The integrality of k-median clustering solutions when input data set has no clustering feature

Yunqiu Guo & Mutiara Sondjaja

New York University

October 26, 2018

Outline

Outline

Background

- What is a clustering problem?
  - Definition of Clustering Problem
  - Example of Clustering Problem
- Brief Intro to Linear Programming and Relaxation
  - k-median clustering algorithm

A Short Definition of Clustering Problem

Clustering : Determine the intrinsic grouping in a set of unlabeled data.

A simple illustration of Clustering Problem

(a) Original points.

(b) Two clusters.

(c) Four clusters.

(d) Six clusters.

Figure 1: Simple clustered results

Clustering

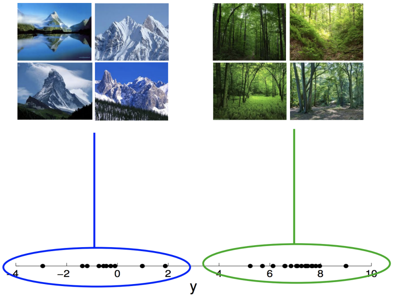


Figure 2: Clustering applied to picture pixels

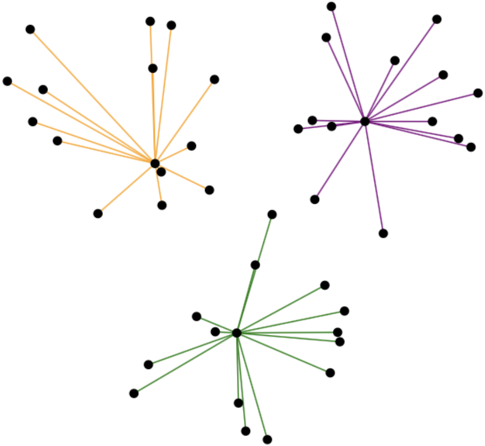
k-median clustering

The k-median clustering objective minimizes the sum of distances from points to their representative data points.

In the k-median (also known as k-medroid) problem, clusters are specified by centers: k representative points from within the set P denoted by $c1, c2, ..., ck$. The corresponding partitioning is obtained by assigning each point to its closest center. The cost incurred by a point is the distance to its assigned center, and the goal is to find k center points that minimize the sum of the costs of the points in P.

$$\min_{c_1, c_2, \ldots, c_k \subset P} \sum_{i=1}^{n} \min_{j=1,2,3,\ldots k} d(x_i, c_j)$$

k-median clustering



Integer Programming Formulation to k-median clustering

$$
\begin{array}{lccc}
\min_{z \in \mathbb{R}^{\mathbb{N} \times \mathbb{N}}} & \sum_{p,q \in P} d(p,q) z_{pq} & & \\
s.t. & \sum_{p \in P} z_{pq} & = & 1 \\
 & z_{pq} & \leq & y_p \\
 & \sum_{p \in P} y_p & = & k \\
 & z_{pq} & \in & \{0,1\} \\
 & y_p & \in & \{0,1\}
\end{array}
$$

In the above linear programming formulation, the variable y_p indicates whether the point p is a center or not, while z_{pq} is 1 if the point q is assigned to p as center, and 0 otherwise.

Relaxation to the *k*-median Linear Programming

$$
\begin{array}{lccc}
\min_{z \in \mathbb{R}^{\mathbb{N}\times\mathbb{N}}} & \sum_{p,q \in P} d(p,q) z_{pq} & & \\
s.t. & \sum_{p \in P} z_{pq} & = & 1 \\
 & z_{pq} & \leq & y_p \\
 & \sum_{p \in P} y_p & = & k \\
 & z_{pq} & \in & [0,1] \\
 & y_p & \in & [0,1]
\end{array}
$$

The above modified linear programming formulation serves as a relaxation to the integer programming formulation, which relax the integer constraints $x_i \in \{0,1\}$ to intervals $x_i \in [0,1]$.

Non-integral solutions to LP generally happens

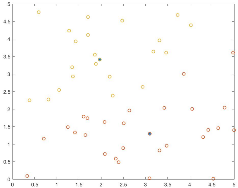
When applying the k-median clustering algorithm to the two-Gaussian-Mixed input data, the output clustering result can be non-integral.
Example from Computational Experiments on G.M.M. input data set:
Optimal solution found : ctr = 4 5 9 23 44
The number of index of centers given should be 3 *under the case* $k = 3$*, however we got give centers after applying the clustering algorithm, in this case the corresponding result vectors are also non-integral.*
Non-integral solution happens.

