2IMW30 - FOUNDATION OF DATA MINING
Assignment 1b Report

Duy Pham (0980384, d.pham.duy@student.tue.nl)
Mazen Aly(0978251, m.aly@student.tue.nl)

(Both students contributed equally in this assignment)

EXERCISE 1

In this assignment, we implement the k-means algorithm for clustering the input data points. Clustering is considered as a unsupervised machine learning problem, that means there no labels in the data that we need to predict as classification, but it helps us getting some insights about the input data.

We developed four methods for initializing the first K cluster centroids (where K is an input) The first one is just choosing the first K data points to be the centroids, the second method is randomized selections of the initial centroids. the third method is k-means++ and the fourth method is using Gonzales algorithm.

For each method, we run k-means for different values of k 3,4,5 and in randomized methods, we run that 5 times. we visualise the results to get more understanding and learning experience as shown below.

In figure **??**, we shows the k-means clustering result on the dataset C2.txt with various initialization strategies and different values of k.

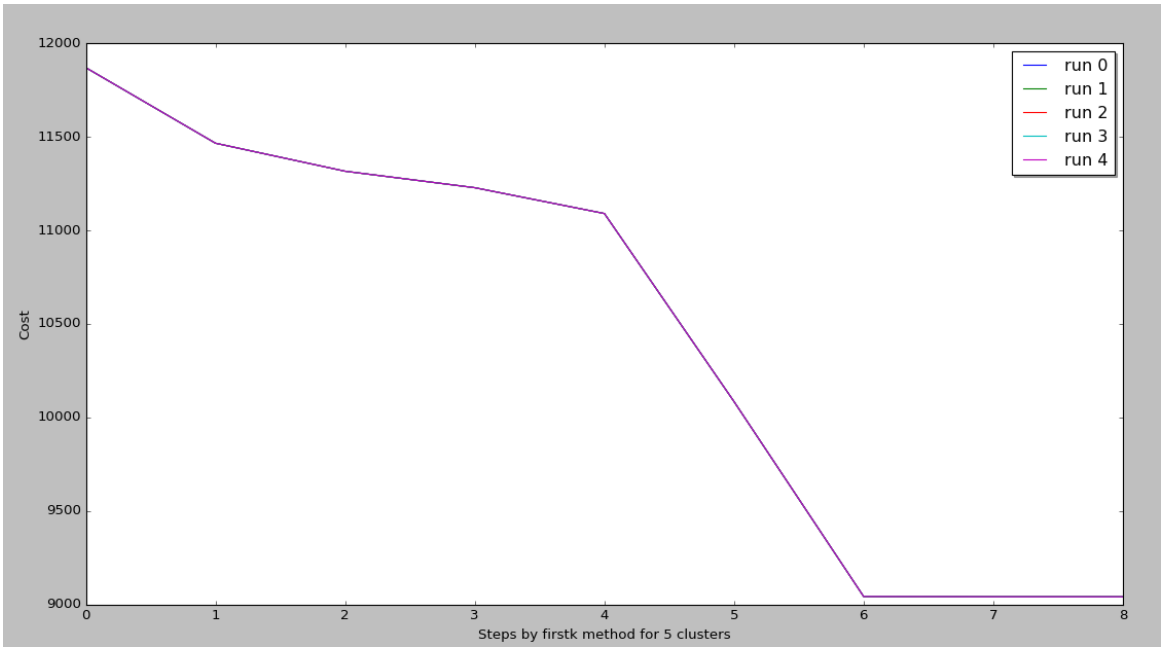


Figure 1

EXERCISE 2

In this excercise, we are proving that barycenter computed in the update step minimizes the cost function.

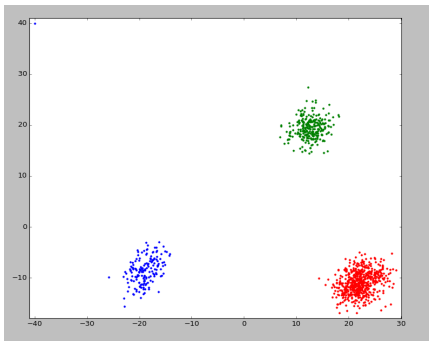
Let's first assume that $d = 1$. Then the cost function is

$$\phi(U, b) = \Sigma_{p_i \in U} ||p_i - b||^2$$

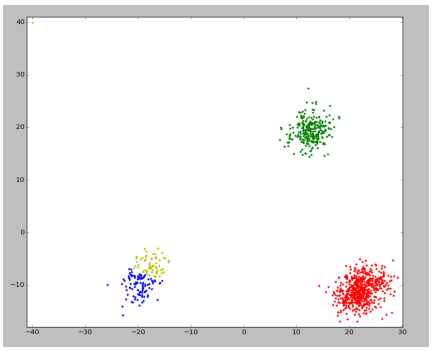
The function is minimized when its derivative is 0. In this case, we are considering the changes of the function with respect to the way we choose the center. Thus, we can take the partial derivative with respect to b.

