

AAIL100: Introduction to Data Science

Ragesh Jaiswal, IITD

The k -means Problem

The k -means Problem

k -means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2.$$

The k -means Problem

k -means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2.$$

- How hard is the 1-means problem?
 - Given $x_1, \dots, x_n \in \mathbb{R}^d$ find a point $z \in \mathbb{R}^d$ such that $f(z) = \sum_i \|x_i - z\|^2$ is minimized.

The k -means Problem

k -means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2.$$

- How hard is the 1-means problem?
 - Given $x_1, \dots, x_n \in \mathbb{R}^d$ find a point $z \in \mathbb{R}^d$ such that $f(z) = \sum_i \|x_i - z\|^2$ is minimized.
 - What is $\frac{\partial f(z)}{\partial z_i}$?
 - So, for what z , $\sum_i \|x_i - z\|^2$ gets minimized?

The k -means Problem

k -means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2.$$

- How hard is the 1-means problem?
 - Given $x_1, \dots, x_n \in \mathbb{R}^d$ find a point $z \in \mathbb{R}^d$ such that $f(z) = \sum_i \|x_i - z\|^2$ is minimized.
 - What is $\frac{\partial f(z)}{\partial z_i}$? $\frac{\partial f(z)}{\partial z_i} = nz_i - \sum_j x_{ji}$
 - So, for what z , $\sum_i \|x_i - z\|^2$ gets minimized? $z = \frac{\sum_i x_i}{n}$
 - $\frac{\sum_i x_i}{n}$ is called the *centroid* of the points $x_1, \dots, x_n$.

The k -means Problem

k -means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2.$$

- How hard is the k -means problem for $k > 1$ when $d = 1$? In other words, how hard is the 1-dimensional k -means problem?

The k -means Problem

k -means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2.$$

- How hard is the k -means problem for $k > 1$ when $d = 1$? In other words, how hard is the 1-dimensional k -means problem?
 - There is a simpler Dynamic Programming algorithms for this problem!

The k -means Problem

k -means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2.$$

- How hard is the k -means problem for $k > 1$ and $d > 1$?
 - NP-hard.

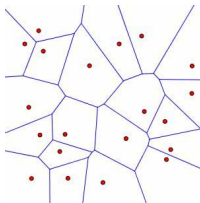
The k -means Problem

k -means

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\text{cost}(T) = \sum_{x \in S} \min_{z \in T} \|x - z\|^2.$$

- How is this problem related to clustering?
 - The k centers induce a *voronoi* partition of $\mathbb{R}^d$ (and hence the given data points).
 - The voronoi partition corresponding to a center z is the region of space whose nearest center (among the k centers) is z .



The k -means Problem

The k -means Algorithm

- The most popular heuristic that used to solve the k -means problem in practice is the k -means algorithm (also known as Lloyd's Algorithm).

k -means Algorithm

- Initialize centers $z_1, \dots, z_k \in \mathbb{R}^d$.
- Repeat until there is no further change in cost:
 - For each j : $C_j \leftarrow \{x \in S \mid z_j \text{ is the closest center of } x\}$.
 - For each j : $z_j \leftarrow \text{Centroid}(C_j)$.

The k -means Problem

The k -means Algorithm

k -means Algorithm

- Initialize centers $z_1, \dots, z_k \in \mathbb{R}^d$.
- Repeat until there is no further change in cost:
 - For each j : $C_j \leftarrow \{x \in S \mid z_j \text{ is the closest center of } x\}$.
 - For each j : $z_j \leftarrow \text{Centroid}(C_j)$.

- Claim 1: For any dataset S , let $\bar{C}_i$ denote the set of centers after the i^{th} iteration of the loop. Then for all i ,
 $\Phi(S, \bar{C}_{i+1}) \leq \Phi(S, \bar{C}_i)$.

Lemma

For any set $S \subset \mathbb{R}^d$ and any $z \in \mathbb{R}^d$,

$$\Phi(S, z) = \text{cost}(S, \text{Centroid}(S)) + |S| \cdot \|z - \text{Centroid}(S)\|^2.$$

The k -means Problem

The k -means Algorithm

k -means Algorithm

- Initialize centers $z_1, \dots, z_k \in \mathbb{R}^d$.
 - Repeat until there is no further change in cost:
 - For each j : $C_j \leftarrow \{x \in S \mid z_j \text{ is the closest center of } x\}$.
 - For each j : $z_j \leftarrow \text{Centroid}(C_j)$.
-
- Claim 1: For any dataset S , let $\bar{C}_i$ denote the set of centers after the i^{th} iteration of the loop. Then for all i , $\Phi(S, \bar{C}_{i+1}) \leq \Phi(S, \bar{C}_i)$.
 - Claim 2: There exists datasets on which the k -means algorithm gives arbitrarily bad solutions.

The k -means Problem

The k -means Algorithm

k -means Algorithm

- Initialize centers $z_1, \dots, z_k \in \mathbb{R}^d$.
 - Repeat until there is no further change in cost:
 - For each j : $C_j \leftarrow \{x \in S \mid z_j \text{ is the closest center of } x\}$.
 - For each j : $z_j \leftarrow \text{Centroid}(C_j)$.
-
- Claim 1: For any dataset S , let $\bar{C}_i$ denote the set of centers after the i^{th} iteration of the loop. Then for all i , $\Phi(S, \bar{C}_{i+1}) \leq \Phi(S, \bar{C}_i)$.
 - Claim 2: There exists datasets on which the k -means algorithm gives arbitrarily bad solutions.
 - Claim 3: There exists datasets on which the k -means algorithm takes a very long time to output an answer.

The k -median Problem

k -median

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\Phi(S, T) = \sum_{x \in S} \min_{z \in T} \|x - z\|.$$



Figure: What is the solution for the 2-median problem for the above 2-D dataset?

The k -median Problem

k -median

Given a set of points $S \subset \mathbb{R}^d$ in a d dimensional Euclidean space, and an integer k , output a set $T \subset \mathbb{R}^d$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\Phi(S, T) = \sum_{x \in S} \min_{z \in T} \|x - z\|.$$

- How hard is the k -median problem?
 - NP-hard.

The k -center Problem

k -center on a Metric Space

Let (X, D) denote a metric space. Given a set of points $S \subset X$, and an integer k , output a set $T \subset X$ of points (called *centers*) such that $|T| = k$ and the following cost function is minimised:

$$\Phi(S, T) = \max_{x \in S} \min_{z \in T} D(x, z).$$



Figure: What is the solution for the 2-center problem for the above 2-D dataset?

End