TP Analyse de données - Projets à choisir

Les données sont accessibles ici dans le dossier *Projet*.

1 : Apprentissage non supervisé - Poissons et mercure

Base de données : poissons.xls.
Quelques informations sur cette base de données : Contamination par le mercure de différents muscles de différentes espèces de poissons. 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 et/ou un bootstrap ; Partitionnement ; Interprétation des résultats.
Remarque : the length, the weight and the mercury concentration in μg/g in the muscle and in the 5 organs.

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 ; polynomial ; splines ; smoothing splines ; forêts aléatoires ; combinaison des méthodes) ; Valider ces méthodes sur la base de données d'apprentissage (validation croisée ou bootstrap) ; Proposer une stratégie et l'appliquer sur les données test (la solution sera donnée *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.

Annabelle Collin - annabelle.collin@inria.fr

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 (prototype et/ou modèle) ; Faire les courbes ROC et calculer le meilleur seuil ; Valider ces méthodes sur la base de données d'apprentissage (validation croisée ou bootstrap) ; Proposer une stratégie et l'appliquer sur les données test (la solution sera donnée *a posteriori* : +1 pour le meilleur groupe !)

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 (prototype et/ou modèle) ; Faire les courbes ROC et calculer le meilleur seuil ; Valider ces méthodes sur la base de données d'apprentissage (validation croisée ou bootstrap) ; Proposer une stratégie et l'appliquer sur les données test (la solution sera donnée *a posteriori* : +1 pour le meilleur groupe !)

5 : Apprentissage supervisé (EDO, Régression, Classification) Volumes tumeurs cérébrales (méningiomes)

Base de données d'apprentissage : meningiomas_train_set.csv
Base de données test : meningiomas_test_set.csv.
Quelques informations sur cette base de données : les données concernent 229 méningiomes **non traités**. Les données concernent le suivi dans le temps des patients. Pour chaque méningiome vous avez :

- des données sur le patient au diagnostique (au premier temps) : âge, sexe, position du méningiome (1-AM/NOS, 2-AM/TB, 3-SW/L , 4-SW/M, 5-PF/P , 6-PostFossa/SO, 7-PS/P, 8-PF/A, 9-TS/T, 10-OpticNerve, 11-NA) dans le crâne,
- des données de suivi du volume : 4 volumes (V0, V1, V2, V3) à 4 temps différents (t0, t1, t2, t3),
- une catégorie qui fait référence à leur agressivité : 1 peu agressif, 2 agressif.

Quelques pistes de travail : (1) Étudier la croissance à l'aide de modèles EDO (croissance linéaire : $V(t)' = c$, croissance exponentielle : $V(t)' = cV(t)$, croissance de type Gompertz : $V(t)' = ce^{-dt}V(t)$) ; (2) Prédire la catégorie Cat en utilisant différentes méthodes de classification (prototype et/ou modèle) ; (3) Prédire le dernier volume $V3$ (attention : il ne faut pas utiliser la catégorie Cat) en utilisant différentes méthodes de régression (linéaire ; polynomial ; splines ; smoothing splines ; forêts aléatoires ; combinaison des méthodes) ; (3bis) Prédire le volume $V2$ (attention : il ne faut pas utiliser la catégorie Cat ni le temps $t3$ et le volume $V3$) ; Proposer une stratégie et l'appliquer sur les données test (la solution sera donnée *a posteriori* : $+1$ pour le meilleur groupe !).