__/__/____

CS - 328 : Intro. to Data Science
Homework - 1

Name - Pushkar Mujumdar
Roll No. - 18110132

1. Given function $d(x,y) = \min_i |x_i - y_i|$

We typically like to have the following metric properties -

(i) $d(x,x) = 0$, $d(x,y) \geq 0$ (where $x \neq y$)
(ii) $d(x,y) = d(y,x)$
(iii) $d(x,y) + d(y,z) \geq d(x,z)$

It is trivial to observe that the property (iii) is not followed by our given ~~distan~~ function

Counter Example →

Suppose

$x = (0, 0)$
$y = (k, 0)$
$z = (k, k)$

$$d(x,y) = \min(|(k-0)|, |0-0|)$$
$$= \min(|k|, 0)$$
$$= 0$$

$$d(y,z) = \min(|k-k|, |0-k|)$$
$$= 0$$

$$d(x,z) = \min(|0-k|, |k-0|)$$
$$= |k|$$

Here, $d(x,y) + d(y,z) < d(x,z)$

$\therefore$ $d(x,y) = \min_i |x_i - y_i|$ is NOT a metric

2. Given cost function →

$$\text{cost}(C_i) = \sum_i \frac{1}{|C_i|} \sum_{x, y \in C_i} \|x - y\|_2^2$$

Let us try to simplify this function

Suppose we fix 'y' in the cost,

$$\text{expr} = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - y\|_2^2 \quad \text{for some } y \in C_i \qquad (1)$$

Let c_i° be the centroid of cluster C_i

$$\therefore \text{expr} = \frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i^\circ + c_i^\circ - y\|_2^2$$

(add & subtract vector c_i°)

$$= \frac{1}{|C_i|} \left[\sum_{x \in C_i} \|x - c_i^\circ\|_2^2 + \sum_{x \in C_i} \|c_i^\circ - y\|_2^2 + \cancel{2(c_i^\circ - y) \sum_{x \in C_i} (x - c_i^\circ)} \right] \quad \to 0$$

$\left(\because c_i^\circ \text{ is centroid of } C_i, \; \sum_{x \in C_i} (x - c_i^\circ) = 0 \right)$

$$= \frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i^\circ\|_2^2 + \frac{\cancel{|C_i|}}{\cancel{|C_i|}} \|c_i^\circ - y\|_2^2$$

$$\therefore \text{expr} = \boxed{\frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i^\circ\|_2^2 + \|c_i^\circ - y\|_2^2} \qquad (2)$$

P.T.O. →

_/_/__

Now, if we sum up this expr. over all $y \in C_i$,

$$\sum_{y \in C_i} \text{expr} \; \cancel{\sum_{y\in C_i} \sum_{x \in C_i} \frac{1}{|C_i|}}$$

$$= \sum_{y \in C_i} \frac{1}{|C_i|} \sum_{x \in C_i} \|x-y\|_2^2$$

from 1 & 2, we have

$$= \sum_{y \in C_i} \left(\frac{1}{|C_i|} \sum_{x \in C_i} \|x - c_i^o\|_2^2 + \|y - c_i^o\|_2^2 \right)$$

$$= \cancel{\frac{1}{|C_i|} |C_i|} \sum_{x \in C_i} \|x - c_i^o\|_2^2 + \sum_{y \in C_i} \|y - c_i^o\|_2^2$$

$$= \sum_{x \in C_i} \|x - c_i^o\|_2^2 + \sum_{y \in C_i} \|y - c_i^o\|^2$$

($\because$ we are iterating over all points in C_i for both sums.)

$$= \boxed{2 \sum_{x \in C_i} \|x - c_i^o\|^2}$$

$$\therefore \sum_{y \in C_i} \sum_{x \in C_i} \frac{1}{|C_i|} \|x-y\|_2^2 = 2 \sum_{x \in C_i} \|x - c_i^o\|$$

