

TP Partie 1 : Arbres CART

Guide pour le TP avec le scénario compagnon, la documentation et l'article :

hal-descartes.archives-ouvertes.fr/hal-01387654v2

Un compte-rendu de TP unique et global portant sur les deux parties, validera les deux journées de formation (à rédiger seul ou en binôme).

1. Les données

1. Charger la librairie `kernlab`
2. Charger le jeu de données `spam` dans R construire les *dataframes* d'apprentissage et de test (qui servira à évaluer les erreurs)

2. Arbres CART

1. Charger la librairie `rpart`
2. Calculer l'arbre par défaut fourni par `rpart`
3. Construire un arbre de profondeur 1 (`stump`) et le dessiner
4. Examiner les splits compétitifs et les splits de substitution
5. Construire un arbre maximal et le dessiner
6. Dessiner les erreurs de validation croisée de la suite de Breiman des sous-arbres élagués de l'arbre maximal et l'interpréter
7. Chercher le meilleur d'entre eux au sens d'une estimation de l'erreur de prédiction par validation croisée
8. Comparer l'arbre par défaut de `rpart` avec celui obtenu par minimisation de l'erreur de prédiction. Idem avec celui minimisant l'erreur de prédiction par la règle du 1 SE
9. Comparer les erreurs des différents arbres obtenus, en apprentissage et en test

3. Arbres CART

1. Charger la librairie `mlbench`
2. Charger le jeu de données `ozone` dans R et le présenter
3. Construire un modèle expliquant la concentration d'ozone, commenter les performances obtenues
4. Dégager les variables importantes