EIGHT U.S. ARMY CONFERENCE ON APPLIED STATISTICS

SHORT COURSE
Statistical Data Mining

Dr. Edward J. Wegman   Dr. Jeffrey L. Solka
George Mason University   Naval Surface Warfare Center

Abstract: Used to extract knowledge hidden from large volumes of raw data, data mining has been applied in recent years to a wide variety of research areas such as medical imaging, high-energy physics, consumer behavior, atmospheric sciences, and security and surveillance. Data mining is an extension of exploratory data analysis and has basically the same goals, the discovery of unknown and unanticipated structure in the data. The chief distinction between the two topics resides in the size and dimensionality of the data sets involved. Data mining in general deals with much more massive data sets for which highly interactive analysis is not fully feasible. In this course we shall discuss the scales of data set sizes and the limits of feasibility for the various data set sizes. We will introduce some visualization tools and indicate how they may be used to accomplish data mining tasks. We shall review some structure finding algorithms including: density estimation and bump hunting; clustering and classification; visual clustering strategies; CART and related methods; time domain time series methods; nonparametric regression including convolution, LOESS and ridges and skeletons methods will be illustrated with application to several data sets. Particular emphasis will be placed on visualization techniques. The course will cover basic techniques used in visual data mining, including parallel coordinates, grand tour, and saturation brushing. These techniques will be illustrated in further discussions on rapid data editing, density estimation, inverse regression, tree-structured decision rules, classification and clustering, structural inference and outlier investigation.