Outline

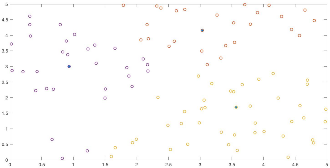
Result of Computational Experiments

1. Two disjoint balls with different radius and equal number of points
2. Two disjoint balls with different radius and equal number of points
3. Two overlapping circles $\in \mathbb{R}^2$
4. Gaussian Mixture Model input
5. Uniformly generated points within a square
6. Two overlapping intervals $\in \mathbb{R}^1$

Uniformly generated points within a square



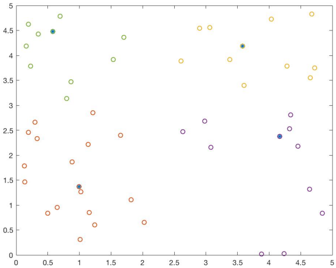
(a) $k = 2$



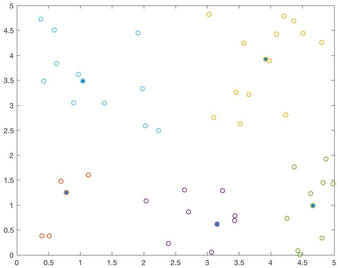
(b) $k = 3$

Figure 3: Uniformly simulated data set within a 5×5 square

Uniformly generated points within a square



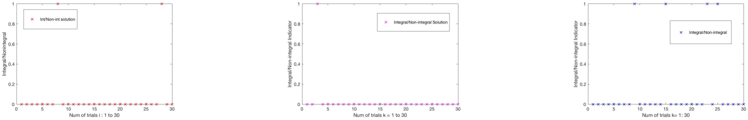
(a) $k = 4$



(b) $k = 5$

Figure 4: Uniformly simulated data set within a 5×5 square

Uniformly generated points within a square



(a) $k = 3$ (b) $k = 4$ (c) $k = 5$

Figure 5: Integrality of cases $k = 3, 4, 5$ for uniformly generated points within a square

Gaussian Mixture Model input

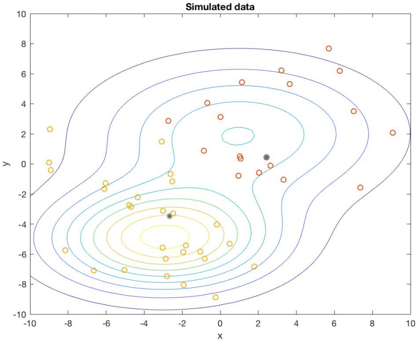
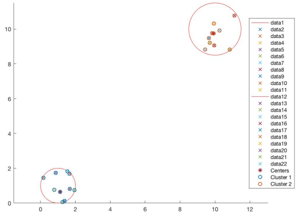
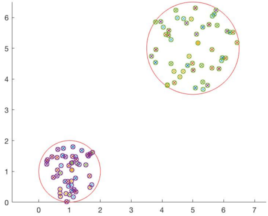


Figure 6: Clustering cases for $k = 2$ GMM

Two disjoint balls with different radius and equal number of points



(a) Trial 1



(b) Trial 2

Figure 7: Two trials with different radius but same number of points

Two disjoint balls with different radius and equal number of points

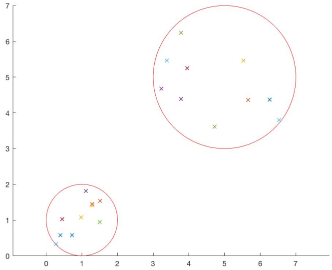


Figure 8: A Clear demonstration 1

Two disjoint balls with different radius and equal number of points

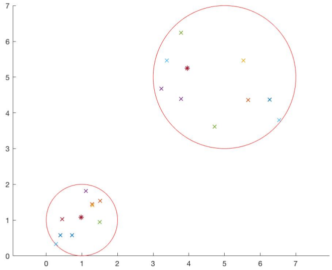


Figure 9: A Clear demonstration 2

Two disjoint balls with different radius and equal number of points

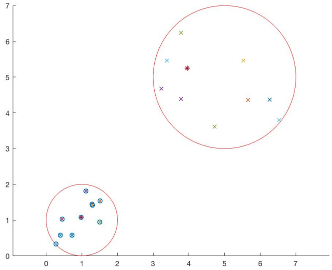


Figure 10: A Clear demonstration 3

Two disjoint balls with different radius and equal number of points

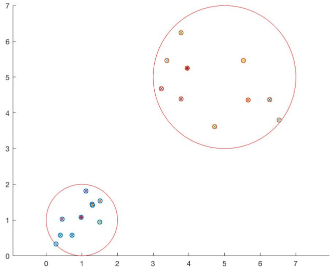
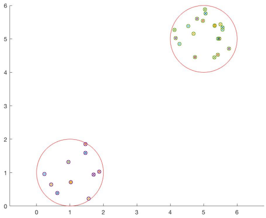
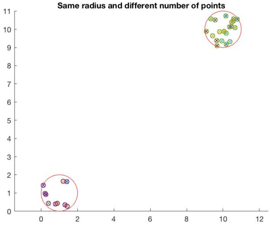


