Developments in Data Mining and Machine Learning: An Overview

David Banks*
Duke University, USA – banks@stat.duke.edu

Abstract: Over the last twenty years, new technology has changed mathematical statistics into computational statistics. Nonparametric regression, Bayesian nonparametrics, and complex classification algorithms have emerged. In parallel, there are new heuristics, such as the value of sparsity, regularization, latent spaces and ensembles. And there are fresh criteria, such false discovery rates, parallelizability, and speed and memory requirements. Finally, our problem set has changed---now, instead of proving assumption-laden theorems, many statisticians pursue high-profile applications such as the Netflix prize, the LinkedIn economic graph challenge, or various information technology applications in genomics, computational advertising, and dynamic network modeling. This lecture attempts to convey a broad-brush overview of the modern research landscape in statistics, with particular attention to how the modern computational methodology informs regression and classification. The ideas are motivated by practical examples, mostly drawn from the information technology industries, although comparable examples exist in medicine, administrative data bases, astrostatistics, and many other fields.

Keywords: Random Forests, Nonparametric Regression, Nonparametric Classification, Boosting, Ensembles, Support Vector Machines, Sparsity