Homework - 1

Chris Francis
18110041
CS 328 - Data Science
Indian Institute of Technology Gandhinagar

February 10, 2021

1

Answer:
The given function $d(x, y) = \min_i |x_i - y_i|$ is not a metric. The following counter-example shows that it doesn't satisfy the third metric property:

$x = (10, 2)$, $y = (1, 2)$, $z = (1, 20)$

$d(x, y) = 0$, $d(y, z) = 0$, $d(x, z) = 9$

Clearly, $d(x, y) + d(y, z) \geq d(x, z)$ is not satisfied here.

2

Answer:
We can use the same Lloyd's algorithm that we used for k-means clustering in this case too, since the given cost function is the same as the cost function in k-means clustering except for a constant factor of 2.

Cost in k-means, $\text{cost}(\mathbb{C}) = \sum_i \sum_{x \in C_i} ||x - c_i||^2$, where c_i is the cluster center, which is also the mean of the points in the cluster.

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

We can show the following relation between the cluster costs:

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

or,

$$2|C_i| \sum_{x \in C_i} ||x - c_i||^2 = \sum_{x,y \in C_i} ||x - y||^2$$

Proof:

$$
\begin{aligned}
2|C_i| \sum_{x \in C_i} ||x - c_i||^2 &= 2|C_i| \sum_{x \in C_i} \left\| x - \frac{1}{|C_i|} \sum_{y \in C_i} y \right\|^2 \\
&= 2|C_i| \sum_{x \in C_i} \left(\|x\|^2 - 2x^T \left(\frac{1}{|C_i|} \sum_{y \in C_i} y \right) + \left\| \frac{1}{|C_i|} \sum_{y \in C_i} y \right\|^2 \right) \\
&= 2|C_i| \sum_{x \in C_i} \|x\|^2 - 4 \sum_{x \in C_i} \sum_{y \in C_i} x^T y + 2|C_i|^2 \left\| \frac{1}{|C_i|} \sum_{y \in C_i} y \right\|^2 \\
&= 2|C_i| \sum_{x \in C_i} \|x\|^2 - 4 \sum_{x,y \in C_i} x^T y + 2 \left\| \sum_{y \in C_i} y \right\|^2 \\
&= 2|C_i| \sum_{x \in C_i} \|x\|^2 - 4 \sum_{x,y \in C_i} x^T y + 2 \sum_{x \in C_i} \sum_{y \in C_i} x^T y \\
&= 2|C_i| \sum_{x \in C_i} \|x\|^2 - 2 \sum_{x,y \in C_i} x^T y \\
&= |C_i| \sum_{x \in C_i} \|x\|^2 - 2 \sum_{x,y \in C_i} x^T y + |C_i| \sum_{x \in C_i} \|x\|^2 \\
&= \sum_{y \in C_i} 1 \sum_{x \in C_i} \|x\|^2 - 2 \sum_{x,y \in C_i} x^T y + \sum_{x \in C_i} 1 \sum_{y \in C_i} \|y\|^2 \\
&= \sum_{y \in C_i} \sum_{x \in C_i} \|x\|^2 - 2 \sum_{x,y \in C_i} x^T y + \sum_{x \in C_i} \sum_{y \in C_i} \|y\|^2 \\
&= \sum_{x,y \in C_i} \|x\|^2 - 2 \sum_{x,y \in C_i} x^T y + \sum_{x,y \in C_i} \|y\|^2 \\
&= \sum_{x,y \in C_i} \left(\|x\|^2 - 2x^T y + \|y\|^2 \right) \\
&= \sum_{x,y \in C_i} \|x - y\|^2 ,
\end{aligned}
$$

$\implies$ the given cost function is just 2 times the cost function in k-means clustering.

$\implies$ we can use Lloyd's algorithm for this case too.