$$\begin{aligned}
\frac{d}{d_b} \Sigma_{p_i \in U} ||p_i - b||^2 &= \Sigma_{p_i \in U} \frac{d}{d_b} ||p_i - b||^2 \\
&= \Sigma_{p_i \in U} 2 \cdot ||p_i - b|| \cdot \frac{||p_i - b||}{p_i - b} \cdot (-1) \\
&= 2\Sigma_{p_i \in U} (b - p_i)
\end{aligned}$$

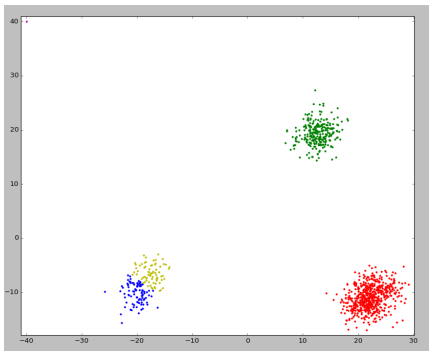
The cost function is minimized when



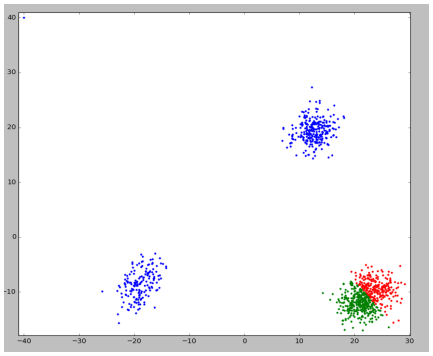
(a) Kmeans by picking first 3 points



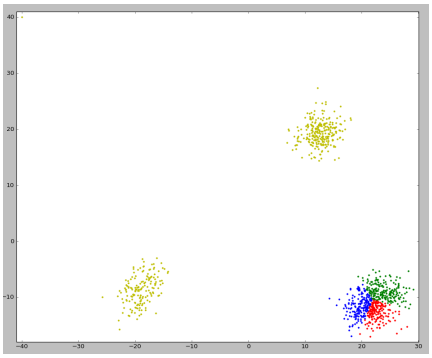
(b) Kmeans by picking first 4 points



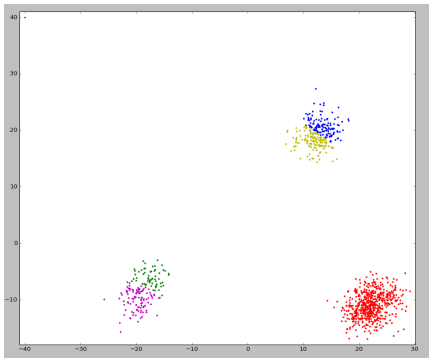
(c) Kmeans by picking first 5 points



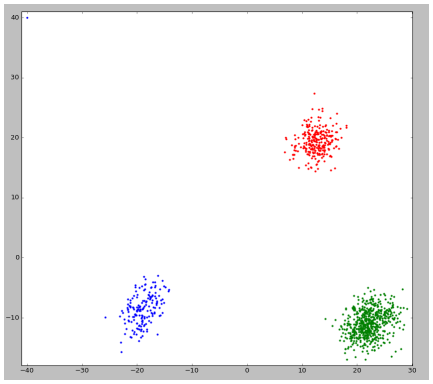
(d) Kmeans by picking first 3 random points



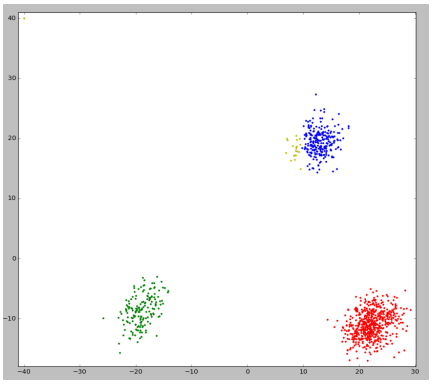
(e) Kmeans by picking first 4 random points



