Transcription de l'épisode 079 – Les résultats (09) – DEFI6M

Données et valeurs manquantes (méthodes complexes d'imputation 3, 4 et 5)

Aujourd'hui, nous allons voir 3 méthodes, 3 solutions possibles qui sont un peu plus complexes que les 2 premières, pour traiter les données, pour imputer les données manquantes.

La 3ème solution serait de remplacer les données manquantes avec un arbre de décision.

Avec cette méthode, les choses commencent à se préciser un peu plus. Au lieu d'attribuer toujours la même valeur pour tous les individus, on va affecter une valeur personnalisée en fonction des autres données existantes. Vous pouvez donc utiliser pour cela un arbre de décision.

Pour aller plus loin, il existe d'ailleurs un package R MissForest qui permet d'imputer les valeurs manquantes grâce à une méthode basée sur des forêts aléatoires.

4ème méthode d'imputation : il s'agit de remplacer les données manquantes par les valeurs les plus proches.

Pour ce faire, on peut aussi utiliser la méthode KNN (dite k- plus proches voisins) pour estimer les valeurs manquantes. Pour cela, pour chaque individu ayant une valeur manquante, on recherche les k-individus les plus proches (en calculant la distance sur les autres variables renseignées) puis on remplace la valeur manquante par la moyenne de ces k-individus.

5ème et dernière méthode : Remplacer les données manquantes avec des algorithmes dédiés.

La problématique des données manquantes est telle que des algorithmes spécifiques ont été développés pour y répondre. Vous retrouverez ces algorithmes implémentés dans certains outils. Dans R, on retrouvera par exemple AMELIA qui est basé sur l'estimation du maximum de vraisemblance et des échantillons bootstrap. Le package Hmsic permet également d'utiliser plusieurs méthodes à base de régression et de bootstrap.

Dans tous les cas, que vous utilisiez une méthode simple ou plus compliquée, le traitement des valeurs manquantes est une partie importante du travail de préparation des données qu'il ne faut absolument pas négliger.

Voilà, c'est terminé pour aujourd'hui et je vous remercie de votre écoute ; je vous invite à visiter mon blog methodorecherche.com et on se retrouve mercredi pour un nouvel épisode du podcast de Methodo Recherche. A mercredi !

Références :

Enders, C. K., & Guilford Press. (2010). Applied missing data analysis. NewYork; London: The Guilford Press.

Abonnez-vous au Podcast suivant votre préférence d'écoute. Vous trouverez toutes les possibilités et les explications à l'URL :

https://methodorecherche.com/subscribe-to-podcast/

En complément, vous êtes libre de vous abonner à ma newsletter et recevoir gratuitement le bonus "6 clés essentielles pour réussir brillamment votre mémoire de recherche (ou votre thèse)".

http://bit.ly/2RsYpll

A très bientôt,

Christophe