TP Analyse de données - Projets à choisir

Sujet 1 : Apprentissage non supervisé - Poissons et mercure

Base de données : poissons.xls.
Quelques informations sur cette base de données : contamination par le mercure dans différents organes de différentes espèces de poissons péchés en Guyanne Francaise, où certaines études ont mis en évidence des imprégnations par le mercure supérieures à la norme OMS dans les cheveux des populations amérindiennes. Certaines données sont manquantes. Elle sont asymétriques et contiennent un effet taille important (problème dans l'ACP).
Quelques pistes de travail : ACP (classique ; transformation des données asymétriques ; pourcentage pour l'effet taille) ; Pour le choix du nombre de composantes mettre en place une validation croisée ; Partitionnement ; Interprétation des résultats.
Remarque : poids (g), longueur (cm), concentration en mercure (μg/g pds sec) dans 6 organes : muscle, branchies, foie, intestins, estomac, reins.

Sujet 2 : Apprentissage supervisé (Regression) - Superconducteurs

Base de données d'apprentissage : superconductivity_data_train.rda
Base de données test : superconductivity_data_test.rda
Quelques informations sur cette base de données : prédire la température critique d'un superconducteur en utilisant 81 variables explicatives.
Quelques pistes de travail : Étude des variables (lesquelles sont importantes ? utilisation ACP ?) ; Appliquer différentes méthodes de régression (linéaire ; stepwise ; lasso ; elasticnet ; forêts aléatoires...) ; Valider et comparer ces méthodes sur la base de données d'apprentissage (validation croisée) avec le critère RMSE ; Proposer une stratégie et l'appliquer sur les données test (les vrais valeurs de températures seront données *a posteriori* : +1 pour le meilleur groupe !)
Remarque - Supraconducteurs :

- Atomic Mass (atomic mass units (AMU)) total proton and neutron rest masses
- First Ionization Energy (kilo-Joules per mole (kJ/mol)) energy required to remove a valence
- Atomic Radius (picometer (pm)) calculated atomic radius
- Density (kilograms per meters cubed (kg/m3)) density at standard temperature and pressure
- Electron Affinity (kilo-Joules per mole (kJ/mol)) energy required to add an electron to a neutral atom
- Fusion Heat (kilo-Joules per mole (kJ/mol)) energy to change from solid to liquid without temperature change
- Thermal Conductivity (watts per meter-Kelvin (W/(m $\times$ K))) thermal conductivity coefficient κ
- Valence (no units) typical number of chemical bonds formed by the element
- Pour certaines données plusieurs infos : wtd = weighted, gmean = geometric mean, std = standard deviation.

Marie Chavent - marie.chavent@u-bordeaux.fr

Sujet 3 : Apprentissage supervisé (Classification) Exploitations agricoles

Base de données d'apprentissage : farm_data_train.rda
Base de données test : farm_data_test.rda
Quelques informations sur cette base de données : les données concernent $n = 1260$ exploitations agricoles réparties en $K = 2$ groupes : le groupe des exploitations saines et le groupe des exploitations défaillantes. On veut construire un score de détection du risque financier applicable aux exploitations agricoles. Pour chaque exploitation agricole on a mesuré une batterie de critères économiques et financiers et finalement $p = 4$ ratios financiers on étés retenus pour construire le score :

- R2 : capitaux propres / capitaux permanents,
- R14 : dette à long et moyen terme / produit brut,
- R17 : frais financiers / dette totale,
- R32 : (excédent brut d'exploitation - frais financiers) / produit brut.

La variable qualitative à expliquer est donc la variable difficulté de paiement (0=sain et 1=défaillant) notée DIFF dans les données.
Quelques pistes de travail : Données équilibrées ; Appliquer différentes méthodes de classification ; Faire les courbes ROC et calculer le meilleur seuil ; Valider ces méthodes sur la base de données d'apprentissage (validation croisée) ; Proposer une stratégie et l'appliquer sur les données test (les vraies classes seront données *a posteriori* : +1 pour le meilleur groupe !)

Sujet 4 : Apprentissage supervisé (Classification) Spams

Base de données d'apprentissage : spam_data_train.rda
Base de données test : spam_data_test.rda.
Quelques informations sur cette base de données : les données concernent 4601 emails répartis en 2 groupes : le groupe des emails qui sont des spams, et le groupe des mails qui sont des hams. Chaque email est décrit par 57 variables quantitatives et 1 variable qualitative ("1=spam" et "0=ham"). Les variables quantitatives indiquent si un mot ou un caractère particulier est apparu fréquemment. Il y a :

- 48 variables du type word_freq_WORD = 100 x (nombre d'apparitions du mot WORD dans le courriel) / nombre total de mots dans le courriel.
- 6 variables du type char_freq_CHAR = 100 x (nombre d'apparition du caractère CHAR) / nombre total de caractères dans le courriel
- 1 variable capital_run_length_average = longueur moyenne des séquences de lettres majuscules consécutives.
- 1 variable capital_run_length_longest = longueur de la plus longue séquences de lettres majuscules consécutives
- 1 variable capital_run_length_total = somme des longueurs des séquences de lettres majuscules consécutives = nombre total de lettres majuscules dans le courriel.

Les données contiennent 39.5% de spams et 60.5% de hams. Elles sont réparties dans deux jeux de données spam_data_train.rda et spam_data_test.rda.
Quelques pistes de travail : Expliquer les variables explicatives ; Données quasi-équilibrées ; Intérêt d'une ACP pour diminuer le nombre de données ; Appliquer différentes méthodes de classification ; Faire les courbes ROC et calculer le meilleur seuil ; Valider ces méthodes sur la base de données d'apprentissage (validation croisée) ; Proposer une stratégie et l'appliquer sur les données test (les vraies classes seront données *a posteriori* : +1 pour le meilleur groupe !)