Figure 11: A Clear demonstration 4

Two disjoint balls with same radius and different number of points



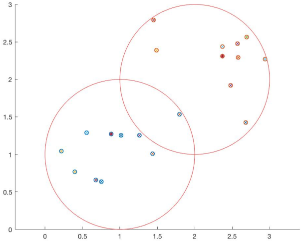
(a) Trial 1



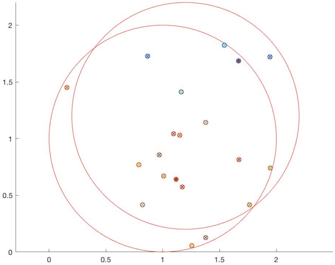
(b) Trial 2

Figure 12: Two trials with same radius and different number of points

Two overlapping circles $\in \mathbb{R}^2$



(a) Trial 1



(b) Trial 2

Figure 13: Two overlapping circles

Two overlapping intervals $\in \mathbb{R}^1$

Surprisingly, over the 1000 trials on k-median clustering for two overlapping intervals, all of the trials get integral solutions.

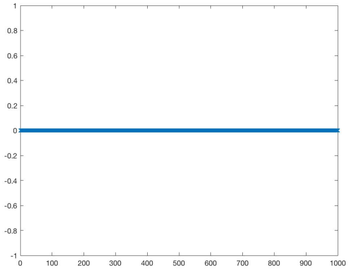


Figure 14: 1000 trials for two overlapped intervals

Outline

Theoretical Proofs

- Conjecture 1 (General k)
- Theorem 2 ($k = 1$)
- Conjecture 3

Conjecture for the General Case k

Given n data points $P = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^1$ and clustering integer k, the k-median LP relaxation always has an integral optimal solution.

Theorem for $k = 1$ case

Given n data points $P = \{x_1, x_2, ..., x_n\} \in \mathbb{R}^1$, the k-median LP relaxation always has an integral solution for $k = 1$ case.

Remark

The data points may not have to be uniformly distributed, it can be generated by any distribution.

Theorem for $k = 1$ case

We prove by contradiction.
Suppose there doesn't exist an optimal fractional solution for the $k-$median objective function.($k = 1$)

$$\min_{z \in \mathbb{R}^{n \times n}} \sum_{p,q} d(p,q) z_{pq} \tag{1}$$

s.t.

$$\sum_{p \in P} z_{pq} = 1, \forall q \in P \tag{2}$$

$$z_{pq} \leq y_p, \forall p, q \in P \tag{3}$$

$$\sum_{p \in P} y_p = 1 \tag{4}$$

$$z_{pq}, y_p \in [0,1], \forall p, q \in P \tag{5}$$

Theorem for $k = 1$ case

Let y^* and z^* be this optimal solution, and let d_{pq} be the objective coefficient vector.
Since the solution is fractional, instead of having either 1 or 0 in the entries of y_p, there exists a series of nonzero entries in y_p vector.

$$y^* = [y_1^*, y_2^*, y_3, ..., y_n]$$

By equation (4), we have

$$\sum_i y_i = 1$$

By equation (3), for z_{pq}, we have the constraints as

$$z_{pq} \leq y_p, \forall p \in \{1, 2, 3, ...n\}$$

Theorem for $k = 1$ case

By equation (2)

$$z_{1q} + z_{2q} + ... + z_{iq} + ... + z_{jq} + ..z_{nq} = 1, \forall q$$

Fix q,if there exists one z_{iq} is strictly small than y_i, then in order to make the total sum still equals 1, there would be some z_{jq} greater than y_j, which violates the constraints. Therefore, the only possibility, is

$$z_{iq} = y_i, \forall i$$

Theorem for $k = 1$ case

Based on this, then we consider the objective function,

$$d_{11}z_{11} + d_{12}z_{12} + ... + d_{1n}z_{1n}$$

$$...$$

$$+d_{i1}z_{i1} + d_{i2}z_{i2} + ... + d_{in}z_{in}$$

$$...$$

$$+d_{j1}z_{j1} + d_{j2}z_{j2} + ... + d_{jn}z_{jn}$$

$$...$$

which is equal to

$$\alpha_1 \sum_{k=1}^{n} d_{1k} + ... + \alpha_n \sum_{k=1}^{n} d_{nk}$$

Now, our goal is to minimize the above objective function,

Theorem for $k = 1$ case

The rest of the proofs is based on the property of linear combination, and get to the contradiction.
-We know that the sum of all αs is equal to 1.
By the property of *linear combination*,
$x \leq \alpha x + (1 - \alpha)y \leq y, \alpha \in [0, 1]$
if we apply this on more terms, we get the result that in order to minimize the objective function, pick the y_i whose corresponding sum is the smallest, and set that y_i to be 1, the rest of them are just assigned to 0, which is an integral vector, contradict with the assumption that y_i is a fractional solution.
Contradiction!
End proof.

