SY19 – A22
TP 10 (noté) : Apprentissage à partir de trois jeux de données réelles

Le but de ce TP est d'appliquer les méthodes vues en cours sur trois jeux de données réelles. Il s'agira de construire des fonctions de prédiction aussi performantes que possible. Les performances sur un jeu de données de test (non fourni) seront prises en compte dans la notation. Pour les données `Communities`, il faudra également présenter une analyse descriptive, déterminer quelles sont les variables qui influent le plus sur le nombre de crimes pour 100.000 habitants, et analyser le sens de cette influence.

1 Datasets

1.1 `Phonemes` dataset

The data were extracted from the TIMIT database (TIMIT Acoustic-Phonetic Continuous Speech Corpus, NTIS, US Dept of Commerce) which is a widely used resource for research in speech recognition. A dataset was formed by selecting five phonemes for classification based on digitized speech from this database. The phonemes are transcribed as follows : "sh" as in "she", "dcl" as in "dark", "iy" as the vowel in "she", "aa" as the vowel in "dark", and "ao" as the first vowel in "water". From continuous speech of 50 male speakers, 4509 speech frames of 32 msec duration were selected, approximately 2 examples of each phoneme from each speaker. Each speech frame is represented by 512 samples at a 16kHz sampling rate, and each frame represents one of the above five phonemes. The breakdown of the 4509 speech frames into phoneme frequencies is as follows :

aa	ao	dcl	iy	sh
695	1022	757	1163	872

From each speech frame, a log-periodogram was computed, which is one of several widely used methods for casting speech data in a form suitable for speech recognition. Thus the data used in what follows consist of 4509 log-periodograms of length 256, with known class (phoneme) memberships.

The learning dataset contains 2250 randomly selected observations (and randomly permuted attributes). It contains 256 columns labelled `X1-X256` and a response column `Y`.

1.2 `robotics` dataset

This files contains data related to the kinematics of a robot arm. There are 4000 learning cases, eight predictors, and one response variable (last column).

1.3 `Communities` dataset

This data set is about communities within the United States. The data combines socio-economic data from the 1990 US Census, law enforcement data from the 1990 US LEMAS survey, and crime data from the 1995 FBI UCR. The response variable is the total number of violent crimes per 100K population (`Per Capita Violent Crimes`). A description of this dataset is contained in file `communities.names`. Some attribute values are missing. The way you handle these missing values will have to be described in your report.

2 Critère de notation et format de remise du devoir

Comme pour le TP4, votre devoir sera noté sur trois critères :

1. variété des méthodes utilisées et rigueur de la méthodologie : 1/3 des points ;
2. performances obtenues sur chaque problème, pertinence de l'analyse des données (données `Communities`) : 1/3 des points ;
3. qualité du rendu écrit : clarté des explications ; correction du français ou de l'anglais ; qualité des tableaux et des figures ; soin dans la présentation du rapport : 1/3 des points.

Vous devrez rendre votre travail **avant le 8 janvier à minuit** sous deux formes :

1. Un rapport écrit au format pdf, éventuellement réalisé avec un *notebook* RStudio, en français ou en anglais, maximum 12 pages, à charger sur Moodle ;
2. Un fichier `Rdata` de données R contenant *uniquement* trois fonctions de noms
   — `prediction_phoneme`
   — `prediction_robotics`
   — `prediction_communities`.
   Chaque fonction admet comme unique argument un *data frame* contenant les données de test et renvoie le résultat de la classification

ou de la régression. Vérifiez la taille de ce fichier : les fichiers trop gros ne pourront être traités. Le fichier sera téléversé à l'adresse `http://maggle.gi.utc` sur le site `maggle` qui calculera les performances de vos algorithmes. Vous êtes limités à 6 essais réussis (6 essais dont les prédictions ont pu être calculées sans erreur).

Exemple de génération d'un fichier `Rdata` :

```
# 1. Apprentissage des modèles.

model.phoneme <- ...
model.robotics <- ...
model.communities <- ...

# 2. Création des fonctions de prédiction

prediction_phoneme <- function(dataset) {
  # Ne pas oublier de charger **à l'intérieur de la fonction** les
  # bibliothèques utilisées.
  library(...)

  # Attention à ce que retourne un modèle en prédiction. Par exemple,
  # la lda retourne une liste nommée. On sélectionne alors les
  # classes.
  predict(clas, test_set)$class
}


prediction_robotics <- function(dataset) {
  ...
}

prediction_communities <- function(dataset) {
  ...
}

# 3. Sauvegarder sous forme de fichier .Rdata les fonctions
# 'prediction_phoneme', 'prediction_robotics', 'prediction_communities'.
# Sauvegarder également les objets utilisés dans ces fonctions
# (`model.phoneme`, `model.robotics` et `model.communities` dans l'exemple) !

save(
  "model.phoneme",
  "model.robotics",
  "model.communities",
```

```
  "prediction_phoneme",
  "prediction_robotics",
  "prediction_communities",
  file = "env.Rdata"
)
```

Remarques :

— Le rapport sera tronqué à 12 pages. Aucune page supplémentaire ne sera pas prise en compte.

— Les fonctions devront s'exécuter automatiquement sans problème. Si ce n'est pas le cas, il ne sera pas tenu compte du résultat.

— Aucun devoir ne sera accepté après la date limite.