Arbres binaires de régression : CART

Travaux Pratiques

Dans ce TP, on revient sur le problème de classification binaire, où $Y_i \in \{-1, 1\}$ est expliqué par p régresseurs $X_i^1, \ldots, X_i^p$. On reprendra les exemples des TP précédents, pour leur appliquer cette fois la méthode de classification par arbres de régression CART : on comparera donc avec les méthodes de régression logistique, du perceptron et des K plus proches voisins vues précédemment.
La séance commencera par une présentation de CART, et par quelques propriétés théoriques. On pourra se référer à [3], chapitre 9.2.

On considérera dans CART les mesures d'impureté suivantes (à minimiser récursivement) :

- Indice de Gini : $2\widehat{p}_k(R)(1 - \widehat{p}_k(R))$
- Entropie croisée : $-\widehat{p}_k(R)\log(\widehat{p}_k(R)) - (1 - \widehat{p}_k(R))\log(1 - \widehat{p}_k(R))$.

- Arbres de régression -

R sait construire et élaguer des arbres de régression grâce au package `tree`.

1. Reprendre l'exemple simulé, et jouer sur le paramètre `control=tree.control` de la fonction `tree` pour générer des arbres de profondeurs différentes.
2. Tester les différents classifieurs obtenus sur de nouvelles données, et estimer leur risque.
3. Mettre en évidence le phénomène d'*overfitting* et l'équilibre biais-variance à trouver. Comparer les résultats obtenus sur les données réelles avec ceux des autres méthodes.
4. Quels résultats obtient-on pour l'explication du risque d'attaques cardiaques ?
5. Effectuer le même genre de comparaison sur les données issues de la base zipcode.

Références

[1] Pierre-André Cornillon, Arnaud Guyader, François Husson, Nicolas Jégou, Julie Josse, Maela Kloareg, Eric Matzner-Lober, and Laurent Rouviere. *Statistiques avec R.* Didact Statistiques. Presses Universitaires de Rennes, 2nd edition, 2010.

[2] Pierre André Cornillon and Eric Matzner-Lober. *Régression avec R.* Springer, Collection Pratique R, 1st edition, 2011.

[3] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning : data mining, inference and prediction.* Springer, 2 edition, 2009.