Examen mi-parcours

Master parcours SSD - UE Analyse Fouille de Données

Automne 2021

Cet examen mi-parcours prend la forme d'un devoir maison, à réaliser **en binôme** pour le **lundi 10 janvier**. Vous devrez me rendre (via la plateforme moodle) :

1. un rapport de 6 pages maximum (sans le code) décrivant les analyses réalisées,
2. un fichier texte contenant les prédictions obtenues à l'issue de l'exercice,
3. un notebook Jupyter (commenté un minimum) permettant de reproduire vos analyses.

Dans cet exercice vous travaillerez sur une problématique de classification binaire visant à reconnaître le statut clinique d'un patient à partir de deux matrices de descripteurs. Plus précisément, vous chercherez à prédire si un patient souffre d'une maladie inflammatoire chronique de l'intestin (IBD - Inflammatory Bowel Disease) à partir :

1. de covariables cliniques décrivant (essentiellement) ses habitudes alimentaires,
2. d'une caractérisation de son microbiote intestinal, c'est à dire de la constitution de sa flore bactérienne intestinale.

Cette problématique offre de grandes perspectives en terme de diagnostic et suscite un intérêt important dans la communauté scientifique, la valeur prédictive apportée par le microbiome restant néanmoins encore à démontrer.

L'objectif de cette étude sera donc de construire un modèle de prédiction le plus performant possible à partir de ces deux sources d'information. Vous aurez à disposition un jeu d'apprentissage à partir duquel vous devrez fournir (à l'aveugle) des prédictions sur un jeu de test. J'évaluerai ensuite les prédictions, et leur qualité contribuera à la note. Précision importante : le critère de performance considéré sera l'aire sous la courbe ROC, en considérant les patients souffrant d'IBD comme catégorie positive.

La constitution du jeu de données est la suivante :

— le jeu d'apprentissage met en jeu 4965 individus, et prend la forme de 3 fichiers :
  — `train-data_clinical.csv` : une matrice de taille 4965×13 contenant 13 covariables cliniques décrivant essentiellement les habitudes alimentaires des patients (e.g., si le patient consomme des légumes). Chacune de ces covariables se présente sous la forme d'un descripteur quantitatif "ordinal" (i.e., une valeur numérique encodant si le patient consomme jamais / un peu / beaucoup de légumes).
  — `train-data_otu.csv` : une matrice de taille 4965×7033 contenant une caractérisation de la flore microbienne des patients. Chaque colonne de cette matrice correspond à une espèce bactérienne. Chaque ligne correspond à un patient, et contient les proportions relatives des différentes bactéries observées chez ce patient (la somme de chaque ligne vaut donc 1).
  — `train-label.txt` : un vecteur de 4965 valeurs donnant le statut clinique des différents patients ("IBD" ou "Healthy").
— le jeu de test se constitue de 850 individus, pour lesquels vous disposerez des mêmes matrices de descripteurs : `test-data_clinical.csv` et `test-data_otu.csv`.

Vous enregistrerez les prédictions obtenues sur le jeu de test dans un fichier texte, à me soumettre, ainsi que votre rapport, via la plateforme moodle comme évoqué ci-dessus. Vous avez toute liberté quant au choix des modèles à considérer, de leurs hyperparamètres et des éventuels pré-traitements à appliquer au jeu de données. Vous serez évalués sur la qualité de vos prédictions, sur la pertinence de votre démarche, et sur l'originalité de votre travail. N'hésitez notamment pas à explorer certaines fonctionnalités de `scikit-learn` peu ou pas vues en cours.