Examen - Outils Python pour Machine Learning

Nicolas Bourgeois

SCIA, S9, 2019-2020

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. 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 `communes.csv`. Un des champs contient les coordonnées spatiales de la commune. Séparez ce champ en deux champs *latitude* et *longitude*. Attention aux types.

2) Ne gardez que les communes métropolitaines (latitude comprise entre 40 et 52, longitude entre -5 et 9).

3) Avec un scatterplot, affichez les différentes communes en fonction de leur latitude et de leur longitude.

BONUS : Le résultat devrait vous paraître un peu déformé car un degré de longitude représente une distance qui décroît à mesure qu'on monte vers le Nord. Essayez de trouver un coefficient par lequel multiplier la longitude pour corriger ce problème. A moins d'être déjà très à l'aise avec geopandas ne perdez pas de temps à comprendre comment gérer rigoureusement cette question pendant l'examen.

4) Entraînez une régression linéaire pour trouver l'altitude à partir de la longitude et de la latitude. Le résultat est-il conforme à votre intuition géographique ? Quelle est l'erreur calculée par la moyenne des moindre carrés ?

5) Affichez le scatterplot pour les data (l'altitude en ordonnée) sur deux graphes (un où l'abscisse est la latitude, un où l'abscisse est la longitude). Identifiez les principales chaînes de montagne. Vous pouvez afficher par-dessus la droite de régression pour une meilleure visualisation, mais ayez conscience que sa valeur à l'origine est arbitraire (c'est l'intersection avec le plan zéro) donc ne vous inquiétez pas si c'est décalé.

6) Le code insee est obtenu par concaténation du numéro de département et du code commune. Pour l'ensemble des communes possédant ces éléments,

rajoutez une colonne code insee et remplissez-la.

7) Chargez le fichier `demographie_par_commune.csv`. Ne gardez que les champs correspondant au code INSEE (LIBGEO) et à la population. Faites une jointure entre vos deux tables. Vérifiez bien qu'il n'y a pas de doublons, et que vous avez à peu près le bon nombre de lignes à la fin.

HINT : il peut y avoir une petite différence à cause des DOM, des arrondissements parisiens/lyonnais et des changements administratifs entre le moment où ont été réalisées les deux tables.

8) A partir de la table jointe et en utilisant un groupby, calculez la population de chaque département.

HINT : Si vous n'arrivez pas à faire correctement la jointure, vous pouvez toujours retrouver le numéro du département directement dans la 2e table en gardant simplement les deux premiers chiffres du code INSEE.

9) Divisez maintenant votre table de communes en échantillon d'entraînement et de test.

10) A l'aide de méthodes supervisées de votre choix, essayez de prédire le numéro du département à partir de la latitude et de la longitude.

11) Pourquoi faut-il utiliser une classification et non une régression pour cette question, bien que la variable de sortie est numérique ?

12) Coloriez la carte de l'exercice 3 à partir du département prédit et affichez le résultat.

13) Affichez les métriques qui vous semblent pertinentes et comparez la qualité des différents algorithmes.

14) Expliquez les différences graphiques de leurs résultats à partir de la façon dont fonctionnent les algorithmes.