DSS Statistics Seminar

November 19, 2021, 15:00

https://uniroma1.zoom.us/j/86881977368?pwd=SWRFcVFjMDZTa0lXZk05TE1zNm5adz09

Passcode: 432940

Advances in Model-based Clustering and Outlier Detection with missing data

Cristina Tortora

San Jose State University

Model-based clustering assumes that the data were generated from a convex combination of densities. The choice of the density function is crucial; the multivariate contaminated normal distribution (MCN) was proposed to model datasets characterized by the presence of outliers. The MCN is a two-component Gaussian mixture; one of the components, with a large prior probability, represents the good observations, and the other, with a small prior probability, the same mean, and an inflated covariance matrix, represents the outliers. Mixtures of MCN distributions can detect outliers and perform cluster analysis improving the clustering performance when compared to normal mixtures and representing an alternative to t mixtures. However, the MCN distribution has two main drawbacks, it uses univariate parameters to model the proportion of outliers and their impact on the inflation parameter, i.e., they are the same for all the variables, and it cannot work when the data have missing values. To overcome this issue, we propose a multiple scaled contaminated normal distribution with p-dimensional proportion of outliers and degrees of contamination, where p is the number of variables and we extended both distributions for data sets with data missing at random.

SAPIENZA
UNIVERSITÀ DI ROMA