

STATISTICS/DEAS (EE) JOINT COLLOQUIUM

WEDNESDAY, NOVEMBER 08, 2006
TALK: 4:00 PM — Cruft Laboratory, Rm. 318

“Feature Selection through Lasso: model selection consistency and the Blasso algorithm”

Bin Yu
Statistics Department
UC Berkeley

ABSTRACT

Information technology advances are making data collection possible in most if not all fields of science and engineering and beyond. Statistics as a scientific discipline is challenged and enriched by the new opportunities resulted from these high-dimensional data sets. Often data reduction or feature selection is the first step towards solving these massive data problems. However, data reduction through model selection or l_0 constrained least squares optimization leads to a combinatorial search which is computationally infeasible for massive data problems. A computationally efficient alternative is the l_1 constrained least squares optimization or Lasso optimization.

In this talk, we first study the model selection property of Lasso in linear regression models. We show that an Irrepresentable Condition on the design matrix is almost necessary and sufficient for the model selection consistency of Lasso for fixed p and $p \gg n$ cases, provided that the true model is sparse. Moreover, we describe the Boosted Lasso (BLasso) algorithm which produces an approximation to the complete regularization path of Lasso. BLasso consists of both a forward step and a backward step. The forward step is similar to Boosting and Forward Stagewise Fitting, but the backward step is new and crucial for BLasso to approximate the Lasso path in all situations. For cases with finite number of base learners, when the step size goes to zero, the BLasso path is shown to converge to the Lasso path. Finally, the Blasso algorithm is extended to give an approximate path for the case of a convex loss function plus a convex penalty.