Problem 1 – K-means clustering

2. µ = [ [2, 3], [2, -1] ]

Iteration 1

Dtrain = { [1,0], [1,2], [3,0], [2,2] }

µ1 = [2,3] and µ2 = [2,−1]

distances = [ [10,2], [2, 10], [10, 2], [1,9] ]

z = [ 1, 0, 1, 0 ]

new\_ µ = [ [1.5, 2], [2, 0] ]

iteration 2

Dtrain = { [1,0], [1,2], [3,0], [2,2] }

µ1 = [1.5, 2] and µ2 = [2, 0]

distances = [ [4.25, 1], [.25, 5], [6.25, 1], [.25, 4] ]

z = [ 1, 0, 1, 0 ]

new\_ µ = [ [1.5, 2], [2, 0] ]

converged -> z = [1, 0, 1, 0], µ = [[1.5, 2], [2,0]]

1. µ = [ [0,1], [3,2] ]

Iteration 1

Dtrain = { [1,0], [1,2], [3,0], [2,2] }

µ1 = [0,1] and µ2 = [3,2]

distances = [ [2,8], [2,4], [10,4], [5,1] ]

z = [ 0, 0, 1, 1 ]

new\_ µ = [ [1, 1], [2.5, 1] ]

iteration 2

Dtrain = { [1,0], [1,2], [3,0], [2,2] }

µ1 = [1,1] and µ2 = [2.5, 1]

distances = [ [1,3.25], [1,3.25], [5,1.25], [2,1.25] ]

z = [ 0, 0, 1, 1 ]

new\_ µ = [ [1, 1], [2.5, 1] ]

converged -> z = [0, 0, 1, 1], µ = [ [1, 1], [2.5, 1] ]

Final results are different since we used two different initial centroids.

1. hw5\_submission.py

I believe alternating minimization on this function would look like this:

1. Running k-means multiple times on the same dataset with the same K, but different random initializations allows us to find the most optimal set of initial conditions. K-means is very dependent on the initial centroid positions you use, and to find the best possible initial centroid positions, we need to run k-means multiple times, and select the initial centroid positions that give us the best data.
2. If we scale all dimensions in our initial centroids and data points by some factor, we would retrieve the same clusters after running k-means. This is because finding a cluster and calculating a centroid are all linear processes that operate on vectors. And because k-means is a distance-based algorithm changing all the distances by the same amount would generate the same clusters, but we might get a higher amount of loss. If we were to scale only certain dimensions, we would receive the same clusters as well, but the exact datapoints where the centroids are located would change.