Field:

Theoretical and Applied Mathematics/Informatics

Session Topic:

Statistics for Large-Dimensional Data

Introductory Speaker:

Hidetoshi SHIMODAIRA, Tokyo Institute of Technology

In statistics/machine learning field, several approaches have been developed to handle the difficulty of large dimensions. In this session, we attempt to illustrate a coherent idea behind them: How can we convert complicated problems in large dimensions into a simple representation?

The reality may be represented in infinite dimensions. A classical approach is to assume a simple mathematical model to describe the data, and this has been very successful until recently. However we are now facing a flood of complicated large data with many variables, and the mathematical model is specified in large dimensions. The classical approach is then suffered from difficulties in computation time and lack of information.

We present three modern approaches to handle the difficulties of large dimensions. Although they are developed independently in statistics/machine learning field, there are some similarities among them in the ways how data is represented. They all utilize simple data representations for effectively limiting the number of dimensions by the data length.

The first approach I am going to show briefly is the plug-in principle. This simply treats the data “as is” without using a mathematical model. We replace the probability distribution of data with the observed histogram of the data. This idea leads to a powerful computer simulation method, called bootstrap resampling, for evaluating the confidence level of data analysis. The idea will be illustrated in a real data analysis of inferring the evolutionary history of mammal species. I also include a recent theoretical development of the methodology, which utilizes the scaling-law of the confidence level by changing the data length.

Kenji Fukumizu will show the second approach, called the kernel trick. The idea is to hold only a matrix representing relations between data elements. The matrix size is always limited by the data length even if the data elements are given in infinite dimensions. Alexandre d'Aspremont will show the third approach, called sparse representation. The idea is inspired by the sparse coding in brain, where only a small number of neurons may activate together. Both these ideas lead to very efficient data analysis.