Convex Clustering of Mixed Numerical and Categorical Data

Carlo Cavicchia

Erasmus School of Economics, Erasmus University of Rotterdam

Clustering analysis is an unsupervised learning technique widely used for information extraction. Current clustering algorithms often face instabilities due to the nonconvex 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. MIDACC combines likelihood-based loss functions for numerical and categorical data, weighted by parameters controlling the importance of both data types on the clustering outcome. The penalty term fuses centroids, and thus allows for clustering of observations. The presence of few parameters characterizes MIDACC and allows the user to tailor the analysis for the problem at hand. For instance, in contrast to, who perform subgradient descent (in the case of numerical data only) for a path for values of the regularization term in order to retrieve an entire clusterpath, we run our algorithm for fixed values of it. We therefore present our framework as a partitional clustering methods and we do not provide the solution in a hierarchical fashion. Another crucial parameter is the one controlling the contribution of the data type; this one is fine tuned to obtain the optimal solution. We present two different implementations for this framework. The first consists of a dedicated subgradient descent algorithm. However, our current implementation follows a majorization-minimization approach (similar to what propose for only numerical data) that results remarkably faster and more efficient. 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.