Parcours Informatique Médicale
DFGSM3
Semestre 2 - 2018-2019

A Neuraz, B Rance, M Wack

15/05/2018

Le sujet comporte 3 volets. Chaque binôme doit réaliser les 3 parties.

Le projet devra être rendu au plus tard le **27 mai 2019 à 20h**.

La soutenance aura lieu le **29 mai 2019** pendant les heures de cours

Dataset

Ce projet est consacré au machine learning. Vous travaillerez sur un dataset public de biopsies à l'aiguille fine de tumeurs du sein. Le dataset est à récupérer à l'adresse suivante : url

Vous devrez le préparer afin de pouvoir l'analyser.

Questions

Analyse descriptive

Dans un premier temps, vous effectuerez l'analyse descriptive du dataset. Vous décrierez la nature de la population de l'étude, et des différentes variables. Si nécessaire, vous identifierez la nature des différentes variable et réfléchirez à leur pertinence. L'analyse sera illustrée par des graphiques réalisés à l'aide de la bibliothèque ggplot2.

Clustering

La première étape de votre exploration par méthodes de machine learning utilisera les techniques de clustering. Vous comparerez les résultats obtenus par trois méthodes de clustering :

- K-means,
- hierarchical clustering
- 1 méthode de votre choix

Vous discuterez de la validité des clusters et leurs cohérences à l'aide de métriques adéquates. Vous permettrez également la visualisation des résultats obtenus à l'aide de représentation graphique (dendrogramme, heatmap, t-sne).

Classification

Vous prendrez garde à conserver une partie du dataset pour l'évaluation du modèle final.

Utiliser au moins 3 approches différentes:

- 1 méthode à base d'arbres de décision
- une autre approche de votre choix

- une méthode d'ensemble learning de votre choix.

Vous devrez proposer et justifier une métrique d'évaluation des modèles adaptée à la question. Vous comparerez les résultats des différents modèles testés. Une attention toute particulière devra être portée à l'interprétation des résultats.

Bonus: Taille du sample, imbalance de classes

Modalités d'évaluation

Vous devrez fournir un rapport d'analyse détaillé et interprété sous la forme d'un fichier RMarkdown. Le fichier de données d'entré devra être le fichier original, et chargé depuis le répertoire de travail. Une présentation orale de 10min + 10min de questions sera également organisée. Une grande importance sera apportée à l'interprétation des résultats.

Coordonnées des enseignants

Maxime Wack	maxime.wack@aphp.fr	01.56.09.23.63 / 06.98.80.60.36
Antoine Neuraz	antoine.neuraz@aphp.fr	01.71.39.65.85 / 06.24.62.23.55
Bastien Rance	bastien.rance@aphp.fr	01.56.09.59.85 / 06.14.89.16.35