UNIVERSITÉ DE NANTES

Master 2 Ingénierie Mathématique

Apprentissage Statistique

Travaux Pratiques

1 Apprentissage sous R

On donne une liste de fonctions utiles sous `R`. Pour les détails d'utilisation, se référer à l'aide.

1.1 Statistiques descriptives

Les fonctions de bases:

- `mean, var, sd, median` etc... donnent respectivement la moyenne, la variance, l'écart-type et la médiane d'un échantillon
- `summary` résume l'information d'un jeux de données. La commande `stat.desc` du package `pastecs` contient encore plus d'options
- `plot, hist, boxplot, lines, abline, curve` etc... pour les représentations graphiques
- `na.omit, complete.cases` pour filtrer les individus contenant des données manquantes

1.2 Régression logistique

La régression logistique étant un cas particulier du modèle linéaire généralisé, elle est gérée par la commande `glm`. Les informations principales de la régression sont données par la commande `summary`. Parmi les autres fonctions importantes,

- `predict` permet de prédire une nouvelle valeur
- `deviance` calcule la deviance d'un modèle
- `confint` calcule des intervalles de confiance sur les paramètres
- `logLik` donne la log-vraisemblance
- `residuals` extrait les résidus

Pour la régression logistique multinomiale (la variable y plus de 2 modalités), utiliser la fonction `multinom` du package `nnet`.

1.3 Classification supervisée

Pour l'analyse factorielle discriminante, le package `DiscriMiner` contient de nombreuses fonctions utiles, notamment

- `desDA` effectue l'analyse discriminante
- `classify` attribue une modalité à de nouvelles données
- `binarize` convertit les facteurs (variables quantitatives) d'un jeu de données en indicatrices
- `betweenCov` estime la matrice de variance inter-classes Ω
- `withinCov` estime la matrice de variance intra-classes Σ
- `totalCov` estime la matrice de variance totale V
- `corRatio` calcule le rapport de corrélation entre une variable qualitative et quantitative
- `groupMeans` et `groupVars` calcule les moyennes et variances de chaque groupe

La classification par estimation de densité se fait facilement par

- `density` pour l'estimation non-paramétrique de la densité par la méthode à noyau
- `knn` du package `class` pour les k plus proches voisins

Pour analyser les résultats, on peut utiliser la fonction `performance` du package `ROCR`, qui nécessite de créer un objet de type prédiction par une commande du type `pred=prediction`($\hat{y}$,y).

1.4 Arbres binaires de décision

Les packages `rpart` et `rpart.plot` contiennent toutes les fonctions importantes pour la construction d'arbres de décision.

- `rpart` construit un arbre de décision à partir des données
- `prune` détermine une suite de sous-arbres emboîtés en faisant varier le paramètre λ du critère pénalisé
- `residuals` calcule le vecteur des résidus
- `predict` prédit une valeur de la variable à expliquer à partir de nouvelles données
- `prp` pour le plot

1.5 Support Vector Machine

La classification par SVM est gérée par la fonction `svm` du package `e1071`. Pour la prédiction et les représentations graphiques, on peut utiliser les commandes `predict.svm` et `plot.svm` du package `SparseM`.

2 Données simulées

On cherche ici à implémenter les méthodes d'apprentissage du cours sur des données simulées. Pour chaque question, les résultats et conclusions doivent être illustrés graphiquement et numériquement (tests, erreur de prédiction etc...). Dans un premier temps, créer une fonction qui simule un jeu de données de taille n contenant deux variables explicatives x_1, x_2 et une variable réponse y avec

- $(x_{1i})_{i=1,\dots,n}$ des réalisations iid de loi exponentielle $\mathcal{E}(1)$
- $(x_{2i})_{i=1,\dots,n}$ des réalisations iid de loi uniforme sur $[0,1]$
- $(y_i)_{i=1,\dots,n}$ des réalisations d'une variable binaire telle que $\mathbb{P}(Y_i = 1 | x_{1i} + 0.5x_{2i} > 1) = 0.9$ et $\mathbb{P}(Y_i = 1 | x_{1i} + 0.5x_{2i} \leq 1) = 0.2$

Quelle est la règle de décision Bayésienne δ^*? Proposer une méthode de classification de y en fonction de x_1, x_2 par:

- régression linéaire en fonction de (x_1, x_2)
- régression logistique
- les k plus proches voisins
- un arbre binaire de décision
- SVM

3 Jeux de données

L'objectif est d'appliquer les différentes méthodes d'apprentissage sur plusieurs jeu de données.

3.1 iris

Le jeu de données `iris` est disponible dans le package `datasets`.

1. Faire une étude descriptive des données. Quelles variables quantitatives vous semblent significatives pour expliquer la variable "espèce"?
2. Représenter sur un même graphique les relations entre "longueur du pétale", "longueur du sépale" et "espèce"
3. Faire la régression logistique de l'appartenance à l'espèce Setosa en fonction de la longueur du pétale. Commenter.
4. Faire la régression logistique de l'appartenance à l'espèce Setosa en fonction de la longueur du sépale et évaluer les performances de la discrimination (graphiques, courbes ROC, matrice de confusion etc...).

5. Faire la régression logistique multinomiale de l'espèce en fonction de la longueur du pétale.

6. Faire l'analyse factorielle discriminante de l'espèce en fonction des autres variables. Représenter les données dans le repère des deux premiers axes discriminants.

7. Proposer une procédure de discrimination par estimation de densité. Justifier le choix de la méthode d'estimation.

8. Implémenter l'algorithme des k plus proches voisins.

3.2 Mode

Les jeux de données `Mode_app` et `Mode_valid` contiennent les prix et temps de trajet pour quatre moyens de transport sur des échantillons de 400 et 53 individus. Connaissant ses données, l'objectif est de prédire le mode de transport choisi.

1. Implémenter différentes méthodes de classifications sur l'échantillon d'apprentissage en excluant les données manquantes, puis prédire le choix du mode de transport sur l'échantillon de validation.

2. Créer une fonction qui impute aléatoirement une valeur à une donnée manquante. Utiliser la loi uniforme sur l'échantillon bien renseigné, puis une loi qui attribue plus de poids aux individus ayant des caractéristiques plus proches.

3. Créer une fonction qui impute une valeur prédite (bruitée ou non) par la régression sur les variables disponibles.

4. Comparer les résultats de classification obtenus avec les différentes méthodes d'imputation.

3.3 Marketing

Le fichier `Marketing.txt` comprend les informations sur des clients d'un centre commercial, dont le revenu annuel `rev`, le sexe, l'état civil `ec`, l'age, le niveau d'étude `edu`, l'occupation professionnelle `prof`, la taille du foyer familial `fam`, la situation du logemant `sit_loge` et le type de logement `typ_log`. L'objectif est de prédire le revenu annuel à partir des autres variables disponibles.

3.4 Heart

Le fichier `heart.txt` comprend les relevés médicaux chez 300 patients sujets à des problèmes cardiaques. Pour chaque patient, on dispose de 6 variables: l'âge, le sexe, la pression sanguine `ps`, le taux de cholesterol `chol`, le rythme cardiaque au repos `rc` et le diagnostique `diag` de risque de problème cardiaque qui varie de 0 (risque faible) à 4 risque très élevé.

3.5 Library `rpart`

La library `rpart` contient plusieurs jeux de données dont `car.test.frame`, `kyphosis` ou `stagec`.