ASSIGNMENT - 1

Discussed with :

Pushkar Mujumdar (18110132)
Harsh Patel (18110062)

1. We can show that $d(x, y) = \min |x_i - y_i|$ does not satisfy the metric property:

$$d(x, y) + d(y, z) = d(z, x)$$

Suppose $x = (0, 0)$
$y = (10, 0)$
$z = (10, 10)$

then, $d(x, y) = 0$
$d(y, z) = 0$
$d(x, z) = 10$

Hence, for these values of x, y, z :
$d(x, y) + d(y, z) < d(x, z)$

Hence, $d(x, y) = \min |x_i - y_i|$ is **NOT** a metric

2.

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

We show, cost ($\mathbb{C}$) = K Means Cost

Let $\vec{k}$ be the mean of all points in the cluster,

$$\Rightarrow \text{Cost}(\mathbb{C}) = \sum_i \frac{1}{|C_i|} \sum_{x,y \in C_i} \|x-k+k-y\|^2$$

$$= \sum_i \frac{1}{2|C_i|} \sum_x \sum_y \|(x-k)-(y-k)\|^2$$

$$\text{Cost}(\mathbb{C}) = \sum_i \frac{1}{2|C_i|} \sum_x \sum_y \left(\|x-k\|^2 + \|y-k\|^2 - 2(x-k)^T(y-k)\right)$$

Solving each summation separately :

$$\rightarrow \sum_x \sum_y \|x-k\|^2 = \sum_x |C_i| \|x-k\|^2$$

$$\rightarrow \sum_x \sum_y \|y-k\|^2 = \sum_y \sum_x \|y-k\|^2 = \sum_y |C_i| \|y-k\|^2$$

$$\rightarrow \sum_x \sum_y 2(x-k)^T(y-k)$$

$$= 2 \sum_x \sum_y x^Ty - x^Tk - k^Ty + k^Tk$$

$$= 2\left(|C_i|^2 k^Tk - |C_i|^2 k^Tk - |C_i|^2 k^Tk + |C_i|^2 k^Tk\right)$$

$$= 0$$

Hence,

$$\text{Cost}(\mathbb{C}) = \sum_i \frac{1}{2|C_i|} \left(\sum_x |C_i| \|x-k\|^2 + \sum_y |C_i| \|y-k\|^2 \right)$$

$$\text{Cost}(\mathbb{C}) = \sum_i \frac{1}{2|C_i|} \left(2 \sum_x |C_i| \|x-k\|^2 \right)$$

(As $\sum_x |C_i| \|x-k\|^2 = \sum_y |C_i| \|y-k\|^2$)

$$\boxed{\text{Cost}(\mathbb{C}) = \sum_i \sum_x \|x-k\|^2}$$

Hence, Cost ($\mathbb{C}$) = Cost of K Means

Algorithm for optimizing Cost ($\mathbb{C}$)

→ We can use any of the K Means objective optimize algorithms such as Lloyd or KMeans ++.

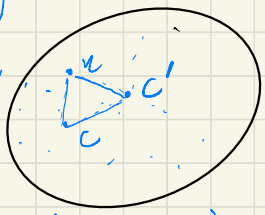
3. Let c be the optimal centre when we allow any arbitrary point to be the centre

Let c' be the optimal centre when we allow only the data points to be the centre.

Let K be the partition created by c centre.

Now,

Partition K



$\forall\ x \in K$

$$d(c', x) \leq d(c, x) + d(c, c') \quad -(1)$$

(as L1 norm follows triangle inequality)

We can choose c' to be the data point in K, st

$$d(c, c') \leq d(c, x) \quad -(2)$$

Using (1) and (2)

$$d(c', x) \leq d(c, x) + d(c, c')$$

$$\leq d(c, x) + d(c, x)$$

$$\boxed{d(c', x) \leq 2\, d(c, x)} \quad \forall\ x \in K$$

Cost of partition K using C' as centre =

$$\sum_{x \in K} d(c', x) \leq 2 \sum_{x \in K} d(c, x)$$

$$= 2 \text{ Optimal cost of clustering}$$

→ Similarly we can do this for all cluster partitions

Hence, we proved that there is two ratio between optimal value when we requir all cluster centres to be data points or allow arbitrary points to be centres.

Lloyd's algorithm for Euclidean k-median

The only change that we see from Lloyd's for K-Means is the criteria for recalculating centers. Here, we take median of the partition for recalculating center.

Lloyd for K-Median

- Choose k points arbitrary from data points
- repeat
  - Assign points to nearest centers
  - recalculate centers by taking median of all points. Then choose the data point closest to the median as center.
- until no (or small #) points change cluster

  or when cluster centers dont shift much.

Let C be number of iteration

Time Complexity : $O(CKNd)$

where K = no. of clusters
N = no. of samples
d = dimension

To show :- Cluster cost decreases with each iteration

→ The first step of an iteration, we recalculate center by putting it as the median.

As we know, median gives the minimum cost when using L1 norm, this step will decrease cost.

→ In the second step, we reassign points to their nearest centers, hence reducing cost of the whole clustering.

Hence, clustering cost always reduces with each iteration.

4. https://github.com/shivam15s/CS-328-HW-1/blob/main/q4.ipynb

5. LINK FOR YOUTUBE VIDEO :

https://youtu.be/KaRMekgtFCY

Challenges :

1. Figuring out the attributes which can lead to good inferences
2. Handling the NaN values.
   Used b fill, f fill for filling