

STATISTICS COLLOQUIUM

FRIDAY, FEBRUARY 16, 2007

TALK: 4:00 PM — SCIENCE CENTER RM. 705

RECEPTION: 5:15 PM — SCIENCE CENTER, 7TH FLOOR

High Dimensional Classification Using Features Annealed Independence Rules

Yingying Fan

Operations Research and Financial Engineering
Princeton University

ABSTRACT

Classification using high-dimensional features arises frequently in many contemporary statistical studies such as tumor classification using microarray or other high-throughput data. The impact of dimensionality on classifications is largely poorly understood. In a seminal paper, Bickel and Levina (2004) show that the Fisher discriminant performs poorly due to diverging spectra and they propose an independence rule to overcome the problem. We first demonstrate that even for the independence classification rule, classification using all features can be as bad as the random guessing due to noise accumulation in estimating population centroids in high-dimensional feature space. In fact, we demonstrate further that almost all linear discriminants can perform as bad as the random guessing. Thus, it is of paramount important to select a subset of important features for high-dimensional classification, resulting in Feature Annealed Independence Rules (FAIR). The conditions under which all the important features can be selected by the two-sample t -statistic are established. The choice of the optimal number of features, or equivalently, the threshold value of the test statistics are proposed based on an upper bound of the classification error. Simulation studies and real data analysis strongly support our theoretical results and demonstrate convincingly the advantage of our new classification procedure.

This talk is based on collaborated work with Professor Jianqing Fan.