TP MAT4102 **2023-2024**

Objectif : le but de ce TP est de mettre en pratique les notions vues en cours sur les arbres de décision et les méthodes d'agrégation et de les comparer pour la classification supervisée.

Partie 1 : ouvrir « TP1_DT_ iris_diabete.ipynb » Jupyter Notebook

Dans cette partie, vous exploiterez la base de données « Iris de Fisher » qui a été prise comme exemple en cours. Vous appliquerez d'abord l'algorithme d'arbre de décision pour une classification multi-classes, puis l'algorithme de forêt aléatoire.

Ainsi, il vous sera demandé :

- De prendre en main la base de donnée, de visualiser les variables descriptives entre autres
- D'appliquer les arbres de décision en utilisant les hyper-paramètres établis par défaut ; puis en fixant leurs valeurs. L'objectif est de bien comprendre leur impact sur les performances de classification

Une autre base de donnée « diabetes.csv » sera aussi utilisée. Vous appliquerez un arbre de décision pour une classification binaire dans un but de détection de diabète chez les individus.

https://www.kaggle.com/datasets/uciml/pima-indians-diabetes-database?resource=download

Partie 2 : ouvrir « TP2_RF_ iris_diabete.ipynb » Jupyter Notebook

Dans cette partie, vous appliquerez les forêts aléatoires pour la classification supervisée.

- Comparer les résultats à ceux obtenus avec les arbres de décision
- Etudier l'impact de certains hyper-paramètres en termes de performances
- Utiliser les forêts aléatoires comme outil de sélection de variables les plus pertinentes, puis d'étudier l'efficacité de cette approche pour la classification.

Partie 3 : ouvrir « TP3_DTversus RF_digits.ipynb » Jupyter Notebook

Dans cette partie, vous exploiterez la base de données « digits de MNIST», une version similaire à celle utilisée dans le TP classification non-supervisée. Vous appliquerez d'abord l'algorithme d'arbre de décision pour une classification multi-classes, les forêts aléatoires, puis Adaboost.

Ainsi, il vous sera demandé :

- De prendre en main la base de donnée, de visualiser les images et les variables descriptives entre autres
- D'appliquer les arbres de décision en faisant plusieurs tirages de l'ensemble d'apprentissage et d'évaluer les performances de classification obtenues
- D'appliquer les forêts aléatoires et de comparer les résultats à ceux obtenus avec les arbres de décision
- D'appliquer Adaboost et de comparer les résultats.

Partie 4 : ouvrir « TP4_Bagging_Boosting.ipynb » Jupyter Notebook

Dans cette partie, vous exploiterez la base de données « breast_cancer », contenant des images histologiques de tumeurs bénignes et malignes (problème à deux classes). Une copie de la ase UCI ML Breast Cancer Wisconsin (Diagnostic) est téléchargeable via le lien suivant: https://goo.gl/U2Uwz2

Le but de cette partie est de comparer les arbres de décisions ainsi que les approches d'agrégation basées sur le Bagging et le Boosting.

Ainsi, il vous sera demandé :

- De prendre en main la base de donnée, de visualiser les variables descriptives entre autres ;
- De comparer les résultats des arbres de décision, forêts aléatoires, bagging, boosting ;
- D'étudier l'impact de certains hyper-paramètres de Adaboost et de visualiser ses paramètres.
- Travail supplémentaire : appliquer sur la base « iris de fisher » la méthode Adaboost construite avec des estimateurs type SVM linaires.

Livrable : rapport comprenant ces différentes parties avec une analyse approfondie des résultats.