Homework - 1

18110079
Kalyan Reddy . S

1) $d(x,y) = \min_i |x_i - y_i|$ is not a matric,

because it doesn't follow,

$$d(x,y) + d(y,z) \geq d(x,z)$$

Ex:-

Conider,

$x = (2,1,3)$

$y = (4,5,1)$

$z = (8,6,7)$

$\therefore d(x,y) = \min(|4-2|, |5-1|, |1-3|)$

$= 2$

$d(y,z) = \min(|4-8|, |5-6|, |1-7|)$

$= 1$

$d(x,z) = \min(|2-8|, |1-6|, |3-7|)$

$= 4.$

$\therefore d(x,z) > d(x,y) + d(y,z)$

$4 > 2 + 1$

$4 > 3$

Hence, $d(x,y) = \min_i |x_i - y_i|$ is not a matric.

2) The best way to solve this problem is to merge points such that it reduces the,

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

<u>Algorithm</u> :-

* Here each point is a cluster. (cluster $= [x_1, x_3, \ldots x_n]$) $x_i \in C$.
* While (no. of cluster > K)

{

Select all pair from cluster :-

{

Calculate cost (C).

If (Cost is minimum)

{ we combine those pair to form cluster }

}

(If, x_i, x_j are minimum)

cluster $= [x_1, x_2, \ldots [x_i, x_j], \ldots]$

}

---

* <u>Calculating Cost</u> :-

→ for i in cluster :-

{ $Cost = Cost + \frac{1}{|i|} \sum_{x,y \in i} \|x-y\|_2^2$

}

3) + Let us concider a single cluster C, where,

x_{opt} is the optimal cluster centre which is a data point.

+ let, C_{opt} be the optimal cluster centre.

$$\text{Cost of } (C) \text{ wrt } x_{opt} = \sum_{x \in C} |x - x_{opt}|$$

(i) For any point $x \in C$, we have $|x - x_{opt}| \leq |x - C_{opt}| + |x_{opt} - C_{opt}|$

$$① \rightarrow \therefore \boxed{\begin{aligned} \sum_{x \in C} |x - x_{opt}| &\leq \sum_{x \in C} \left(|x - C_{opt}| + |x_{opt} - C_{opt}| \right) \\ &\leq \sum_{x \in C} |x - C_{opt}| + |C| \left(|x_{opt} - C_{opt}| \right) \end{aligned}}$$

We know, $|x_{opt} - C_{opt}|$ is the minimum distance, $\forall x \in C$. since x_{opt} is optimal centre.

$$② \rightarrow \therefore \boxed{|C| \left(|x_{opt} - C_{opt}| \right) \leq \sum_{x \in C} |x - C_{opt}|}$$

Sub, ② in ①.

$$\begin{aligned} \sum_{x \in C} |x - x_{opt}| &\leq \sum_{x \in C} |x - C_{opt}| + \sum_{x \in C} |x - C_{opt}| \\ &\leq 2 \sum_{x \in C} |x - C_{opt}| \end{aligned}$$

$\therefore$ It varies by <u>atmost a factor of 2!</u>

3) $\therefore$ From above we can say that for every cluster we can have optimal centre, whose cost is atmost factor of 2 of real optimal centre.

Lloyd's algorithm variant :-

* Choose K- centers.
* Iterate :-
    - → add points to the cluster centre based on minimum $|x_i - C_K|$, where C_K = centres.
    - → Find median of all clusters and repeat above process with medians as centre.

* The clustering cost is always decreasing, since, we are reducing the cost in every iteration, by adding points to low cost cluster.

Scanned with CamScanner