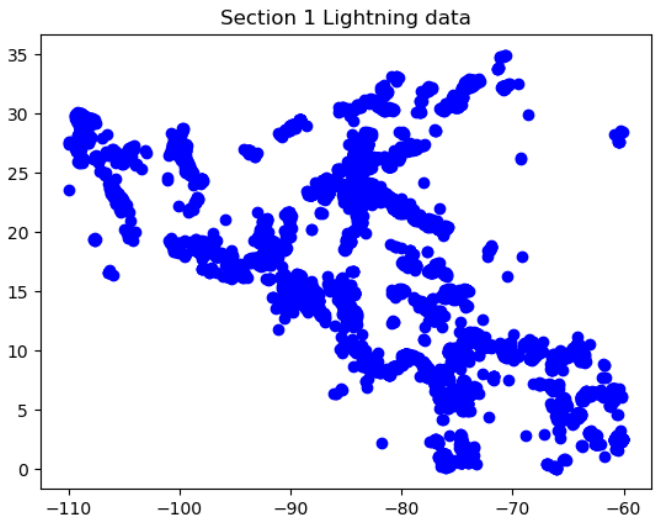
Assignment 4 – KMeans

Seah Zu Xiang

1. Lightning Data



2. Objective function

$$E = \frac{1}{n}\sum_{i=1}^{K}\sum_{O\in C_i} d(O, medoid_i)$$

$$where\ n = number\ of\ samples, K = number\ of\ clusters, C_i = cluster\ of\ i$$

This objective function is used to find the variance of the data for each cluster, to see if the datas are near the medoid. In general, we want to minimize the result of the function to find the optimal number of clusters because that would mean the datas are properly separated and clustered to clusters where other datas are also similar to it.

3. Assigning objects

$$\forall O \in Samples, \exists O \in Cluster_i |\ distance(O, Medoid_i) == argmin$$

An object belongs to the cluster which has the lowest distance between the object and its respective medoid.

Find the medoid closest to the object and assign it to that cluster. In this case, we are using Euclidean distance since the object has 2 features.

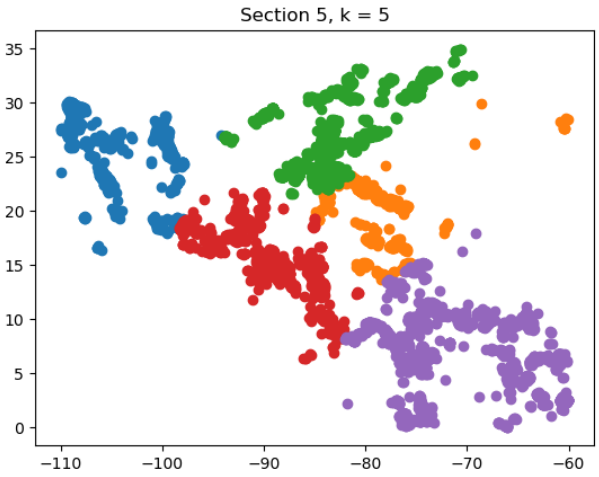
4. Updating Means

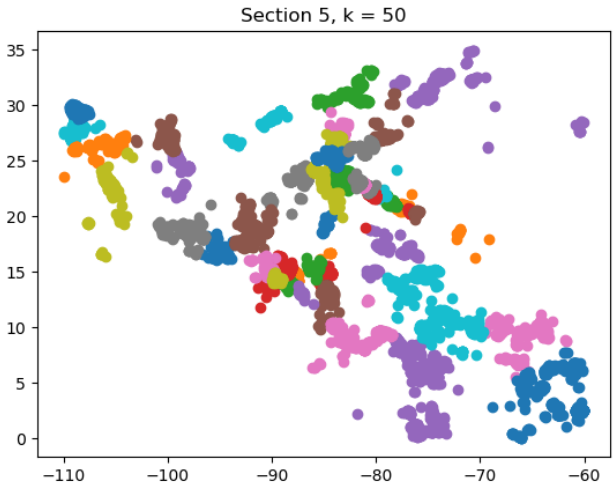
$$Medoid_i = \frac{1}{N}\sum_{n=1}^{N} O_n \in C_i$$

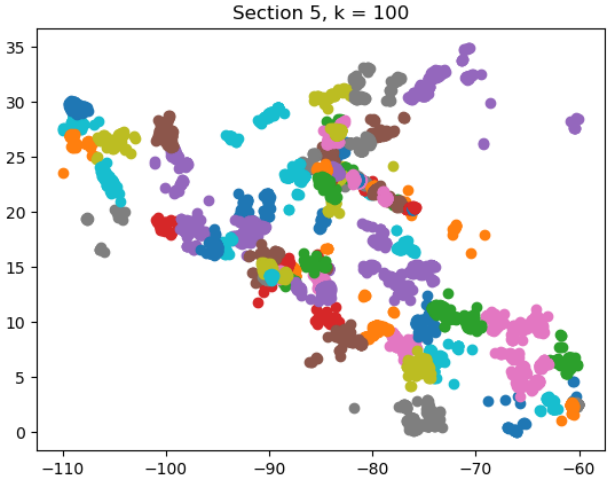
$$where\ N\ = number\ of\ samples\ in\ that\ cluster, O = object\ and\ C = cluster$$

To update the medoid, we do that by finding the total sum of all the data added together and then finding the mean of that data dividing the values with the number of data. We use this to find the new medoids to confirm if the data has converged because if none of the medoid changes, we know the data has converged.

5. Choosing K

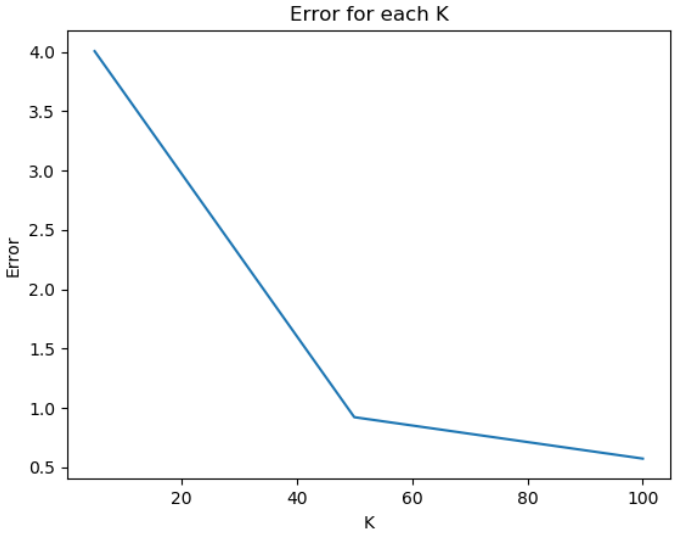






To achieve these plots, I first assigned all samples to the zero cluster and chose medoid using the K-first samples.

I then group the samples into new clusters using the formula in section 3 which groups samples to medoids closest to them. I then check if the samples remained in the same clusters. If they did, the samples have converged, if not I will calculate the new medoids using the formula in section 4, which takes the mean of the points in that cluster.



For k = 5, the error I got is 4.006, for k = 50, error = 0.921 and for k = 100, error = 0.5725.

From the graph above, we can tell the error drops a significant amount from k = 5 to k = 50, meaning samples have much lesser variance and are much closer to its local neighbours, making it a better k. However, the difference from k = 50 to k = 100 is significantly smaller. The error will always get lower as K decreases, however lower does not necessarily means its better. The best K to choose from would be the k that has a huge decrease in error, so k = 50 is the best k to choose.