abstract

COMPUTER SCIENCE AND DISCRETE MATHEMATICS SEMINAR I
Topic:

Speaker:

Affiliation:

Date:

Time/Room:

A popular practical method of obtaining a good estimate of the error rate of a learning algorithm is k-fold cross-validation. Here, the set of examples is first partitioned into k equal-sized folds. Each fold acts as a test set for evaluating the hypothesis learned on the other k-1 folds. The average error across the k hypotheses is used as an estimate of the error rate. Although widely used, especially with small values of k (such as 10), the cross-validation method has heretofore resisted theoretical analysis due to the fact that the k distinct estimates have inherent correlations between them. With only sanity-check bounds known, there is no compelling reason to use the k-fold cross-validation estimate over a simpler holdout estimate.

Conventional wisdom is that the averaging in cross-validation leads to a tighter concentration of the estimate of the error around its mean. We show that the conventional wisdom is essentially correct. We analyze the reduction in variance of the gap between the cross-validation estimate and the true error rate, and show that for a large family of stable algorithms cross-validation achieves a near optimal variance reduction factor of (1 +o(1))/k . In these cases, the k different estimates are essentially independent of each other.

To proceed with the analysis, we define a new measure of algorithm stability, called mean-square stability. This measure is weaker than most stability notions described in the literature, and encompasses a large class of algorithms including bounded SVM regression and regularized least-squares regression. For slightly less stable algorithms such as t-nearest-neighbor, we show that cross-validation leads to an O(1/\sqrt{k}) reduction in the variance of the generalization error.