

IMPLÉMENTATION D'ALGORITHMES D'APPRENTISSAGE STATISTIQUE EN R: CAS DES ARBRES DE DÉCISION CART

Vous aurez probablement besoin pour ce TP des librairies R suivantes :

rpart, rpart.plot, tree

Partie 1 : Arbres CART de classification.

- (1) Importer la librairie **rpart** et charger le jeu de données *iris* dans R.
- (2) Consulter l'aide de cette base de données afin de la comprendre, et visualiser la.
- (3) Afficher un résumé des données : combien y a-t-il d'espèces et quelles proportions de chacune des espèces sont représentées ?
- (4) Modélisation : on va utiliser la fonction *rpart* pour créer un arbre de classification.
 - i) Comprendre l'utilisation de *rpart* : consulter l'aide de la fonction et...
 - a) : ...comprendre chacun de ses arguments ;
 - b) : ...comprendre ce qu'elle retourne (attributs de l'objet de classe **rpart**) ;
 - c) : ...comprendre sur quels paramètres vous jouerez en tant qu'utilisateur ;
 - ii) Implémenter l'algorithme : expliquer l'espèce d'iris à partir des mesures sur les pétales et sépales.
 - a) construire un arbre **maximal**, et afficher les résultats dans le terminal R en tapant le nom de l'objet créé ;
 - b) interpréter les résultats affichés : retrouver les taux d'erreur et les prévisions par noeud. Combien cet arbre a-t-il de feuilles ?
 - c) afficher graphiquement l'arbre avec les commandes **plot** et **text** ; Quel est le critère de segmentation qui semble le plus discriminant et à quel niveau ?
 - d) accéder aux attributs de l'objet créé. Combien y en a-t-il ? Les comprendre un par un. Finalement, quelle variable explicative est la plus importante ?
 - e) on se focalise sur l'attribut **cptable**. Que signifient les colonnes de cette matrice ? Afficher graphiquement avec **plotcp**, quel phénomène apparaît ?
 - f) on va élaguer l'arbre depuis ce résultat : quelle stratégie adopter ?
 - g) élaguer l'arbre maximal grâce à la commande **prune**, au meilleur niveau de complexité détecté. Combien de feuilles possède l'arbre élagué ?
 - h) afficher graphiquement l'arbre optimal : est-ce bien un sous-arbre de l'arbre maximal ? Quelle erreur minimise cet arbre ? Cet arbre est-il meilleur en pouvoir explicatif ou prédictif ?
 - i) effectuer des prévisions pour un jeu quelconque de caractéristiques sur les pétales et sépales.
- (5) Instabilité des arbres : retirer 5 observations des données (3%) et répéter la construction de l'arbre maximal puis de l'arbre optimal. Obtenez-vous le même résultat ?

Partie 2 : Arbres CART de régression.

- (1) Charger le jeu de données *cu.summary* dans le terminal R.
- (2) Consulter l'aide des données pour les comprendre, et visualiser. Quelle dimension ?
- (3) Afficher un résumé des données. Nous tenterons d'expliquer le prix des véhicules en fonction de l'ensemble des autres caractéristiques fournies.
- (4) Construire l'arbre maximal : quel argument de la fonction **rpart** faut-il absolument modifier par rapport à précédemment ?
- (5) Afficher les résultats numériques et interpréter : retrouver par exemple la valeur de l'homogénéité de la racine. À quoi correspond cette valeur ? Retrouver également la prévision du prix dans la racine.
- (6) Afficher graphiquement l'arbre maximal.
- (7) Poursuivre les mêmes étapes que dans la première partie pour son élagage, en illustrant le phénomène de surapprentissage.
- (8) De quelle taille est votre arbre optimal ? Quelle est (en pourcentage) la variance non-expliquée par votre estimateur ? Interpréter opérationnellement vos résultats.