9.11 Euclidean Clustering

The stochastic block model, although having fascinating phenomena, is not always an accurate model for clustering. The independence assumption assumed on the connections between pairs of vertices may sometimes be too unrealistic. Also, the minimum bisection of multisection objective may not be the most relevant in some applications.

One particularly popular form of clustering is k-means clustering. Given n points $x_1, \ldots, x_n$ and pairwise distances $d(x_i, x_j)$, the k-means objective attempts to partition the points in k clusters $A_1, \ldots, A_k$ (not necessarily of the same size) as to minimize the following objective[35]

$$\min \sum_{t=1}^{k} \frac{1}{|A_t|} \sum_{x_i, x_j \in A_t} d^2(x_i, x_j).$$

A similar objective is the one in k-medians clustering, where for each cluster a center is picked (the center has to be a point in the cluster) and the sum of the distances from all points in the cluster to the center point are to be minimized, in other words, the objective to be minimized is:

$$\min \sum_{t=1}^{k} \min_{c_t \in A_t} \sum_{x_i \in A_t} d(x_i, c_t).$$

In [ABC+15] both an Linear Programming (LP) relaxation for k-medians and a Semidefinite Programming (SDP) relaxation for k-means are analyzed for a points in a generative model on which there are k disjoint balls in $\mathbb{R}^d$ and, for every ball, points are drawn according to a isotropic distribution on each of the balls. The goal is to establish exact recovery of these convex relaxations requiring the least distance between the balls. This model (in this context) was first proposed and analyzed for k-medians in [NW13], the conditions for k-medians were made optimal in [ABC+15] and conditions for k-means were also given. More recently, the conditions on k-means were improved (made optimal for large dimensions) in [IMPV15a, IMPV15b] which also coined the term "Stochastic Ball Model".

For P the set of points, in order to formulate the k-medians LP we use variables y_p indicating whether p is a center of its cluster or not and z_{pq} indicating whether q is assigned to p or not (see [ABC+15] for details), the LP then reads:

$$\begin{array}{ll} \min & \sum_{p,q} d(p,q) z_{pq}, \\ s.t. & \sum_{p \in P} z_{pq} = 1, \quad \forall q \in P \\ & z_{pq} \leq y_p \\ & \sum_{p \in P} y_p = k \\ & z_{pq}, y_p \in [0,1], \quad \forall p, q \in P. \end{array}$$

the solution corresponds to an actual k-means solution if it is integral.

The semidefinite program for k-means is written in terms of a PSD matrix $X \in \mathbb{R}^{n \times n}$ (where n is the total number of points), see [ABC+15] for details. The intended solution is

$$X = \frac{1}{n} \sum_{t=1}^{k} \mathbf{1}_{A_t} \mathbf{1}_{A_t}^T,$$

[35] When the points are in Euclidean space there is an equivalent more common formulation in which each cluster is assign a mean and the objective function is the sum of the distances squared to the center.

where $\mathbf{1}_{A_t}$ is the indicator vector of the cluster A_t. The SDP reads as follows:

$$\begin{array}{cc} \min_X & \sum_{i,j} d(i,j) X_{ij}, \\ s.t. & \mathrm{Tr}(X) = k, \\ & X\mathbf{1} = \mathbf{1} \\ & X \geq 0 \\ & X \succeq 0. \end{array}$$

Inspired by simulations in the context of [NW13] and [ABC$^+$15], Rachel Ward observed that the k-medians LP tends to be integral even for point configurations where no planted partition existed, and proposed the conjecture that k-medians is tight for typical point configurations. This was recorded as Problem 6 in [Mix15]. We formulate it as an open problem here:

Open Problem 9.3 *Is the LP relaxation for k-medians tight for a natural (random) generative model of points even without a clustering planted structure (such as, say, gaussian independent points)?*

Ideally, one would like to show that these relaxations (both the k-means SDP and the k-medians LP) are integral in instances that have clustering structure and not necessarily arising from generative random models. It is unclear however how to define what is meant by "clustering structure". A particularly interesting approach is through stability conditions (see, for example [AJP13]), the idea is that if a certain set of data points has a much larger $k-1$-means (or medians) objective than a k-means (or medians) one, and there is not much difference between the k and the $k+1$ objectives, then this is a good suggestion that the data is well explained by k clusters.

References

[AJP13] M. Agarwal, R. Jaiswal, and A. Pal. k-means++ under approximation stability. *The 10th annual conference on Theory and Applications of Models of Computation*, 2013.

[ABC$^+$15] P. Awasthi, A. S. Bandeira, M. Charikar, R. Krishnaswamy, S. Villar, and R. Ward. Relax, no need to round: integrality of clustering formulations. *6th Innovations in Theoretical Computer Science (ITCS 2015)*, 2015.

[IMPV15a] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. On the tightness of an sdp relaxation of k-means. *Available online at arXiv:1505.04778 [cs.IT]*, 2015.

[IMPV15b] T. Iguchi, D. G. Mixon, J. Peterson, and S. Villar. Probably certifiably correct k-means clustering. *Available at arXiv*, 2015.

[Mix15] D. G. Mixon. Applied harmonic analysis and sparse approximation. *Short, Fat Matrices Web blog*, 2015.

[NW13] A. Nellore and R. Ward. Recovery guarantees for exemplar-based clustering. *Available online at arXiv:1309.3256v2 [stat.ML]*, 2013.

18.S096 Topics in Mathematics of Data Science
Fall 2015