Consignes pour les cours d'apprentissage supervisé

Cours 1 et cours 2 :

1. lire les diapositives de cours : principes généraux en apprentissage supervisé
2. suivre les vidéos ou lire chapitre 1 et 2 du livre (tout est indiqué en diapo #19)
3. prendre connaissance de jeu de données qui sera utilisé en TP

Cours 3 :

1. lire les diapositives de cours : compléments sur les familles de modèles d'apprentissage supervisé
2. prendre connaissance des documents sur SVM et Ensemble learning

TP1 : voir ci-dessous

Cours 4 et 5 :

1. lire les diapositives de cours : enjeux en apprentissage supervisé
2. prendre connaissance des documents sur interprétabilité, équité et vie privée

TP2 : voir ci-dessous

Evaluation :

Compte-rendu de TP (document pdf) à réaliser en binôme et à déposer sur moodle

Consignes pour les TP d'apprentissage supervisé – 26/11/2023

Objectifs : mettre en place un processus d'apprentissage complet sur le dataset ACSincome pour l'état de Californie (tâche de classification binaire) et analyser les résultats obtenus.

Avant de commencer le TP : créez un environnement virtuel (sous conda en salle de TP) en suivant la même procédure que celle indiquée pour les TP de clustering. Ajoutez les packages `numpy`, `scipy`, `matplotlib`, `scikit-learn` dans ce nouvel environnement.

Attendus Partie 1 :

(0) Comprendre et préparer les données

- En pratique, le dataset est gros, il peut y avoir besoin de ne travailler que sur une fraction du dataset au moins pour mettre au point le code (puis de lancer sur l'ensemble du dataset si possible). Par exemple avec un shuffle du dataset puis sélection d'une fraction (10%, 20%, etc)

```
from sklearn.utils import shuffle
X_all, y_all = shuffle(X_all, y_all, random_state=1)
# only use the first N samples to limit training time
num_samples = int(len(X_all)*0.1)
X, y = X_all[:num_samples], y_all[:num_samples]
```

- Standardisez les données si besoin
- Séparez train set et test set

Les méthodes d'apprentissage étudiées dans ces TP sont les suivantes : SVM, RandomForest, AdaBoost et GradientBoosting. Il est important d'avoir compris le principe de ces méthodes. Pour chaque méthode, vous devez :

(1) Mettre en place une validation croisée
(2) Evaluer la qualité d'un modèle d'apprentissage en utilisant différentes métriques (accuracy, classification_report, confusion_matrix)
(3) Mettre en place une recherche des bons hyperparamètres (gridsearchCV)

Puis

(4) Analysez et comparez les résultats des différents modèles d'apprentissage pour le compte-rendu de TP
(5) Peut-on avoir des prédictions pertinentes à partir des modèles appris sur les données de l'état de Californie pour les états du Nevada et du Colorado ?

Attendus Partie 2 (26/11/2023)

Les données complémentaires nécessaires pour la suite sont fournies sur la page moodle.

(1) **Explicabilité des modèles.** Considérez les meilleurs modèles d'apprentissage obtenus dans la partie 1 pour chaque famille de modèles testés (SVM, Random Forest, Adaboost, Gradient Boosting)
   a. A partir des données d'entrainement (features et labels) calculez les corrélations entre chacune des features et le label.
   b. Calculez les corrélations entre chacune des features et le label prédit. Comparez avec les corrélations précédentes
   c. Evaluez l'importance de chaque feature en utilisant la méthode `permutation_importance` de scikitlearn. Vous pouvez également récupérer les informations fournies (lorsque c'est le cas) par les différents modèles
   d. Que vous apportent ces analyses dans la compréhension des prédictions obtenues ?

(2) **Equité des modèles.** Considérez les meilleurs modèles d'apprentissage obtenus dans la partie 1 pour chaque famille de modèles testés (SVM, Random Forest, Adaboost, Gradient Boosting).
    a. Considérez tout d'abord que la feature 'SEX' est une feature sensible. Calculez la matrice de confusion pour chaque valeur de cet attribut (et pour chaque modèle)
    b. Calculez 2 métriques d'équité statistique (en train et en test) et déerminez l'écart entre les différents individus du jeu de données.
    c. Recommencez le processus en refaisant un apprentissage à partir des données dans lesquelles la feature 'SEX' est retirée. Donnez ensuite les valeurs d'équité statistique.
    d. Que vous apportent ces analyses dans la compréhension du jeu de données et sur l'équité statistique par rapport au genre ?
    e. Analysez une ou deux métriques d'équité statistique (en fonction de la feature SEX du dataset,"male" or "female") en train et en test en gardant ou en supprimant cette feature SEX du dataset
    f. Reprendre le travail effectué en considérant que la feature sensible est la feature 'RAC1P'.
(3) Synthétisez vos analyses pour le compte rendu de TP