

Analysis of Variance of Cross Validation Estimators of the Generalization Error of Computer Algorithms

Marianthi Markatou, Hong Tian, George Hripcsak
Columbia University, Johnson & Johnson, Columbia University

Abstract: We bring together methods from statistics and machine learning to study the problem of estimating variances of cross validation estimators of the generalization error of computer algorithms. Providing an estimator of the variance of the cross validation estimator of the generalization error is a difficult problem, particularly if one wants to take into account various sources of variability. We provide a general framework which allows us to treat this problem as a problem in approximating the moments of a statistic. For the simple case of predicting the sample mean and when the loss functions are smooth, we show that the variance of the cross validation estimator of the generalization error is a function of the moments of the random variables Y , Z where Y denotes the cardinality of the intersection of two difference training sets and Z denotes the cardinality of the intersection of the corresponding test sets. We extend these results to regression, kernel regression and classification, and illustrate the methods through a health services research example on the hospital length of stay of patients that have undergone colon cancer surgery.