UE - Science des données numériques

TP 6
Analyse et prédiction des infections COVID-19

Date limite: 30 novembre 2020, 15:00

Des cas de nouveaux coronavirus ont été signalés pour la première fois à Wuhan, dans la province du Hubei, en Chine, en décembre 2019 et se sont depuis propagés dans le monde entier. Des études épidémiologiques ont indiqué une transmission interhumaine en Chine et ailleurs.

Des données épidémiologiques sont nécessaires pendant les épidémies émergentes pour mieux surveiller et anticiper la propagation de l'infection.
L'ensemble de données a été rendu public le 20 janvier 2020 contenant différentes informations sur les patients : clinique, démographique et géographique.

Il peut être téléchargé gratuitement sur (https://github.com/beoutbreakprepared/nCoV2019) ou directement du lien GoogleDrive en format csv:

https://docs.google.com/spreadsheets/d/1itaohdPiAeniCXNlntNztZ_oRvjh0HsGuJXUJWET008/edit#gid=0

Vous pouvez aussi le télécharger ici :

https://docs.google.com/spreadsheets/d/e/2PACX-1vQU0SIALScXx8VXDX7yKNKWWPKE1YjFlWc6VTEVSN45CklWWf-uWmprQIyLtoPDA18tX9cFDr-aQ9S6/pubhtml#

Le dataset est également disponible sur la plateforme kaggle (https://www.kaggle.com/allen-institute-for-ai/CORD-19-research-challenge), choisissez le fichier COVID19_line_list_data.csv

Ce fichier sera aussi disponible sur le Teams via vos professeurs.

Note: *pour charger un fichier.csv avec Python, vous devez utiliser la function csv_read() de la librairie Pandas.*

Il y a aussi un répertoire Github ou vous trouverez les dernieres données collectées :

https://github.com/beoutbreakprepared/nCoV2019

La situation épidémiologique du COVID-19 est en constante évolution. Le dossier de données d'archive disponible via le référentiel Github est mis à jour quotidiennement. Chacune des lignes représente un cas et un ID individuels.

Une description des champs de la base de données est présentée dans cet article: : *Epidemiological data from the COVID-19 outbreak, real-time case information*

L'objectif de ce TP est d'explorer ces données pour en extraire des connaissances afin d'aider la communauté à mieux comprendre la propagation du COVID-19.

Le TP est composé de l'ensemble des questions suivant :

Afin d'analyser l'ensemble des données, vous devez identifier et extraire certaines informations statistiques sur les données, par exemple : le type de données, les valeurs manquantes, les valeurs aberrantes, la corrélation entre les variables, etc.

Dans le cas des valeurs manquantes, vous pouvez les remplacer par la moyenne, la médiane ou le mode de la variable concernée.

1. Calculez les corrélations entre les variables. Quelles sont variables les plus corrélées avec la cible ('result')? Expliquez les résultats.

2. Visualisez les données en deux dimensions en passant par l'ACP (analyse en composantes principales). Pouvez-vous utiliser une autre méthode ?

Dans la suite, nous utilisons une méthode d'apprentissage automatique afin de prédire la classe : les patients sont soit «décédés» ('died') soit «sortis» ('discharged') de l'hôpital. Vous pouvez utiliser la classification par K-Nearest Neighbours (K-NN), l'arbre de decision ou le classificateur Bayes.

3. Les résultats obtenus doivent être validés en utilisant certains indices externes comme l'erreur de prédiction (matrice de confusion et précision) ou d'autres comme Rappel, F-Measure, …

4. Utilisez la régression pour prédire l'âge (age) des personnes en fonction d'autres variables. Vous avez le choix sur ces variables explicatives ? Comment choisissez-vous ces variables ? Calculez la qualité de la prédiction à l'aide de l'erreur MSE (Mean Squared Error).

5. Appliquer trois méthodes de clustering (K-means, NMF et CAH) sur l'ensemble de données pour segmenter les personnes en différents groupes. Utilisez l'index de Silhouette pour connaître le meilleur nombre de clusters.

6. Visualisez les résultats à l'aide de scatter pour analyser visuellement la structure de clustering des trois méthodes.

7. Les données sont déséquilibrées. Vous pouvez les équilibrer en réduisant aléatoirement la classe majoritaire. Supposons que vous extrayez aléatoirement des échantillons équilibrés. Comment les résultats de la prédiction changeront-ils?

8. Comment pouvez-vous mieux gérer ce déséquilibre entre les classes ?

9. Pour trouver les meilleurs paramètres pour les modèles, l'algorithme Greedy Search peut être utilisé, disponible dans la bibliothèque scikit-learn. Expliquez l'algorithme et utilisez-le pour les modèles d'apprentissage choisis afin de trouver les meilleurs paramètres.

10. Présentez et expliquez le formalisme algorithmique et mathématique de la méthode qui donne les meilleurs résultats. Expliquez tous les paramètres de la méthode utilisée et leur impact sur les résultats.