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 sur les modèles linéaires = http://www-sop.inria.fr/members/Alexis.Joly/01_Linear_Lasso.pdf

Lien de la version en ligne de ce document: https://docs.google.com/document/d/1n2mWvNMlYF6ob-CrmkaZ2LNh9DhST0lJXiJRBQPr7OQ/edit?usp=sharing

1. **Collecte de données**
    a. Quelles données personnelles seraient d'après vous utiles pour prédire le sexe d'une personne ?

    b. Quelles données personnelles seraient d'après vous utiles pour prédire l'animal préféré d'une personne ?

    c. Remplissez le sondage: https://framaforms.org/miashs-td-personal-data-1574696576

    d. Télécharger les données: *Lien fourni le jour du TD*

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 vous inspirant de

https://scikit-learn.org/stable/tutorial/statistical_inference/supervised_learning.html#supervised-learning-tut

   b. Ecrivez un code pour apprendre un modèle de régression linéaire pour prédire le sexe. Quel est le coefficient de régression ? Quels sont les variables d'entrée ayant le plus d'importance selon ce modèle ?
   c. Faites de même pour prédire l'animal préféré.

4. **Sélection de variables avec RFE**
   a. Ecrivez un code permettant de sélectionner les 3 variables les plus explicatives avec l'algorithme Recursive Features Elimination: https://scikit-learn.org/stable/auto_examples/feature_selection/plot_rfe_digits.html
   b. Que fait l'algorithme RFE ? Le modèle obtenu est-il meilleur que le modèle linéaire complet sur les données de test ? Que pourrait-on faire pour choisir le bon nombre de variables à conserver ?

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

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 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 ?