Projet Data Science en R

S8 – 2019/2020

Objectif du projet

Le but de ce projet est de réinvestir l'ensemble des compétences présentées dans le module de Data Science. Vous utiliserez le langage R et Rstudio pour générer un document en format pdf à la fin de votre projet.

Notez qu'une attention particulière devra être apportée à l'interprétation des résultats obtenus

Modalité du projet :

Equipe de 4 personnes

L'ensemble de la partie Statistique sera regroupé dans un fichier nommé **Stat*_nom1_nom2_nom3_nom4*.pdf** avec les noms des 4 membres du groupe (vous inclurez également dans le titre du document les noms des 4 étudiants)

L'ensemble de la partie DataMining sera regroupé dans un fichier nommé **DM*_nom1_nom2_nom3_nom4*.pdf** avec les noms des 4 membres du groupe (vous inclurez également dans le titre du document les noms des 4 étudiants)

Les 2 pdf générés seront à charger sur itslearning au plus tard **pour le mardi 7 avril 2020 23h59 (pénalités *en cas de retard)***

I – Partie Statistique

a) Choix d'un jeu de données

Dans ce projet, vous avez la possibilité de choisir un ou deux jeux de données pour votre étude. Néanmoins, les *datasets* doivent contenir suffisamment de données pour pouvoir réaliser un modèle linéaire (ANOVA, ANCOVA ou régression linéaire) ainsi qu'une analyse en composantes principales (ACP).
Une première source de données est disponible sur http://www.rdatamining.com/resources/data. Vous pouvez également utiliser les datasets disponibles sur R. Pour avoir une liste des jeux de données, il suffira d'exécuter la commande *library(help="datasets").* Pour accéder à un dataset, on peut utiliser la fonction *data()* ; par exemple, pour accéder aux données *airquality*, on écrit *data(airquality)*.

b) Description des données

Il faut d'abord présenter le contexte du jeu de données et la problématique que vous souhaitez traiter grâce à ce jeu de données. Ensuite, présentez les valeurs observées des différentes variables utilisées, grâce à des tableaux ou des graphiques.

c) Analyse en Composantes principales (ACP)

Il faut réaliser une analyse en composantes principales et interpréter les résultats.

d) Modèle linéaire

Il faut réaliser une analyse de la variance (ANOVA), analyse de covariance (ANCOVA) ou une régression linéaire, en suivant toute la démarche de modélisation statistique.

II – Partie Data-Mining

a) Clustering :

Récupérer sur itslearning et étudier le fichier « 2DClustersS8.csv », en recherchant les meilleurs algorithmes, et en recherchant, avec chacun d'entre eux, les paramètres optimaux. Analyser et comparer les résultats obtenus

b) Classification avec les algorithmes étudiés en cours :

Récupérer sur itslearning et étudier le fichier « voteUSA.csv », recensant les votes d'un échantillon de citoyens américains aux élections de 1984, en y associant leurs avis sur les différents sujets d'actualité de l'époque.

Le but de cette analyse est de valider un modèle de prédiction : peut-on prédire le vote d'un citoyen américain, en fonction de ses avis sur les sujets d'actualité ? (Ce modèle serait alors potentiellement transposable aujourd'hui, en se basant bien sûr sur les sujets d'actualité d'aujourd'hui : à partir des avis émis sur les réseaux sociaux, il deviendrait aujourd'hui envisageable, pour un parti, de repérer les citoyens pouvant voter pour eux à la prochaine élection, afin de pouvoir les contacter et les inciter à aller voter…)

On se placera dans cette étude à titre d'exemple du côté démocrate. Vu du camp démocrate, la prédiction sera considérée comme « positive » si elle identifie un démocrate. **Votre objectif sera ainsi d'obtenir une sensitivité maximale, tout en maintenant dans tous les cas une précision globale de la classification supérieure à 50%**.

Vous réserverez 80% des données à l'apprentissage, et 20% à vos tests**.** Recherchez les meilleurs algorithmes, et rechercher, avec chacun d'entre eux, les paramètres optimaux, dans le but de prédire au mieux la colonne « Vote ». Analyser et comparer les résultats obtenus.

Choisir l'algorithme et les paramètres optimaux identifiés, et prédire en conséquence les classes du fichier « voteUSAtoPredict.csv ». Afficher le tableau de résultat avec la 1ère colonne de données, et la prédiction que vous proposez.

c) Règles d'association :

Transformer le cas échéant le jeu de données que vous avez choisi dans la partie Statistique, et rechercher les règles d'association les plus pertinentes relatives à vos données