Examen - Outils Python pour Machine Learning

Nicolas Bourgeois

SCIA, S9, 2020-2021

L'examen dure 2 heures. Tous les documents sont autorisés. Le rendu s'effectue soit sous la forme d'un notebook ipython, soit sous la forme d'un script python avec des commentaires. Toutes les librairies sont autorisées. Toute question pour laquelle le code ne tourne pas ou ne renvoie pas une réponse au bon format ne sera pas lue. ***Tous les noms de fichiers doivent contenir votre NOM et votre PRENOM.*** *Prenez le temps de faire les choses bien, il n'y a pas du tout besoin d'aller au bout pour valider l'examen. Les questions ne sont pas indépendantes mais certaines peuvent être sautées.*

1) Chargez le fichier `joueurs.csv` et ne conservez que les lignes sans valeurs manquantes. Attention : pour la colonne "nombre d'enfants" un 0 n'est pas une valeur manquante, mais pour d'autres colonnes si.

2) Ne conservez que les personnes âgées entre 20 et 40 ans, et séparez entre un échantillon d'apprentissage et un échantillon de test.

3a) Essayez de produire une régression entre l'âge et le nombre d'enfants. Evaluez-en la qualité selon la méthode qui vous semblera pertinente.

3b) Même question mais en séparant l'effectif par modalité de situation conjugale.

3c) Représentez le résultat (scatterplot + droite de régression) dans le cas de la modalité "Marié(e), Pacsé(e)".

4a) Considérez maintenant la variable nombre d'enfants comme catégorielle, et entraînez un arbre de décision sur l'ensemble des variables. Calculez les indicateurs usuels de qualité.

4b) Affichez l'arbre.

5) Chargez le fichier `communes.csv`. Faites un nettoyage sommaire de la colonne "résidence" du premier fichier, et effectuez un left join (le fichier joueurs étant à gauche) entre

les deux tables en utilisant cette colonne.

6) En divisant en latitute/longitude la colonne appropriée du fichier de communes, provoquez un scatterplot des positions des joueurs sur une carte de France.

7) Même question, mais en utilisant une somme de façon à ce que le rayon des points soit proportionnel au nombre de joueurs dans le ville.

8) Avec la méthode de votre choix, effectuez une classification non supervisée des joueurs et donnez une interprétation de chaque classe.