SY19 – A21
TP 10 (noté) : Apprentissage à partir de trois jeux de données réelles

Le but de ce TP est de construire des fonctions de prédiction aussi performantes que possible à partir de trois jeux de données réelles, qui sont brièvement décrits ci-dessous. Pour les données de location de vélos, il faudra également déterminer quelles sont les variables qui influent le plus sur le nombre de locations, 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. It contains 256 columns labelled `X1-X256` and a response column `Y`.

1.2 `Letter recognition` dataset

The objective is to identify each of a large number of black-and-white rectangular pixel displays as one of the 26 capital letters in the English alphabet. The character images were based on 20 different fonts and each letter within these 20 fonts was randomly distorted to produce a file of 20,000 unique stimuli. Each stimulus was converted into 16 primitive numerical attributes (statistical moments and edge counts). The training set is composed of 10,000 randomly selected observations.

The dataset contains 17 columns labelled `X1-X16` and a response column `Y`. The meaning of the attributes is as follows (but the attributes have been randomly permuted) :

— horizontal position of box
— vertical position of box
— width of box
— height of box
— total # on pixels
— mean x of on pixels in box
— mean y of on pixels in box
— mean x variance
— mean y variance
— mean x y correlation
— mean of x * x * y
— mean of x * y * y
— mean edge count left to right
— correlation of x-edge with y
— mean edge count bottom to top
— correlation of y-edge with x

1.3 `Bike rental` dataset

Bike sharing systems are a new generation of traditional bike rentals where the whole process from membership, rental and return back has become automatic. Through these systems, the user is able to easily rent a bike from a particular position and return it back at another position. There are currently about over 500 bike-sharing programs around the world with a total of over 500 thousands bicycles. There is great interest in these systems due to their important role in traffic, environmental and health issues.

The bike-sharing rental process is highly correlated to the environmental and seasonal settings. For instance, weather conditions, precipitation, day of week, season, hour of the day, etc., can affect rental behaviors. The dataset is related to the two-year historical log corresponding to years 2011 and 2012 from Capital Bikeshare system, Washington D.C., USA. It contains aggregated daily data with the corresponding weather and seasonal information. The training set is composed of the data for 2011.

The data file contains 14 columns with the following variables :

— `instant` : record index
— `dteday` : date

— `season` : season (1 : spring, 2 : summer, 3 : fall, 4 : winter)
— `yr` : year (0 : 2011, 1 :2012)
— `mnth` : month (1 to 12)
— `holiday` : weather day is holiday or not
— `weekday` : day of the week
— `workingday` : if day is neither weekend nor holiday is 1, otherwise is 0.
— `weathersit` :
  — 1 : Clear, Few clouds, Partly cloudy, Partly cloudy
  — 2 : Mist + Cloudy, Mist + Broken clouds, Mist + Few clouds, Mist
  — 3 : Light Snow, Light Rain + Thunderstorm + Scattered clouds, Light Rain + Scattered clouds
  — 4 : Heavy Rain + Ice Pallets + Thunderstorm + Mist, Snow + Fog
— `temp` : Normalized temperature in Celsius. The values are divided to 41 (max)
— `atemp` : Normalized feeling temperature in Celsius. The values are divided to 50 (max)
— `hum` : Normalized humidity. The values are divided to 100 (max)
— `windspeed` : Normalized wind speed. The values are divided to 67 (max)
— `cnt` : count of total rental bikes

The task is to predict bike rental count (variable `cnt`) based on the environmental and seasonal settings, using at least four machine learning methods studied in the class.

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 (1/3 des points) ;
3. de la 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 devoir **avant le vendredi 7 janvier à minuit** sous deux formes :

1. un rapport écrit au format pdf de maximum 12 pages à charger sous Moodle,
2. un fichier `Rdata` de données R contenant (uniquement) trois fonctions de noms :
   — `prediction_phoneme`
   — `prediction_letter`
   — `prediction_bike`.

   Chaque fonction admet comme unique argument un *data frame* contenant les données de test et renvoie le résultat de la discrimination/régression.

   Vérifiez la taille de ce fichier : les fichiers trop gros ne pourront pas être traités.

   Ce fichier de données devra ensuite être chargé sur un site dédié [1] qui calculera

1. `http://maggle.gi.utc`

les performances de vos algorithmes. Vous êtes limités à 6 essais réussis (6 essais dont les prédictions ont pu être obtenues sans erreur).

Exemple de génération d'un fichier `.Rdata` :

```
# 1. Apprentissage des modèles.

model.phoneme <- ...
model.letter <- ...
model.bike <- ...

# 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_letter <- function(dataset) {
  ...
}

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

# 3. Sauvegarder sous forme de fichier .Rdata les fonctions
# `prediction_phoneme`, `prediction_letter`, `prediction_bike`.
# Sauvegarder également les objets utilisés dans ces fonctions
# (`model.phoneme`, `model.letter` et `model.bike` dans l'exemple) !

save(
  "model.phoneme",
  "model.letter",
  "model.bike",
  "prediction_phoneme",
  "prediction_letter",
```

```
  "prediction_bike",
  file = "env.Rdata"
)
```

Remarques :

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

— Aucun devoir ne sera accepté après la date limite.