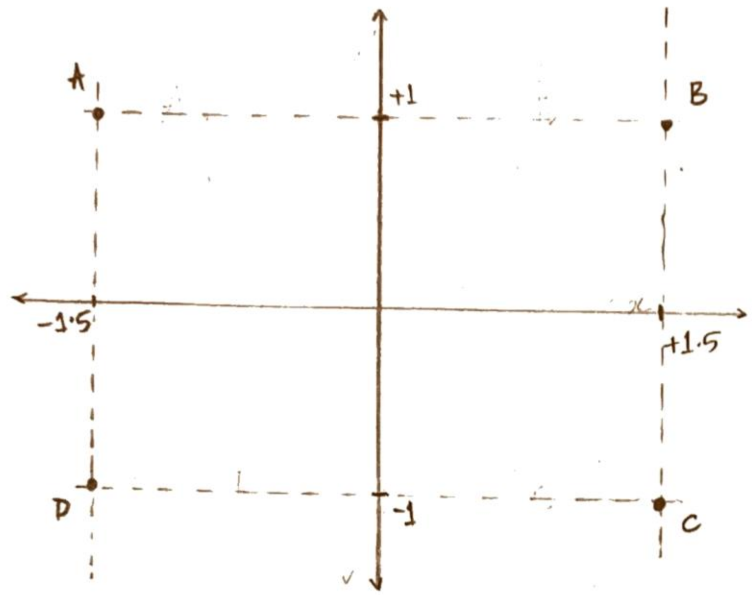
Problem 4

Given:

$$\min_{c_1,\dots c_k} \sum_{i=1}^{n} \|x_i - c(x_i)\|_2^2$$

To Prove: K-means algorithm does not always find the optimal solution to the above objective




Step 0: centroid arrangements

① A & D ② B & C

Step 1: compute objective

$= 2 \cdot \|1-0\|_2^2 = 2$

Step 2: Optimal objective $= 2+2=4 = \text{OPT}$

Now consider a bad initialization:

Step 0: centroid arrangements

① A & B ② C & D

Step 1: compute objective

$= 2 \cdot \|1.5\|_2^2 = 4.5$

In this case, the next step of the algorithm does not change any point assignments because A & B are closer to $(0,1)$ than $(0,-1)$

Then, when we recalculate the cluster centroids, that doesn't change either because $(0,1)$ is the center of A & B, same for C & D. Hence, the algorithm converges

Now, we wish to generalize this example to

p dimensions

n data points

k clusters

4 types of points :

$(-t-0.5, t)$ & 0 for all other dimensions

$(t+0.5, t)$

$(-t-0.5, -t)$

$(t+0.5, -t)$

$$\text{Points} = \left\{ \begin{array}{l} (-t-0.5, t) \\ (t+0.5, t) \\ (-t-0.5, -t) \\ (t+0.5, -t) \end{array} : t \in [1, T] \right\}$$

This could work for any $t \geqslant 0$ ; 4T points ; 2T clusters

The optimal objective function cost =

All the optimal clusters are the midpoints of the pair of points that have the same y-value & are symmetrically reflected across the y-axis; i.e.

- $(-t-0.5, t) <> (-t-0.5, -t)$ ; centroid location is $(-t, -0.5, 0)$
- $(t+0.5, -t) <> (t+0.5, t)$ ; centroid location is $(t+0.5, 0)$

Thus, optimal value of objective would be :

$$\sum_{t=1}^{T} 2 \cdot t^2 = 2 \sum_{t=1}^{T} t^2$$

In the non-optimal case, centroids would be located:

- $(-t-0.5, t) <> (t+0.5, t)$; center is $(0, t)$
- $(-t-0.5, -t) <> (t+0.5, -t)$; center is $(0, -t)$

The value of the objective function in this case

$$\sum_{t=1}^{T} 2\cdot(t+0.5)^2 = 2\sum_{t=1}^{T} t^2 + 2\sum_{t=1}^{T}(0.25+t)$$

$$= 2\sum_{t=1}^{T} t^2 + \sum_{t=1}^{T}(t+0.5)$$

$$> 2\sum_{t=1}^{T} t^2 \quad \because\ t>0\ \&\ T\geq 1$$

Thus, since we have proved suboptimal objective is greater in this setup, we have proved k-means is not ALWAYS optimal.