Extend to $k = 2$ case

Given n data points uniformly generated from two overlapped intervals $\in \mathbb{R}^1$,the k-median LP relaxation always has an integral solution for $k = 2$ case.

1. Extend the proof scheme for $k = 1$ and the reason for its failure
2. Possible Proof by defining understanding of dual variable of the LP

Idea of Dual Approach

Recall that the Primal LP and the Dual LP of our interest:
Primal LP:

$$\min_{z \in \mathbb{R}^{n \times n}} \sum_{p,q \in P} d(p,q) z_{pq}$$

$$s.t. \sum_{p \in P} z_{pq} = 1, \forall q \in P$$

$$z_{pq} \leq y_p, \forall p, q \in P$$

$$\sum_{p \in P} y_p = k$$

$$z_{pq}, y_p \in [0, 1]$$

Idea of Dual Approach

Dual LP:

$$\max_{\alpha \in \mathbb{R}^n} \sum_{q \in P} \alpha_q - k\xi$$

$$s.t. \alpha_q \leq \beta_{pq} + d(p, q), \forall p, q \in P$$

$$\sum_{q} \beta_{pq} \leq \xi, \forall p \in P$$

$$\beta_{pq} \geq 0, \forall p, q \in P$$

Idea of Dual Approach

Recall the Fact: (Weak Duality)
For any feasible Primal/Dual solution, the primal objective is greater or equal than the dual objective,

$$Primal\ Obj \geq Dual\ Obj$$

(Here in our study case, the primal feasible solution variable is y, z and the dual feasible solution variable is α, β, ξ.)

Idea of Dual Approach

Observe:
If we can find feasible dual solutions s.t.

$$\text{Primal Obj.} = \text{Dual Obj.}(*)$$

then the primal solution y, z is optimal.

Strategy to prove optimality of primal solution: - To produce the dual solution that satisfies (*).

Idea of Dual Approach

- In the approach given by the prior work [Awasthi et al., 2015], the dual variable α represents the **distance thresholds**.
- Intuitively, a point in the set A_j can only "see" other points within a distance α_j.
- When the input data has clear clustering structure.(i.e. the input data set satisfies the condition of *separation*,*center dominance* defined in [Awasthi et al., 2015]), the *dual certificate* α can be easily found as a distance slightly larger than the cluster centers' distance.

Directly applying to the input data has no clustering feature will lead to failure because with different input data feature, the separation can vary. The value of α_q for each cluster may vary.

Prior work

Lemma

Consider sets $A_1, ..., A_k$ with $n_1, ..., n_k$ points respectively. If $\exists \alpha_1, ..., \alpha_k$ s.t. for each $s \in A_1 \cup ... \cup A_k$,

$$\frac{1}{k}([\sum_{i=1}^{k}[n_i\alpha_i - \min_{p\in A_i}\sum_{q\in A_i} d(p,q)]) \geq \sum_{q\in A_1}(\alpha_1 - d(s,q))_+$$

$$+ ... + \sum_{q\in A_k}(\alpha_k - d(s,q))_+$$

then the k-median LP is integral and the partition in clusters $A_1, ..., A_k$ is optimal.

Prior Work

Lemma

In order for the above inequality to hold, the dual variable α has the following property:

- *Each center sees exactly its own cluster i.e.* $(\alpha_j - d(c_j, q))_+ > 0$ *if and only if* $q \in A_j$
- *RHS attains its maximum in the centers* $c_1, c_2, ...c_k$.
- *Each of the terms* $n_i\alpha_i - \min_{p\in A_i} \sum_{q\in A_i} d(p,q)$ *in the average in the LHS are the same*

Idea of Dual Approach

- Dual Certificate we've tested:
  - set the dual certificate variable α - equal in each "cluster-like group"
  - set the dual certificate variable α - not equal in each "cluster-like group"
  - all α value experimented is less than 1, i.e. $\alpha < 1$ and $d < 2$. (Based on the experiment of disjoint uni-circles, constraint of the value of dual certificate is $\alpha > 1$, the distance between cluster centers is $d > 2$).
- Test results for 1D

Outline

Conclusion & Future Work

- Geometric meaning of dual certificate α
- $\mathbb{R}^1$, $k > 2$
- Overlapping circles on $\mathbb{R}^d$

Outline

References I

Awasthi, P., Bandeira, A. S., Charikar, M., Krishnaswamy, R., Villar, S., and Ward, R. (2015).
Relax, no need to round: integrality of clustering formulations.
In *ITCS'15—Proceedings of the 6th Innovations in Theoretical Computer Science*, pages 191–200. ACM, New York.

Thanks for your time!