Now, if we take distinct pairs (x, y),

$$\sum_{x, y \in C_i} \frac{1}{|C_i|} \|x-y\|_2^2 = \frac{1}{\cancel{2}} \times \cancel{2} \sum_{x \in C_i} \|x - c_i^o\|$$

$\therefore$ Our given cost function

$$\text{cost}(C) = \sum_i \frac{1}{|C_i|} \sum_{x,y \in C_i} \|x-y\|_2^2$$

simplifies to

$$\cancel{\text{cost}(C) = \sum_i \|x - c_i^o\|}$$

$$\text{cost}(C) = \sum_i \sum_{x \in C_i} \|x - c_i^o\|$$

where c_i^o is the centroid of cluster C_i

As we can see, the cluster costs are same as the cost function we defined for K-means clustering.

$\therefore$ The algorithm for this objective is Lloyd's algorithm.

Iterate →

(i) Find current centers of the partitions
(ii) Assign points to the nearest centers
(iii) Recalculate the centers of clusters.

~~3. Suppose~~

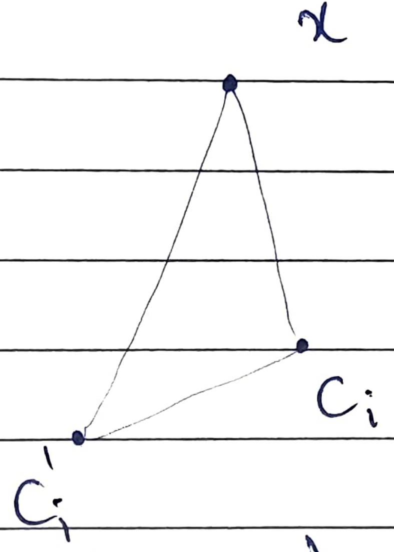
3. Given → The cost function for clustering uses L1 norm (doesn't take squares of distances while summing up).

Suppose the optimal center when we allow arbitrary points to be centers be C_i and the optimal center when we require centers to be data points be C'_i for the i^{th} cluster P_i.

We will choose C'_i such that from all the points in cluster P_i, C'_i is the closest to C_i

ie. $d(C_i' - C_i) < d(x - C_i)$ —— (1)

$\forall$ points $x \in P_i$ and where 'd' is the L1 norm.



(where 'd' is L1 norm)

Using Δ inequality, we have

$$d(c_i' - x) \leq d(c_i' - c_i) + d(c_i - x)$$

from 1, we can say

$$d(c_i' - x) \leq 2d(c_i - x) \tag{2}$$

If we take summation over all $x \in$ cluster P_i, we have

$$\sum_{x \in P_i} d(C_i' - x) \leq 2 \sum d(C_i - x).$$

$\therefore$ (Cost when datapoint is center) $\leq$ 2 (Cost when allow arbitrary center)

Hence Proved 2-ratio.

- Based on the above observation, we can propose an algorithm with initialization similar to Lloyd's algorithm and after recalculating the centers, choose the datapoint closest to the calculated ~~cluster~~ center as the cluster center.

- However, we need $O(|P_i|)$ steps to find the closest point to the ~~center~~ calculated center according to the distance metric.

- Instead, we can do an exhaustive search to find the data-point which minimizes cluster cost for the distance metric in $O(|P_i|^2)$ steps by pre-calculating pairwise distances. This will work for both L1 norm and Euclidean distance costs.

- The clustering cost will stop decreasing after a certain point. So, we will be running the above proposed algorithm ~~till~~ & only while the clustering cost decreases.

Algorithm →

1. Make k partitions
2. While clustering cost decreases
   (i) In each cluster, exhaustive search to find the datapoint which minimizes the sum of distances to make it the cluster center
   (ii) Reassign each point to the closest cluster center

Question 1, 2 and 3 were discussed with –

- Harsh Patel (18110062)
- Shivam Sahni (18110159)

Question 4 and 5 Github link

https://github.com/pmujumdar27/CS328_assignment1

Question 5 youtube video link

https://youtu.be/JV9Mu24ZzV0