partie 3.

- Les données sont à choisir sur le site https://archive.ics.uci.edu/ml/datasets.html.
- Il doit y avoir au moins deux classes, au moins dix variables et au moins 200 données. Evitez de prendre un ensemble de données avec plus de 20000 examples. Toute exception violant ces contraintes reste possible sur demande à l'enseignant.
- Il faudra comparer trois modèles d'apprentissage de votre choix avec une procédure de validation croisée pour déterminer le meilleur modèle (voir partie 2).

Ce projet pourra être réalisé en binôme. Un rapport et le code devront être envoyés à l'enseignant par e-mail le **10 juin**, au plus tard. Vous devrez défendre votre projet par une présentation orale de 20 minutes. Préparez donc quelques diapos illustrant vos résultats. Les présentations commenceront lundi 11 juin et continueront le 18 juin (selon l'horaire du cours) et peut-être un autre jour, selon une plage horaire supplémentaire à déterminer.