

Title:

Risk Estimation For Classification Trees

Author(s):

Daniel A. Bloch, Richard A. Olshen, and Michael G. Walker

Technical Report number (Dept. of Statistics, Stanford Univ.):
215

Date:

February 2001

Abstract:

The purpose of this paper is to evaluate techniques for bias reduction of estimates of risk both globally and within terminal nodes of $CART^R$ classification trees, and to justify a method due Leo Breiman presented in Section 5.4 of Classification and Regression Trees as an estimator of risk within terminal nodes. The Breiman estimator has two free parameters, and an empirical Bayes method is put forth for estimating them. The book gives some evidence that the method is a good one, but only some. What is more, to the best of our knowledge no argument has been advanced, there or elsewhere, as to why the approach is as worthwhile as it is. Here we give an explanation, part heuristic but mostly mathematics, why it should be successful in the many examples for which it is. In addition, we give numerical evidence from simulations in the two-class case. Our simulations include ordinary resubstitution and seven other methods of estimation, including Breiman's. They involve 14 sampling distributions, all but one simulated and the remaining concerning *E. coli* promoter regions. We report on varying minimum node sizes of the trees we construct; prior probabilities and misclassification costs; and, when relevant, the numbers of bootstraps or cross-validations. Breiman's local risk estimator depends, in part, on the cross-validated global estimate of risk. A variation of Breiman's method in which repeated cross-validation is employed to estimate global rates of misclassification seems to be the most accurate overall from among those techniques we studied. Exceptions occur in cases for which the Bayes risk of the Bayes rule is small, for example when the unconditional probability of misclassification is less than .1 and also is much less than half the expected loss computed from prior probabilities and misclassification costs alone (the "no data Bayes rule"). For them, either a local bootstrap .632 estimate or Breiman's method modified to use a bootstrap estimate of the global misclassification rate is most accurate although the Breiman variant using repeated cross-validation is competitive for these data sets as well.