Machine Learning

Projet noté (première partie)

Ce projet a pour but de faire pratiquer les méthodologies caractéristiques de l'approche « Data Science » pour la classification de données. Il se déroulera en trois parties spécifiques associées aux modèles suivants :

1. modèles de base vus au cours ;
2. ensembles ;
3. réseaux convolutifs.

Pour cette première partie un ensemble de données est imposé et un autre est libre. Les données imposées s'appellent *wisconsin* (diagnostic de cancer du sein) : https://sci2s.ugr.es/keel/category.php?cat=clas.

Les données libres peuvent se trouver sur le site précédent ou dans https://archive.ics.uci.edu/ml/index.php. 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 exemples. Toute exception violant ces contraintes reste possible sur demande à l'enseignant.

Pour chaque ensemble de données il faudra comparer trois modèles de classification avec *Scikit Learn* :

1. Le plus proche voisin (**PPV**) ;
2. Les arbres de décision (**AD**) ;
3. Le Perceptron Multi Couche (**PMC**).

Certains modèles d'apprentissage « souffrent » fortement des facteurs d'échelle ; il est donc conseillé de normaliser les données. Spécifiquement, regarder dans `sklearn.preprocessing`.

L'évaluation des modèles se fera en utilisant la procédure de validation croisée répétée dix fois (cf. *Cours04 - Apprentissage supervisé PPV.pptx*, diapo 38). Dans *Scikit Learn* regarder dans `sklearn.model_selection`.

Notez que tous les modèles présentent un certain nombre de paramètres. Comment choisit-on ces paramètres ? Par exemple pour le PPV on pourrait comparer le TCC (taux de classification correct sur l'ensemble d'apprentissage et sur l'ensemble de test) pour plusieurs valeurs du paramètre définissant le nombre de voisins à utiliser.

La liste des classes de *Scikit* se trouve à la page : http://scikit-learn.org/stable/modules/classes.html.

Pour le modèle **PPV** : https://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html. Pour les **AD** : http://scikit-learn.org/stable/modules/generated/sklearn.tree.DecisionTreeClassifier.html. Pour les **PMC** : http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.

Ce travail est noté ; il devra être déposé dans *Cyberlearn*. Une archive contenant le rapport et le code devra avoir un nom indiquant les deux auteurs, si le travail est effectué en binôme. Par exemple : Kunzli_ Pierre_et_Bologna_Guido.zip.

Date de rendu : au plus tard le **lundi 9 décembre**.