Electronic Journal of Statistics > Vol. 1 (2007) open journal systems

Variable importance in binary regression trees and forests

Hemant Ishwaran

Abstract

We characterize and study variable importance (VIMP) and pairwise variable associations in binary regression trees. A key component involves the node mean squared error for a quantity we refer to as a maximal subtree. The theory naturally extends from single trees to ensembles of trees and applies to methods like random forests. This is useful because while importance values from random forests are used to screen variables, for example they are used to filter high throughput genomic data in Bioinformatics, very little theory exists about their properties.

Keywords: CART, random forests, maximal subtree.

Full Text: PDF

Ishwaran, Hemant, Variable importance in binary regression trees and forests, Electronic Journal of Statistics, 1, (2007), 519-537 (electronic). DOI: 10.1214/07-EJS039.

References

[1] Breiman L., Friedman J.H., Olshen R.A., and Stone C.J. Classification and Regression Trees. Wadsworth, Belmont, California, 1984. MR0726392

[2] Breiman L. Random forests. Machine Learning, 45:5–32, 2001.

[3] Friedman J. Greedy function approximation: a gradient boosting machine. Ann. Statist., 29, 1189–1232, 2001. MR1873328

[4] Liaw A. and Wiener M. Classification and regression by randomForest. R News, 2:18–22, 2002.

[5] Ishwaran H. and Kogalur U.B. Random survival forests for R. To appear in R News, 2007.

[6] Ishwaran H., Kogalur U.B, Blackstone E.H. and Lauer, M.S. Random survival forests. Cleveland Clinic Technical Report, 2007.

[7] Breiman L. Statistical modeling: the two cultures. Stat. Science, 16:199–231, 2001. MR1874152

[8] Lunetta K.L., Hayward L.B., Segal J. and Eerdewegh P.V. Screening large-scale association study data: exploiting interactions using random forests. BMC Genetics, 5:32, 2004.

[9] Bureau A., Dupuis J., Falls K., Lunetta K.L., Hayward B., Keith T.P., Eerdewegh P.V. Identifying SNPs predictive of phenotype using random forests. Genetic Epidemiology, 28:171-182, 2005.

[10] Diaz-Uriarte R. and Alvarez de Andres S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics, 7:3, 2006.

[11] van der Laan M. Statistical inference for variable importance. International J. Biosatist., 2:1008, 2006. MR2275897

[12] Strobl C., Boulesteix A.L., Zeileis A., and Hothorn T. Bias in random forest variable importance measures: Illustrations, sources and a solution. BMC Bioinformatics, 8:25, 2007.

[13] Breiman L. Population theory for boosting ensembles. Ann. Stat., 32:1–11, 2004. MR2050998

[14] Ng V.W. and Breiman L. Bivariate variable selection for classification problem. Technical report 692: Berkeley Statistics Department, 2005.

[15] Chambers J. M., Cleveland W. S., Kleiner B., and Tukey P. A. Graphical Methods for Data Analysis. Wadsworth, Belmont, California, 1983.

[16] Liaw A. and Wiener M. randomForest 4.5-18. http://cran.r-project.org, 2007.