Traitement numérique des données

PROJET

Veuillez envoyer avant le 10/05/2020 à 23h59 le code R qui vous a permis de réaliser ce projet, sous forme d'un document markdown.

IMPORTANT: Ce projet sera aussi présenté sous forme d'une soutenance.

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

E-mail : rafika.boutalbi@parisdescartes.fr et mickael.febrissy@parisdescartes.fr

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. Nous nous intéressons à une problématique d'actualité, à savoir l'épidémie du virus COVID-19. L'objectif de ce projet est d'analyser les données disponibles sur cette pandémie afin d'essayer de comprendre l'évolution ainsi que la dynamique de propagation du virus.

Partie 1 : Analyses descriptives

Nous souhaitons effectuer une analyse descriptive des données disponibles à propos du COVID-19 par département en France. Les données considérées proviennent de la base de données "Chiffres-clés" disponible sur le site "data.gouv.fr" auxquelles nous avons ajouté des informations sur la localisation géographique des départements. La période étudiée dans ce projet est celle du 1er Avril 2020 au 14 Avril 2020.

1. Télécharger le jeu de données (http://up5.fr/hBfL8) et le charger sous R.
2. Réaliser une brève présentation du jeu de données (dimensions, types de variables, etc) puis une analyse descriptive des variables présentes au moyen des différents outils vus en cours et TP (diagramme en secteurs, boxplot, histogramme, etc).
3. En vous inspirant du tutoriel suivant (http://up5.fr/cwU9e) qui permet l'affichage de données sur des cartes géographiques à partir de coordonnées GPS (latitude et longitude), afficher chacune des variables «deces_total», «reanimation_total», «hospitalises_total» et «gueris_total» sur une carte géographique. Mettre en avant les données en utilisant des codes couleurs (jaune faible, rouge fort) ou alors en faisant varier les tailles des points. Analyser les résultats obtenus.

4. **Bonus**: Pour ceux d'entre vous qui sont intéressés, vous pouvez utiliser le package leaflet.minichart, qui est compatible avec ggplot (voir le tutoriel suivant http://up5.fr/MgJ5o).

Partie 2 : Prédiction du nombre de décès

Dans cette deuxième partie, nous souhaitons étudier et prédire le taux de mortalité des personnes atteintes du COVID-19 (variable «deces_total») selon les variables «reanimation_total», «hospitalises_total» et «gueris_total».

1. Calculer la corrélation entre les différentes variables et afficher la matrice de corrélation à l'aide du package « corrplot ». Commenter les résultats.

2. Visualiser les nuages de points entre chaque variable et la variable à prédire «deces_total». Quelles sont les variables qui permettent d'expliquer au mieux la variable à prédire ?
3. Diviser votre jeu de données en deux ensembles apprentissage/test avec une proportion de 80%-20%.
4. Réaliser une régression linéaire entre les variables explicatives «reanimation_total», «hospitalises_total», «gueris_total», et la variable à expliquer «deces_total» sur l'ensemble d'apprentissage. Afficher et commenter les quatre graphiques de diagnostic.
5. Commenter ensuite les résultats obtenus (les coefficients de régression liés à chaque variable).
6. Calculer les erreurs MAE (Mean Absolute Error) et MSE (Mean Squared Error) de l'ensemble d'apprentissage, puis celles de l'ensemble de test.
7. Que peut-on déduire de la comparaison des résultats des erreurs de l'ensemble d'apprentissage et de test ? Avec quelle précision pouvons-nous prédire le nombre de décès pour un département donné ?

Partie 3 : Clustering des départements selon la dynamique de propagation du virus

À travers cette partie nous souhaitons comprendre la dynamique de propagation du virus et voir s'il existe des départements pour lesquels le nombre de personnes guéries, atteintes et décédées du COVID-19 sont similaires.

1. Réaliser une ACP. Dans un premier temps, visualiser le nuage des individus et le cercle des corrélations suivant les composantes principales obtenues. Qu'observez-vous ? Il conviendra d'interpréter de façon rigoureuse ces premiers résultats (le pourcentage d'inertie, les contributions des variables et des individus, etc)
2. Dans un second temps, projeter les individus (à savoir les départements) dans le plan d'inertie maximum et les colorier en utilisant la variable «deces_total» (voir l'exemple dans la section *Gradients de couleurs pour un graphique en nuage de points* sur le lien suivant http://up5.fr/k90qK). Interpréter maintenant la disposition des individus en fonction de cette variable.
3. Réaliser diverses classifications non-supervisées (clustering) au moyen des différents algorithmes étudiés en cours : K-means, CAH avec les 4 critères (lien minimum, moyen, maximum, Ward). Expliquer comment vous avez choisi le nombre de classes.
4. Présenter et réaliser une étude comparative entre les résultats des différents algorithmes sous forme d'un tableau. Projeter les groupes obtenus par chaque méthodes (clusters) sur le nuage des individus et interpréter les groupes obtenus.
5. Que peut-on déduire de l'analyse des résultats des trois parties ?

Bonus

Le jeu de données « Fatality » (à télécharger via http://up5.fr/fatalitydata dont la documentation est disponible via http://up5.fr/fatalitydatadocumentation) arbore des données sur le trafic autoroutier des Etats-Unis qui furent récoltées par état et sur plusieurs années. Afin de synthétiser l'information,

Traitement numérique des données

nous souhaitons classer les 336 observations en 3 groupes et ce en considérant uniquement les variables quantitatives : "mrall", "beertax", "mlda", "vmiles", "unrate" et "perinc".

Par la suite nous chercherons à interpréter les caractéristiques de ces groupes suivant l'ensemble des variables (quantitatives et qualitatives) au moyen des différents outils statistiques vus en cours (boxplots, diagramme circulaire, tendances, etc). Vous pouvez par exemple étudier la corrélation des variables suivant les valeurs de chaque groupe. Ci-dessous vous trouverez des questions non contractuelles visant à vous aiguiller lors de vos interprétations :

- L'âge minimum pour boire de l'alcool semble t-il lié au nombre de morts dans certains groupes ?
- Les groupes où les taxes de bière les plus faibles sont-ils ceux pour lesquels le nombre de morts est le plus élevé ? ceux avec le taux de chômage le plus élevé ?
- Les groupes où l'on boit le plus sont-ils ceux où les revenus sont les plus élevés ? ceux avec le moins de condamnations ?
- La mortalité est-elle équivalente pour tous les groupes ? Quels sont les états où la mortalité est la plus élevée ? etc.

Proposer une étude visant à obtenir cette synthétisation de l'information au moyen des différentes méthodes d'analyse non supervisée/classification (CAH avec les quatre critères, K-means) et exploratoire (ACP) vues en cours. La qualité de l'interprétation, la pertinence et l'originalité de votre étude seront primordiales. Il ne s'agit pas de quantité ! Il n'est pas nécessaire de présenter une interprétation pour les partitions de chaque algorithme ou de s'étendre sur des variables non discriminantes entre les groupes. Présentez les résultats de l'algorithme qui, à votre sens, semble obtenir les groupes les plus cohérents. Néanmoins, vous devrez justifier brièvement votre choix par rapport aux autres méthodes de classification.