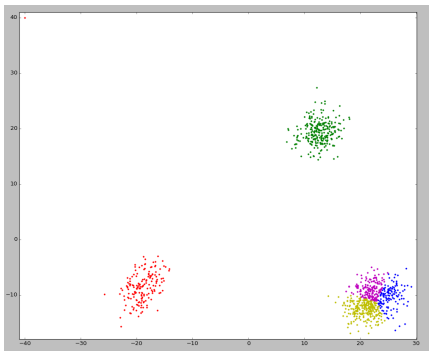
(f) Kmeans by picking first 5 random points



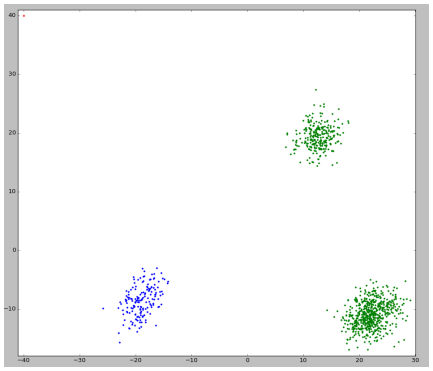
(g) Kmeans++ with k = 3



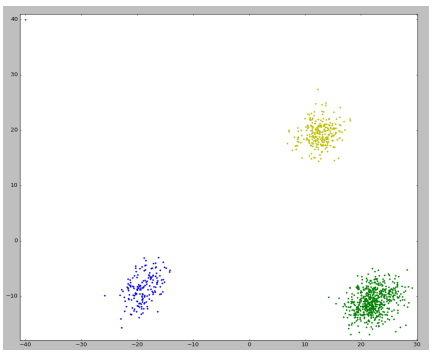
(h) Kmeans++ with k = 4



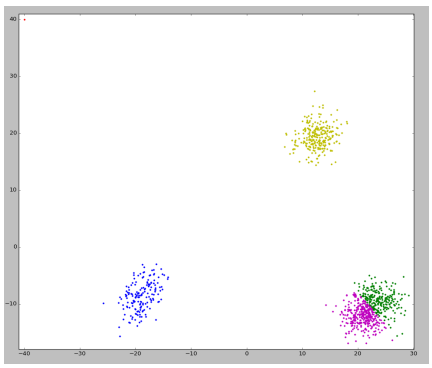
(i) Kmeans++ with k = 5



(j) Kmeans using Gonzales algorithm to choose the seeds, k = 3



(k) Kmeans using Gonzales algorithm to choose the seeds, k = 4



(l) Kmeans using Gonzales algorithm to choose the seeds, k = 5

Figure 2: K-means results with different settings

$$
\begin{aligned}
2\Sigma_{p_i \in U}(b - p_i) &= 0 \\
\Sigma_{p_i \in U} b &= \Sigma_{p_i \in U} p_i \\
b &= \frac{1}{n} \Sigma_{p_i in U} p_i
\end{aligned}
$$

where $n = |U|$.

The update step also assigns each point to the nearest center, which itself minimizes the total cost because the cost of each point is minimized.

When $d > 1$, the update step in each cluster is independent from the other clusters. Thus, the sum of the costs of each clusters is also minimized when each individual cost is minimized.

Therefore, the update step actually minimizes the cost function.

EXERCISE 3

In this exercise, we show an example that the cosine distance does not follow the triangle inequality.

Let p, q be 2 vectors where the angle α between them is $\frac{\pi}{2}$, and r be the bisector of that angle. Thus, the angle β between p and r and the angle γ between r and q are both $\frac{\pi}{4}$. We have:

$$
\begin{aligned}
dist(p, q) &= 1 - cosine(\alpha) = 1 \\
dist(p, r) + dist(r, q) &= 1 - cosine(\beta) + 1 - cosine(\gamma) \\
&= 2 - \sqrt{(2)} < 1
\end{aligned}
$$

which means the cosine distance does not follow the triangle inequality in this case.

This completes the example.

EXERCISE 4

a

In this exercise, we consider another variant of the clustering problem, the k-median problem. The idea is similar to k-means in the sense that the distance from each point to its center is minimized.

The k-median approach can be more robust against noise because noise does not affect the order of the data, so the medians are not modified if the noise increases. If we use k-means, noise can biase the centroids.

b

We use a similar algorithm to k-means, just replace the means by the medians. Thatis, when updating a cluster, we take the point having the medians of the features as coordinates as the new center. This approach is obviously not optimal in term of cost-minimization, as the solution is proven in Exercise 2. So we will experiment the actual cost we get in the set C3.txt.