Università Commerciale Luigi Bocconi

Department of Decision Sciences

Statistics Seminar

A new modern approach to statistical data mining using information complexity and the genetic algorithm

Hamparsum Bozdogan

The University of Tennessee, USA

Monday, 14 September 2009
2:30pm Room N31 Piazza Sraffa 13 Milano

Tuesday, 15 September 2009
9:30am Room N30 Piazza Sraffa 13 Milano

Abstract

In the information age we live in today, with increasingly sophisticated technology for gathering and storing data, many organizations and businesses collect massive amounts of data at accelerated rates and in ever-increasing detail. Massive data sets pose a great challenge to many cross-disciplinary fields. Such data sets have large number of dimensions and often have huge numbers of observations. They are categorical, discrete, quantitative, and often are mixed-data types. This high dimensionality and different data types and structures have now outstripped the capability of traditional statistical methods, data analysis, graphical and data visualization tools.

Researchers, practitioners, and Ph.D. students at all levels need new and modern highly clever model selection tools as a key tool embedded into fast and efficient algorithms to reduce dimensionality and to choose best subset of variables in regression, classification, and cluster analysis to mention a few.

To address these challenges, this workshop seminar will present a new statistical model selection tools to analyze potentially massive data sets with the goal of identifying a parsimonious model structure to fit to the data.

To this end, during the first day of the seminar, we will present:

- A Short Overview of Data Mining & Statistical Methodology During the Last Five Years
- The General Theory and Applications of a New Information Complexity (ICOMP) Class Criteria for Model Selection
- Simultaneous Model Selection in Multivariate Mixture Model Cluster Analysis Using Information Complexity and the Genetic Algorithm: M3

During the second day of the seminar we will present:

- New Advances in Regression Model Selection Using Information Complexity and the Genetic Algorithm
- Lab Applications of Model SelectionHowever, we argue that, since parameter uncertainty can have a big influence on spreads, Bayesian approaches are more appropriate in this context and allow us to partially explain the "credit spread puzzle".

This is joint work with Samuel Malone (Universidad de los Andes) and Enrique ter Horst (Euromed Management).

Department of Decision Sciences

Via Röntgen 1 - 20136Milano

Tel. 02 5836.5632
Fax 02 5836.5630