On high-dimensional CV and APE

Ching-Kang Ing

National Tsing Hua University

Abstract

Cross validation (CV) has been one of the most popular methods for model selection. By splitting n data points into a training sample of size n_{c} and a validation sample of size n_{v} in which n_{v}/n approaches 1 and n_{c} tends to infinity, Shao (1993) showed that subset selection based on CV is consistent in a regression model of p candidate variables with p<< n. However, in the case of p >> n, not only does CV's consistency remain undeveloped, but subset selection is also practically infeasible. Instead of subset selection, in this talk, we suggest using CV as a backward elimination tool for excluding redundant variables that enter regression models through high-dimensional variable screening methods such as LASSO, LARS, ISIS, and OGA. By choosing a n_{v} such that n_{v}/n converges to 1 at a rate faster than the one suggested by Shao (1993), we establish selection consistency of the proposed method. Accumulated prediction error (APE), on the other hand, can be viewed as a counterpart of CV in situations where a random split of data is pointless (e.g., when data are serially correlated). While APE’s behavior in the case of p<<n has been well understood, no results have been reported regarding its performance in the high-dimensional case. To fill this gap, we provide a high-dimensional amendment of APE and justify its asymptotic validity. Simulation evidence will also be furnished if time permits.