SY19 – A19
TP 10 (noté) : Apprentissage à partir de trois jeux de données réelles

Le but de ce TP est de construire des prédicteurs aussi performants que possible à partir de trois jeux de données réelles, qui sont brièvement décrits ci-dessous. Les deux premiers jeux de données doivent impérativement être traités. Le troisième jeu de données (images naturelles) est facultatif ; il permettra d'avoir un bonus de points (entre 0 et 10 points selon le travail réalisé).

1 Jeux de données

1.1 Données astronomiques

Il s'agit de descriptions de 5000 objets astronomiques de trois types : étoiles, galaxies et quasars. Chaque objet est décrit par 17 attributs et une variable de classe.

Signification des attributs :

— objid = Object Identifier
— ra = J2000 Right Ascension (r-band)
— dec = J2000 Declination (r-band)
— u, g, r, i, z = response of the 5 bands of the telescope.
— run = Run Number
— rereun = Rerun Number
— camcol = Camera column
— field = Field number
— specobjid = Object Identifier
— class = object class (galaxy, star or quasar object)
— redshift = Final Redshift
— plate = plate number
— mjd = MJD of observation
— fiberid = fiber ID

1.2 Rendement du maïs

Les données concernent le rendement du maïs en France, dans les différents département sur plusieurs années. L'objectif est de prédire le rendement à partir de données climatiques. Il y a 2300 individus et 58 variables. Signification des variables :

- yield_anomaly : variable à prédire représentant l'anomalie de rendement de maïs (une valeur positive indique un rendement plus élevé qu'attendu, une valeur négative indique une valeur perte de rendement par rapport à la valeur attendue), exprimée en tonne par ha.
- year_harvest : année (anonyme) de récolte (1 à 57)
- NUMD : numéro (anonyme) indiquant le département (de 1 à 94).
- ETP_1,..., ETP_9 : Evapotranspiration potentielle moyenne mensuelle par année et par département (1= janvier, 9=septembre)
- PR_1,..., PR_9 : Précipitation cumulée mensuelle par année et par département (1= janvier, 9=septembre)
- RV_1,..., RV_9 : Rayonnement moyen mensuel par année et par département (1= janvier, 9=septembre)
- SeqPR1,...,SeqPR9 : Nombre de jours de pluie mensuel par année et par département (1= janvier, 9=septembre)
- Tn_1,...,Tn_9 : Température minimale journalière moyenne mensuelle par année et par département (1= janvier, 9=septembre)
- Tx_1,...,Tx_9 : Température maximale journalière moyenne mensuelle par année et par département (1= janvier, 9=septembre)
- IRR : variable comprise entre 1 et 5 liée à la fraction de surface agricole irriguée dans chaque département. La valeur 1 indique une fraction faible, la valeur 5 indique une fraction élevée de surface irriguée. Ces valeurs sont indicatives car établies sur la base d'information collectée pendant une seule année.

1.3 Images naturelles

Il s'agit d'images naturelles (au format JPEG) représentant des voitures, des chats et des fleurs. La tâche consiste à prédire le contenu de nouvelles images appartenant à l'un de ces trois types.

2 Critère de notation et format de remise du devoir

Comme pour le TP3, 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 travail **avant le 16 janvier à minuit** sur Moodle sous forme d'une archive zip (impérativement) contenant exactement *trois* fichiers :

1. Rapport écrit au format pdf, éventuellement réalisé avec un *notebook* RStudio, en français ou en anglais, maximum 8 pages, ou 12 pages si les trois problèmes sont traités (nom de fichier : `rapport.pdf`)
2. Un fichier `classifieurs.R` contenant trois fonctions, de noms
   — `classifieur_astronomie`
   — `regresseur_mais`
   — `classifieur_images`.
   Chaque fonction admet comme unique argument un *data frame* contenant les données de test.
3. Un fichier de données R contenant l'environnement nécessaire à l'exécution des trois fonctions ci-dessus (nom de fichier : `env.Rdata`).

Exemple de fichier `classifieurs.R` :

```
classifieur_astronomie <- function(dataset) {
# Chargement de l'environnement
load("env.Rdata")
# Mon algorithme qui renvoie les prédictions sur le jeu de données
 # ‘dataset‘ fourni en argument.
 # ...
    return(predictions)
}

regresseur_mais <- function(dataset) {
# Chargement de l'environnement
load("env.Rdata")
# Mon algorithme qui renvoie les prédictions sur le jeu de données
# ‘dataset‘ fourni en argument.
# ...
    return(predictions)
}

classifieur_images <- function(list) {
# Chargement de l'environnement
# 'list' est une vecteur de chaînes de caractères contenant les chemins d'accès
# aux fichiers contenant les images de test
n<-length(list)
for(i in 1:n){
```

```
 # lecture de l'image dans le fichier de nom (avec chemin d'accès) list[i]
 }
# Mon algorithme qui renvoie les prédictions sur le jeu de données
# ‘dataset‘ fourni en argument.
# Le vecteur 'prediction' de longueur n contient les chaînes de caractère 'car',
# 'cat' et 'flower'
# ...
    return(predictions)
}
```

Remarques :

— Le rapport sera tronqué à 8 pages (12 pages si les trois problèmes sont traités). 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.