Data Mining

2022-2023

ETUDE DE CAS

Projet par groupe de 2 étudiants
Remise du dossier par mail (rapport Word et script R)
Date : le 31 mars 2023
veronique.cariou@oniris-nantes.fr

Descriptif des données

Ces données sont issues de l'excellent livre *Elements of Statistical Learning* de Hastie, Tibshirani & Friedman (2009) et portent sur l'analyse de phonèmes. Les données ont été extraites de la base de données TIMIT qui est une ressource largement utilisée pour la recherche en reconnaissance vocale. Le projet porte sur la discrimination de cinq phonèmes transcrits comme suit: "sh" comme dans "she", "dcl" comme dans "dark", "iy" comme la voyelle dans "she", "aa" comme la voyelle dans "dark", et "ao" comme première voyelle dans "water". A partir de 50 discours enregistrés, 4509 trames vocales d'une durée de 32 ms ont été sélectionnées avec environ 2 exemples de chaque phonème pour chaque locuteur. Le tableau de données résultant contient 4509 lignes et 256 colonnes intitulées "x.1" - "x.256" issues du log-périodogramme effectué sur les différentes trames. Deux dernières colonnes indiquent le phonème prononcé et le locuteur. Chaque trame est identifiée suivant qu'il s'agit de l'apprentissage (train) ou le test (test) dans la colonne locuteur.

Document Word avec comme plan indicatif le suivant ainsi que le fichier R associé

- Présentation du contexte de l'étude et présentation du jeu de données
- Découpage du jeu de données en un échantillon d'apprentissage et test suivant la variable locuteur.
- Analyse exploratoire non supervisée : Classification des trames vocales sur la base des 256 variables "x.1" - "x.256" sur l'échantillon d'apprentissage (train). Caractérisation des classes obtenues
- Analyse supervisée : Discrimination des phonèmes sur la base des variables "x.1" - "x.256" par analyse factorielle discriminante et PLS-DA (avec ou sans sélection de variables) sur la base du jeu de données d'apprentissage (train). Interprétation et comparaison des résultats. Validation des résultats sur l'échantillon test (test).
- Synthèse de l'étude

Une attention particulière sera portée sur l'interprétation des résultats trouvés.