COMPSTAT TUTORIAL 2010

Complex Data - Symbolic Data Analyses

L. Billard, University of Georgia

Abstract

Complex data can take a myriad of formats covering a vast number of different settings. This tutorial will focus on one type of non-standard data that differs from the standard classical data, that is, so-called symbolic data.

Symbolic data appear in numerous settings, in all avenues of the sciences and social sciences, from medical, industry and government experiments, and data collection pursuits. Some data are inherently symbolic. Some, perhaps most, arise as the result of the massive datasets that emerge from contemporary computer capacity. Such datasets have to be aggregated in some meaningful way (with the actual aggregation being instructed by the scientific questions of interest). An introduction to symbolic data and how such data can be analysed will be presented. Classical data on p random variables are represented by a single point in p-dimensional space $\Re^p$. In contrast, symbolic data with measurements on p random variables are p-dimensional hypercubes in $\Re^p$, or a cartesian product of p distributions, broadly defined. There are many possible formats for symbolic data.

Basic descriptions of symbolic data and how they contrast with classical data are covered first. For example, it may not be possible to give the exact cost of an apple (or shirt, or product, or ...), or the exact pulse rate measurement, but only its value in the range [66, 74], (say). We note also that an interval value of [66, 74] differs from the interval [68, 72] even though they both have the same midpoint value of 70. A classical analysis using the same midpoint (70) would lose the information that these are two differently valued realizations with different internal variations.

Methodologies for obtaining basic descriptive statistics for random variable whose values are symbolic valued, viz., a histogram and its associated empirical probability distribution, along with the empirical mean, variance, and covariance, will be presented. This will be followed by methodologies that deal with clustering, principal components, and regression (if time permits, including regression methodologies for handling taxonomy tree structures and hierarchy tree structures), respectively.

These methods are extensions of well-known classical theory applied or extended to symbolic data. Our approach assumes knowledge of the classical results, with the focus on the adaptation to the symbolic data setting. Therefore, only minimal classical theory is provided.

REFERENCES

Billard, L. and Diday, E. (2006). *Symbolic Data Analysis: Conceptual Statistics and Data Mining* John Wiley, Chichester

Bock, H. -H. and Diday, E. (eds.) (2000). *Analysis of Symbolic Data: Exploratory Methods for Extracting Statistical Information from Complex Data.* Springer-Verlag, Berlin.

Diday, E. and Noirhomme, M. (eds.). (2008). *Symbolic Data and the SODAS Software.* John Wiley, Chichester.

TOPIC Details

1. Symbolic Data
   - Multi-valued, lists, categorical data
   - Interval-valued
   - Modal-valued (Modal categorical, Histograms, Distributions, ...
   - How do symbolic data arise?
2. Descriptive Statistics
   - Sample mean, variance, covariance
   - Histograms, joint histograms
   - Rules
3. Principal Components
   - Vertices method
   - Data matrices
   - Variance-covariance matrices
   - Diagnostics
   - Weights
   - Centers method

4.Clustering

   - Distance Measures, Similarity/Dissimilarity Matrix
     - Multi-valued data
     - Interval-valued data
     - Mixed-valued data
   - Types of clusters
   - Building algorithms
   - Partitions
   - Divisive Clustering
   - Hierarchy-Pyramid Clusters

5. Regression
   - Linear regression
   - Taxonomies, Hierarchies