Sesión Especial 12

Experimental design techniques applied to treatment of Big Data

Organizadores

- Jesús López Fidalgo (Universidad de Navarra)
- José Antonio Moler Curiel (Universidad Pública de Navarra)

Descripción

With the onslaught of huge quantities of data and the need to extract information from them with an appropriate model, model selection has become tremendously important for research. Experimental design attempts to properly organize this information. This session aims to merge the two fields in search of a powerful tool for optimizing information collection and analysis. The novelty of this session is found in joining approaches from both modelling and experimental design theory.

Programa

Jueves, 7 de febrero (mañana)

11:30 – 12:00	John Stufken (Charles Wexler Professor, Arizona State University) *Information-Based Optimal Subdata Selection*
12:00 – 12:30	Chiara Tommasi (University of Milan) *Optimal design theory: A device to select a good sample from big data*
12:30 – 13:00	Jesús López-Fidalgo (University of Navarre) *Model-Robust Classification in Active Learning*
13:00 – 13:30	Noèlia Viles (Data Scientist at Nestlé) *A practical view from a Mathematician Data scientist in a company*

Information-Based Optimal Subdata Selection

John Stufken

Charles Wexler Professor, Arizona State University

jstufken@asu.edu

Abstract. The focus of this presentation is on the analysis of data with a very large number of cases, n, and a modest number of variables, p. The size of n, or the lack of access to a sufficiently powerful computing platform, might necessitate or suggest the use of subdata, which consists of only some of the cases. How should one select such subdata? In the linear regression context, several subsampling methods have been proposed (Ma, Mahoney and Yu, 2015) to obtain such subdata, along with appropriate estimation methods. Such methods have been referred to as algorithmic leveraging methods.
Also in the context of linear regression, Wang, Yang and Stufken (2018) proposed a deterministic method for subdata selection, referred to as Information-Based Optimal Subdata Selection (IBOSS). This method borrows ideas from design of experiments to select subdata that provides "maximum information".
In this presentation, I will briefly introduce the different methods for subdata selection, with an emphasis on the IBOSS method. Selected results and comparisons from Wang, Yang and Stufken (2018) will be presented. I will conclude with a brief discussion of remaining challenges in this area of research.

Referencias

[1] P. Ma, M. W. Mahoney, & B. Yu. A statistical perspective on algorithmic leveraging. Journal of Machine Learning Research, 16:861–911, 2015.

[2] Wang, H. Y., Yang, M. & Stufken, J. Information-Based Optimal Subdata Selection for Big Data Linear Regression. Journal of the American Statistical Association, 2018 (in press).

Optimal design theory: A device to select a good sample from big data

Chiara Tommasi

University of Milan (Italy)

chiara.tommasi@unimi.it

Abstract. Big Data are generally huge quantities of digital information accrued automatically and/or merged from several sources and rarely result from properly planned population surveys. A Big Dataset is herein conceived as a collection of information concerning a finite population. Since the analysis of an entire Big Dataset can require enormous computational effort, we suggest selecting a sample of observations and using this sampling information to achieve the inferential goal. Instead of the design-based survey sampling approach (which relates to the estimation of summary finite population measures, such as means, totals, proportions...) we consider the model-based sampling approach, which involves inference about parameters of a super-population model. This model is assumed to have generated the finite population values, i.e. the Big Dataset. Given a super-population model we can apply the theory of optimal design to draw a sample from the Big Dataset which contains the majority of information about the unknown parameters of interest.

Joint work with Laura Deldossi, University Cattolica del Sacro Cuore (Italy)

Model-Robust Classification in Active Learning

Jesús López-Fidalgo

University of Navarre

fidalgo@unav.es

Abstract. We aim to develop a theory of model-robust classification, and a methodology for applying this to large data sets such as arise in machine learning. The general idea is that there is a (large) population of explanatory variables, which can be easily sampled. With probability $a(x;t)$, an item with covariates x belongs to group A and with probability $1-a(x;t)$ it belongs to group "B". We suppose that the determination of the appropriate group, given x, is difficult and expensive, so that the investigator wishes to sample from x in a manner which is more effcient than random sampling (sometimes termed "passive learning").

Joint work with José Antonio Moler (University of Navarre) and Douglas P. Wiens (University of Alberta, Canada)

A practical view from a Mathematician Data scientist in a company

Noèlia Viles

Data Scientist at Nestlé, Barcelona, Cataluña

noelia.viles@gmail.com

Abstract. Each day, more and more organizations are opening up their doors to big data and increasing the value of a data science team who knows how to tease actionable insights out of gigabytes of data. We will see the benefits of data science in the companies, what a data scientist do in a company and how can add value to a business.