The University of Chicago

Department of Statistics

Seminar

Moulinath Banerjee
Department of Statistics
University of Michigan

"Confidence Sets for Split Points in Decision Trees"

Monday, November 17, 2003 at 4:00 PM
133 Eckhart Hall, 5734 S. University Avenue

ABSTRACT

Bühlmann and Yu (2002) recently obtained the limit distribution of the least squares estimator of split points in decision trees (CART). Their result is an instance of cube–root asymptotics with a non–normal limit and raises the interesting question of how to construct confidence sets for split–points. The bootstrap fails in this setting and the standard Wald type confidence intervals are found to be grossly inaccurate in terms of coverage probability. In this talk, we present a new method of constructing (asymptotic) confidence sets for the split point by inverting a centered residual sum of squares statistic in the decision tree problem under milder assumptions (in particular, we allow heteroscedastic errors). In contrast to the Wald-type confidence sets, the ones obtained via inversion are substantially more accurate, even at moderate sample sizes. This is illustrated using results from simulation experiments. We also study how to obtain confidence sets for split points for hazard functions in the setting of right–censored survival data.

The motivation for developing this new approach comes from the problem of phosphorus pollution in the Florida Everglades. Ecologists have suggested that split–points provide a phosphorus threshold at which biological imbalance occurs, and the lower endpoint of the confidence set may be interpreted as a level that is protective of the ecosystem. This is illustrated using data from a Duke University Wetlands Center phosphorus dosing study in the Everglades.

The above is joint work with Ian McKeague.