A Robust Spectral Clustering Algorithm for Sub-Gaussian Mixture Models with Outliers

We consider the problem of clustering datasets in the presence of arbitrary outliers. Traditional clustering algorithms such as k-means and spectral clustering are known to perform poorly for datasets contaminated with even a small number of outliers [García-Escudero and Gordaliza, 1999, Li et al., 2007, Rujeerapaiboon et al., 2017]. In this paper, we develop a provably robust spectral clustering algorithm that applies a simple rounding scheme to denoise a Gaussian kernel matrix built from the datapoints, and uses vanilla spectral clustering to recover the cluster labels of data points. We analyze the performance of our algorithm under the assumption that the “good” data points are generated from a mixture of sub-gaussians (we will call these “inliers”), while the noisy (outlier) points can come from any arbitrary probability distribution. For this general class of models, we show that the asymptotic mis-classification error decays at an exponential rate in the signal-to-noise ratio, provided the number of outliers grow slowly compared to the number of inlier points. Surprisingly, the derived error bound matches with the best-known bound [Fei and Chen, 2018, Giraud and Verzelen, 2018] for semidefinite programs (SDPs) under the same setting without outliers, under suitable conditions on the outlier points. We conduct extensive experiments on a variety of simulated and real world datasets to demonstrate that our algorithm is less sensitive to outliers compared to other state-of-the-art algorithms proposed in the literature, in terms of both accuracy as well as scalability.