TD Apprentissage supervisé à partir de données personnelles

L'objectif de ce TD est d'illustrer les enjeux de l'apprentissage supervisé à partir de données personnelles. Concrètement, nous allons essayer d'estimer si il est possible de prédire le sexe et l'animal préféré d'une personne à partir de données personnelles numériques simples (collectées en classe) telles que la taille, l'âge, la pointure de chaussure, etc.

Pré-requis: cours intro ML =
http://www-sop.inria.fr/members/Alexis.Joly/01_Linear_Lasso.pdf
http://www-sop.inria.fr/members/Alexis.Joly/02_CrossValidation_slides.pdf
http://www-sop.inria.fr/members/Alexis.Joly/03_Linear_Lasso.pdf

1. **Collecte de données**
    a. Quelles données personnelles seraient d'après vous utiles pour prédire le genre d'une personne (homme ou femme) ? Ou bien son animal préféré ? Ou son nombre d'amis sur facebook ?
    b. Remplissez le sondage: https://framaforms.org/td-personal-data-1602059191

    c. Télécharger les données: http://www-sop.inria.fr/members/Alexis.Joly/data-2020.csv

2. **Environnement de travail et chargement des données**
    a. Téléchargez le notebook sur Google collab (Fichier>télécharger le fichier .ipynb): https://colab.research.google.com/drive/1in7KLJD192u3THJxJ6eAaIW5UjRfu5jH
    b. Ouvrir le fichier .ipynb via jupyter-notebook
    c. Exécutez la cellule "Import data" avec le bon nom de "input_file". A quoi servent les différentes commandes ?

3. **Apprentissage d'un modèle de régression linéaire (moindres carrés)**
    a. Ecrivez un code permettant de découper les données en 75% train / 25% test en utilisant la fonction train_test_split de sklearn.model_selection
    b. Ecrivez un code pour apprendre un modèle de régression linéaire pour prédire la taille à partir des autres variables. Quel est le coefficient de régression sur l'ensemble d'apprentissage et sur celui de test (utiliser la fonction .score() de

`LinearRegression)` ? Quels sont les poids du modèle associés à chaque variable d'entrée (utiliser .coef_) ?

c. Faites de même pour le nombre d'amis sur facebook
d. Ré-essayer avec un nouveau split aléatoire. Que constatez-vous ? Peut-on faire confiance aux modèles entraîner ?

4. **Sélection de variables avec RFE**
    a. Ecrivez un code permettant de sélectionner les 2 variables les plus explicatives avec l'algorithme Recursive Features Elimination: https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_digits.html (pour la taille et le nb d'amis facebook)
    b. Que fait l'algorithme RFE ? Le modèle obtenu est-il meilleur que le modèle linéaire complet sur les données de test ?
    c. Quelles les sont les 2 variables sélectionnées par RFE et quel est le poids asssocié (utiliser rfe.ranking_ et rfe.estimator_.coef_) ? Cela vous parait-il logique (pour la taille ? pour le nombre d'amis facebook) ?
    d. Essayer avec 1 variable ou 3 variables. Que pourrait-on faire pour choisir le bon nombre de variables à conserver ?
    e. Ré-essayer avec un nouveau split aléatoire. Que constatez-vous ? Peut-on faire confiance aux modèles entraîner ?

5. **Lasso & Ridge Path**
    a. Ecrivez un code permettant de calculer et d'afficher les chemins Lasso et/ou Ridge et/ou ElasticNet en vous inspirant de:

https://scikit-learn.org/stable/auto_examples/linear_model/plot_lasso_coordinate_descent_path.html#sphx-glr-auto-examples-linear-model-plot-lasso-coordinate-descent-path-py

Trick1: N'oubliez pas de standardiser les variables au préalable !! Par exemple avec la fonction preprocessing.StandardScaler()

Trick2: Vous pouvez rendre le code générique pour calcul le chemin LASSO de régression de n'importe quelle variable

6. **Lasso with AIC or BIC criterion**
    a. Que fait la fonction LassoLarsIC de sklearn ?
    b. Affichez les valeurs des critères AIC et BIC en fonction de alpha en vous inspirant de:

https://scikit-learn.org/stable/auto_examples/linear_model/plot_lasso_model_selection.html

7. **Conclusion**
    a. Parmis tous les modèles de régression appris lequel a le meilleur score (coefficient de détermination) ?
    b. Le score de ce modèle est-il une bonne estimation du score que l'on peut espérer obtenir avec ce modèle sur un nouvel ensemble de test ? Pourquoi ?