Algorithm: (Lloyd's algorithm)

Initialize centers randomly

While stopping criteria not attained, iterate {
    Assign points to nearest centers
    Recalculate centers
}

Stopping criteria:

1. when no (or small number) points change cluster
2. when cluster centers don't shift much

3

Answer:
Since the k-medians problem is defined in a similar way to the k-means problem, except that we do not take the squares of the distances when summing up, the cost for k-medians is as follows:

Cost for k-medians clustering, $\text{cost}(\mathbb{C}) = \sum_i \sum_{x \in C_i} ||x - c_i||$, where c_i is the cluster center, which is also the median of the points in the cluster. The median is computed in each single dimension. For example, the median of the following points: $(1, 2, 3), (4, 5, 6), (10, 4, 2)$ would be $(4, 4, 3)$ since 4 is median in the first dimension, 4 is the median in the second dimension and 3 is the median in the third dimension. Clearly, the cluster center computed in this manner may not lie in the dataset.

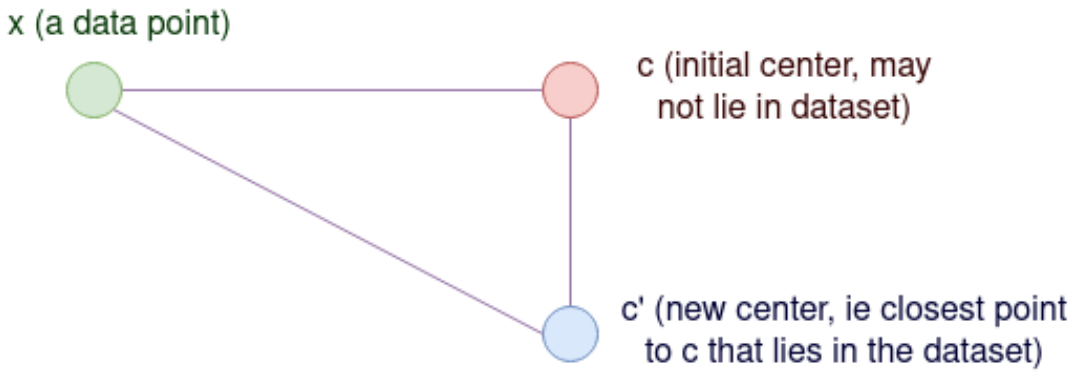
We can show that there is at most a factor of two ratio between the optimal value when we either require all cluster centers to be data points or allow arbitrary points to be centers,

which can be written as:

cost when all cluster centers are required to be data points $\leq 2 \times$ cost when arbitrary points are allowed to be cluster centers

When we require all cluster centers to be data points, we first compute centers similar to the other case(arbitrary points allowed case), and then replace them with the points closest to them in the data set.

Proof:



x is any arbitrary data point. c is a cluster center computed by finding the median in each single dimension as mentioned earlier. c' is the new cluster center that will replace c' when you want all centers to be points in the dataset, which means c' is the closest point to c that is in the dataset.

we know that

$||x - c'|| \leq ||x - c|| + ||c - c'||$

since c' is the closest point to c that is in the dataset, $||c - c'|| \leq ||c - x||$

$\implies ||x - c'|| \leq ||x - c|| + ||c - x||$

$\implies ||x - c'|| \leq 2||x - c||$

$\implies \sum_i \sum_{x \in C_i} ||x - c_i'|| \leq 2 \sum_i \sum_{x \in C_i} ||x - c_i||$

ie, cost when all cluster centers are required to be data points $\leq 2 \times$ cost when arbitrary points are allowed to be cluster centers

Variant of the Lloyd's algorithm for Euclidean k-median

Initialize centers randomly

While stopping criteria not attained, iterate {
    Assign points to nearest centers
    Recalculate centers(centers are found by calculating the median of each cluster in each single dimension as mentioned earlier)
    if centers have to be data points {
        replace each center with the point closest to it in the dataset

```
    }
}
```

Stopping criteria:

1. when no (or small number) points change cluster
2. when cluster centers don't shift much

Since the algorithm follows directly from Lloyd's algorithm we can say that the clustering cost is always non-increasing (either decreasing or no change).

4

Link to Google Colab notebook:
`https://colab.research.google.com/drive/1OyOIvDSGYnc48Z9CqM-mQnuUJKM_cLPn?usp=sharing`

5

Link to Google Colab notebook:
`https://colab.research.google.com/drive/18MDdznIH4U0rWwCofdTtJHawPWtvHnP6?usp=sharing`

Link to recorded video: `https://youtu.be/d9rjh6-Qhfs`

Note

The Jupyter Notebooks are also present in the following GitHub repo:
`https://github.com/frank-chris/CS-328-Assignments`