Machine Learning

Projet noté (première partie)
Modèles de base et ensembles de modèles

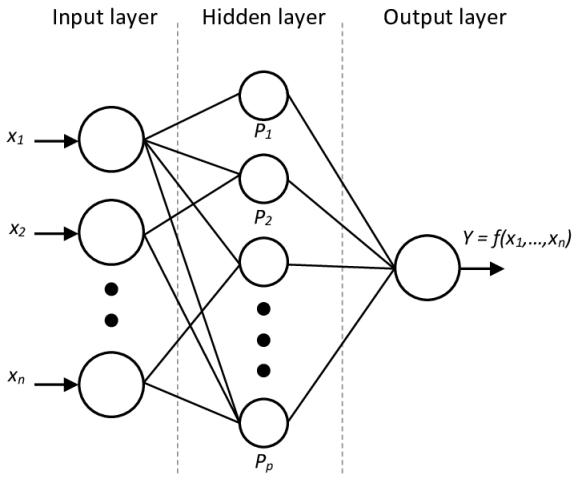
Ce projet a pour but de faire pratiquer les méthodologies caractéristiques de l'approche « Data Science » pour la classification de données. Trois exercices devront être réalisés avec les modèles suivants :

1. Perceptrons Multi-Couches (PMC) ;
2. ensembles de modèles ;
3. réseaux convolutifs.

Pour cette première partie un jeu de données est imposé et un autre est libre. Les données imposées sont dénommées *heart* (maladie du cœur ; 270 données) : https://sci2s.ugr.es/keel/category.php?cat=clas. Les modèles devront être évalués par validation croisée.

Exercice 1.

Dans cet exercice il faudra programmer (un peu) ; ce que vous avez réalisé dans la série 3 sera utile. Vous allez utiliser des PMC avec une seule couche intermédiaire. Au début de l'apprentissage tous les poids seront initialisés aléatoirement dans l'intervalle -0.5...0.5 ; ensuite il n'y aura que les poids entre la couche intermédiaire et la couche de sortie qui pourront varier par la règle du Perceptron.



Questions :

a) Pour les deux jeux de données, évaluer le taux de classifications correctes par validation croisée pour un nombre de neurones cachés égal à : 10 ; 50 ; 100 ; 200 ; 500 ; 1000. Il faudra donner la moyenne de chaque validation croisée et l'écart type.

b) Est-ce que le taux de classifications correctes est bien au-dessus de la proportion de la classe majoritaire ?

c) Avec la librairie *Scikit*, créer des PMC avec une seule couche cachée et répondre à a) en utilisant un nombre de neurones cachés égal à 5 ; 10 ; 20. En outre, pour chacune de ces configurations il est demandé de choisir un nombre d'itérations égal à 500 ; 1000 ; 2000.

Exercice 2.

Programmer le Bagging pour les PMC dont les poids entre la couche d'entrée et la couche cachée ne varient pas. Dans cet exercice les données sont de votre choix. Chaque ensemble de modèles contiendra 25 PMC contenant le même nombre de neurones cachés. Par expérimentation vous pourrez déterminer le nombre de neurones cachés et le nombre d'itérations pour l'entraînement.

Questions :

a) Evaluer le taux de classifications correctes par validation croisée (moyenne, écart type)

b) Avec la librairie *Scikit*, créer des *Random Forests* et comparer les résultats par validation croisée.

Indications (à lire impérativement).

- Vous allez vous familiariser avec la librairie *scikit-learn* (Python) disponible sur les machines du 4è étage ou sur internet. La distribution *conda* est particulièrement conseillée (voir https://scikit-learn.org/stable/install.html).

- 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.

- 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. *Cours03 - Apprentissage supervisé PPV.pptx*, diapo 38).

- La liste des classes de *Scikit* se trouve à la page : http://scikit-learn.org/stable/modules/classes.html.
- Pour les **PMC** : http://scikit-learn.org/stable/modules/generated/sklearn.neural_network.MLPClassifier.html.
- Pour les **Random Forests** : https://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html.

Rendu.

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 : Leblanc_ Quentin_et_Bologna_Guido.zip.

Dans le **rapport** il n'est pas demandé d'expliquer les modèles. Il sera simplement question de montrer les résultats expérimentaux. On s'attend à voir la description des traitements réalisés sur les données

avant l'apprentissage et puis des tableaux ou des graphiques montrant les résultats obtenus par validation croisée. Enfin, il faudra réaliser une analyse des résultats sous forme de discussion.

Date de rendu : au plus tard le **mardi 14 décembre 2021**.