PhD program in Statistics

DSS Statistics Seminar

April 28, 2023, 12:00

https://uniroma1.zoom.us/j/86881977368?pwd=SWRFcVFjMDZTa0lXZk05TE1zNm5adz09
Passcode: 432940

Convex clustering of mixed numerical and categorical data

Carlo Cavicchia
Econometric Institute, Erasmus University Rotterdam

Clustering analysis is an unsupervised learning technique widely used for information extraction. Current clustering algorithms often face instabilities due to the non-convex nature of their objective function. The class of convex clustering methods does not suffer from such instabilities and finds a global optimum for the clustering objective. Whereas convex clustering has previously been established for single-type data, real-life data sets usually comprise both numerical and categorical, or mixed, data. Therefore, we introduce the mixed data convex clustering (MIDACC) framework. We implement this framework by developing a dedicated subgradient descent algorithm. Through numerical experiments, we show that, in contrast to baseline methods, MIDACC achieves near-perfect recovery of both spherical and non-spherical clusters, is able to capture information from mixed data while distinguishing signal from noise, and has the ability to recover the true number of clusters present in the data. Furthermore, MIDACC outperforms all baseline methods on a real-life data set.

SAPIENZA
UNIVERSITÀ DI ROMA