Traitement numérique

PROJET

Ce projet a pour but de mettre en pratique une approche initiale des différentes analyses supervisées et non supervisées abordées en cours, au moyen du logiciel et langage de programmation statistique R.

Partie 1

Télécharger puis importer le jeu de données disponible à l'adresse suivante http://up5.fr/Gnf9l

Veillez à consulter la description associée au jeu de données. Ce dernier est initialement adapté à un cadre supervisé, il conviendra donc de repérer et d'omettre la variable représentant les labels lors des analyses non supervisées.

1. Réaliser une présentation brève du jeu de données (taille, etc.) puis une analyse descriptive des variables présentes au moyen des différents outils utilisés en cours (moyenne, écart-type, boxplot, etc.). Les variables sont-elles homogènes ? Qu'envisagez-vous dans le cas contraire ? Qu'observe-t-on pour la variable « Monetary..c.c..blood. » ? Justifier votre réponse.
2. Visualiser les individus au sein d'un plan factoriel des composantes obtenues par une réduction de la dimension. Il conviendra d'interpréter de façon rigoureuse ces premiers résultats (plan factoriel des individus – qu'observe-t-on dans le nuage de points? cercle des corrélations, etc.) et de justifier le choix des différents paramètres utilisés dans la fonction (PCA).
3. Définir une partition de groupes d'individus (un clustering) au moyen des différents algorithmes étudiés en cours (l'algorithme des centres mobiles Kmeans, la classification ascendante hiérarchique (CAH) avec les 4 critères). Présenter et interpréter les résultats à travers une étude comparative suivant les différentes méthodes, pour cela, faire appel aux différents outils de comparaison utilisés en TP (table de confusion). Vous comparerez donc les résultats de chaque méthode avec la variable des vrais labels.
4. Projeter les classes obtenues dans le plan factoriel issues des composantes et interpréterez chaque groupe d'individus obtenu à l'aide des boites à moustaches (boxplots) et des statistiques descriptives des variables. Les dendrogrammes (segmentés en fonction du nombre de clusters désirés) issus des classifications hiérarchiques suivant les différents critères seront appréciés.
5. Récupérer le jeu de données à l'adresse suivante http://up5.fr/29wou et refaire les questions 1-4.

Partie 2

Télécharger le jeu de données disponible à l'adresse suivante http://up5.fr/A3Mhy

La matrice de données (matrice documents-termes) est contenue dans le dictionnaire « dtm » du ficher « .mat », vous retrouverez le vecteur des labels des individus au sein du dictionnaire « classid ».

1. Appliquer l'algorithme des centres mobiles sur le jeu de données pour une partition correspondant au nombre de classes contenues dans « classid » et comparez-la avec la vraie partition. Que peut-on observer?
2. À l'aide du package « igraph », construire un graphe des documents en réalisant la matrice d'adjacence (matrice exclusivement binaire qui traduit la relation entre deux documents par la valeur 1).
3. Visualiser le graph obtenu, puis appliquer l'algorithme de détection de communautés basé sur la modularité. Projeter les groupes obtenus sur le graph. Comparer la partition obtenue avec la vraie partition ? avec la partition obtenue par Kmeans ? Que peut-on observer?

Traitement numérique

Partie 3

Télécharger le jeu de données ***Combined Cycle Power Plant*** disponible à l'adresse suivante : http://up5.fr/iJ-bj

Charger ensuite les données qui se trouvent dans le fichier excel « Folds5x2_pp.xlsx » à l'aide du package « XLConnect ».

Ce jeu de données représente la production d'électricité (PE) sur une durée de 6 ans (entre 2006 et 2011) ainsi que quatre variables qui influent sur la production d'électricité à savoir la température (T), la pression (AP), l'humidité (RH) et le vide d'échappement (V). Le but est de pouvoir prédire la production électrique et de distinguer les variables qui influencent le plus la production d'électricité.

1. Réaliser une analyse descriptive des 5 variables, commenter les résultats.
2. Afficher sur un même graphique, les 200 premiers points de chacune des quatre variables ainsi que la variable à prédire PE.
3. Calculer la corrélation entre les différentes variables, interpréter les résultats.
4. Visualiser les nuages de points entre tous les couples de variables, quelles sont les deux variables qui permettent de mieux expliquer la variable à prédire PE ?
5. Réaliser une régression linéaire entre ces deux variables en question et la variable PE séparément, afficher les résultats de chacune des deux régressions sur le nuage de points. Commenter ensuite les résultats obtenus (paramètres de la régression **a** et **b**).
6. Calculer les deux erreurs (fonction couts) MSE et MAE pour les deux régressions. Que peut-on en déduire de la comparaison des résultats des deux régressions ?

Veuillez envoyer avant le 03/05/2018 à 23h59 le code R qui vous a permis de réaliser ce projet, accompagné d'un rapport de 5 pages au maximum.

PS : Le projet est à réaliser en binôme. Les projets similaires de groupes différents seront sanctionnés par la note 0.

E-mail : mickael.febrissy@parisdescartes.fr

rafika.boutalbi@parisdescartes.fr