Arbres binaires de régression : CART

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. On pourra se référer à [2], chapitre 9.2 pour plus de détails sur les arbres. La source la plus détaillée sur le sujet étant le livre fondateur [1]. 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`. Une alternative si ce package n'est pas disponible ou non installé est d'utiliser le package `rpart`. Dans la suite on utilisera plutôt ce dernier. Pour plus d'aide sur ce dernier package voir par exemple http://www.statmethods.net/advstats/cart.html .

1. Reprendre l'exemple simulé des TP précédents, et jouer sur le paramètre `rpart.control` (notamment sur `maxdepth`) de la fonction `rpart` 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 de classification vue jusqu'à présent.
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] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth Statistics/Probability Series. Wadsworth Advanced Books and Software, Belmont, CA, 1984. 1

[2] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer, New York, second edition, 2009. http://www-stat.stanford.edu/~tibs/ElemStatLearn/. 1