Boston University Statistics Seminar Series

# Variance Estimators of Cross-Validation Estimators of the Generalization Error

Marianthi Markatou
Dept of Biostatistics
Columbia University

Thursday, November 2, 2006, 4:00-5:00pm
Mathematics and Computer Science (MCS) Building, Room 149
111 Cummington Street, Boston

Tea and Cookies at 3:30pm in MCS 153

**Abstract:** We bring together methods from two different disciplines, machine learning and statistics, in order to address the problem of estimating the variance of cross-validation estimators of the generalization error. Specifically, we approach the problem of variance estimation of the CV estimators of the generalization error of computer algorithms as a problem in approximating the moments of a statistic. The approximation illustrates the role of training and tests sets in the performance of the algorithm. It provides a unifying approach to evaluation of various methods used in obtaining training and tests sets and it takes into account the variability due to different training and test sets. For the simple problem of predicting the sample mean and in the case of smooth loss functions, we show that the variance of the CV 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 different training sets and Z denotes the cardinality of the intersection of two different test sets. We prove that the distribution of these two random variables in hypergeometric and we compare our estimator with the estimator proposed by Nadeau and Bengio (2003). We extend these results to the regression case and the case of absolute error loss, and indicate how the methods can be extended to the classification case and the general case of kernel regression.

---

For directions and maps, please see *http://math.bu.edu/research/statistics/statseminar.html.*