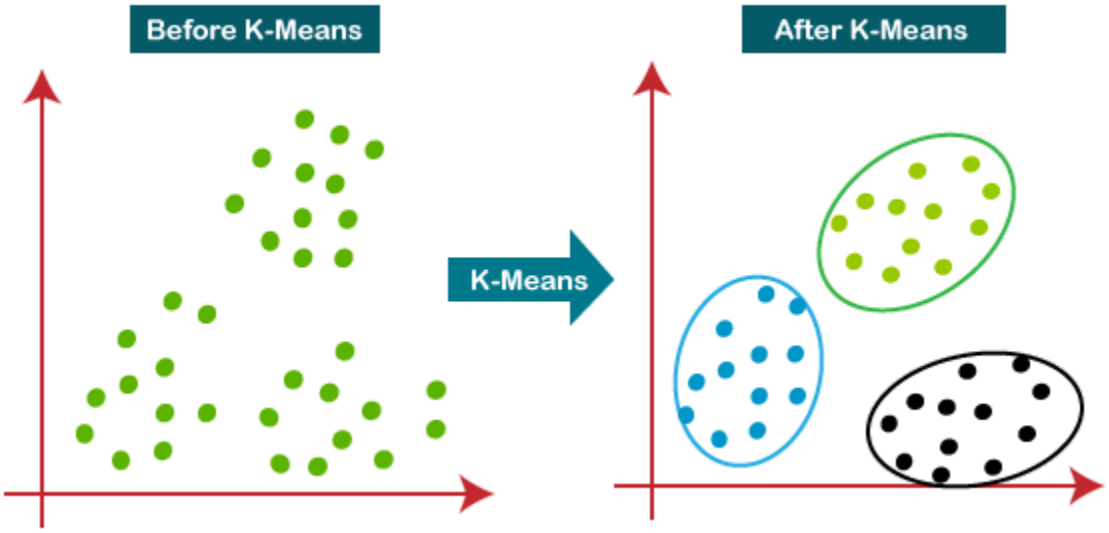
High Performance Computing
Assignment 2

2022 - 2023

Read the requirement carefully and submit Assignment deliverables as mentioned in the pdf

K-means Clustering

is a method of vector quantization that aims to partition *n* observations into *k* clusters in which each observation belongs to the cluster with the nearest mean (cluster centers or cluster centroid), serving as a prototype of the cluster [1]



K means algorithm is very popular and used in a variety of applications such as market segmentation, document clustering, image segmentation and image compression, etc. The goal usually when we undergo a cluster analysis is either:

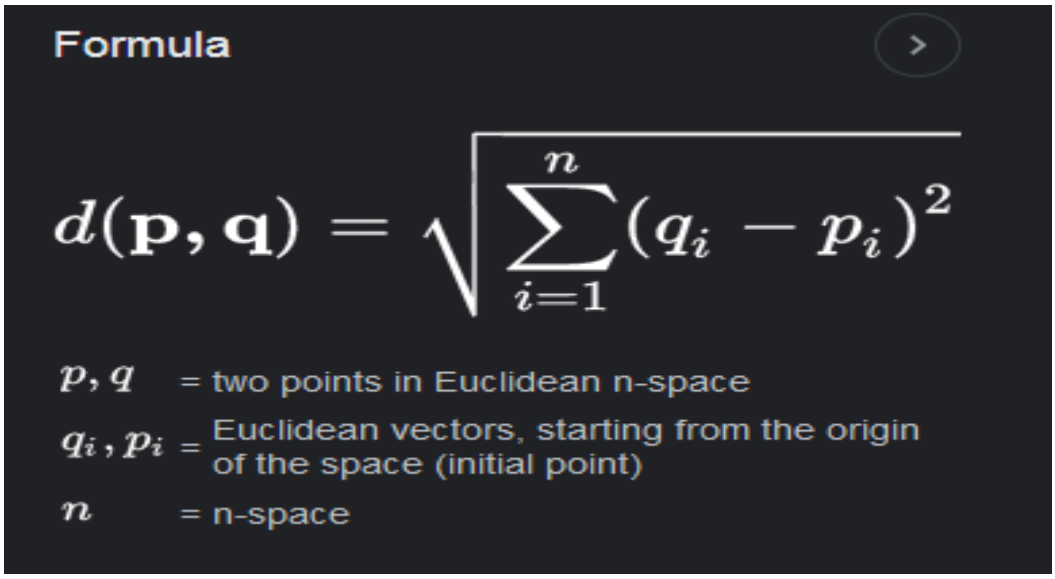
1. Get a meaningful intuition of the structure of the data we're dealing with.
2. Cluster-then-predict where different models will be built for different subgroups if we believe there is a wide variation in the behaviors of different subgroups. An example of that is clustering patients into different subgroups and building a model for each subgroup to predict the probability of the risk of having a heart attack.

The main algorithm steps as follow: [2][3]

Algorithm 1 k-means algorithm

1: Specify the number k of clusters to assign.
2: Randomly initialize k centroids.
3: **repeat**
4: **expectation:** Assign each point to its closest centroid.
5: **maximization:** Compute the new centroid (mean) of each cluster.
6: **until** The centroid positions do not change.

One of popular methods to calculate the distance between point and its closed centroid is **Euclidean distance** [4]



Requirements

You are requested to implement the k means algorithm using the open MP function only.

Tips:

1. Number of data points is known (you are flexible to define how many data points you would use in your program).
2. The datapoint are in 2 dimension i.e, each point is represented as (x,y)
3. You would use only openMP apis. You shouldn't use mpi apis.
4. You should read your points from the data file in your c script (points are defined by you).
5. The program output is as follow
   Cluster 1:
   (x1,y1)
   (x2, y2)
   Where x is the location of point in x dimension and y is the location of point in y dimension and cluster is in which cluster this point belongs to.
6. Number of threads equals the number of clusters you will pass to the program.
7. You are flexible to choose the number of times the algorithm should be repeated until you reach your stopping criteria (you have flexibility to choose maximum number of iteration and/or choose the threshold of centroid difference error).
8. To calculate the distance use the mean square error equation.

Suggested Pseudo K means algorithm using open MP would be as follow:

Step	Executor	Grade
1. Read The data file	Main Thread	10
2. Initiate 2 random numbers for each thread/cluster . the 2 random numbers are representing (x, y) initial of cluster centroid	Main Thread	10
3. Calculate the distance between each point and cluster centroid. using the formula mentioned in page 1	Forked Threads	20
4. Filter each point distances depending on minimum value	Main Thread	20
5. Calculate the mean for each cluster as new cluster centroid	Forked Threads	20
6. Repeat steps 3 to 5 with the new cluster centroid until the ending criteria you specified.		20
Total Grade		**100 ***

* The grade including:

1. Fully utilize for the thread (as possible)
2. Accuracy value of applying algorithm and algorithm result (as possible)
3. Code format (must be formatted and well organized)
4. If your code is not running you will lose half of the grades
5. If your code doesn't show output you will lose ¼ grades
6. If the output does not match the informed output you will lose grade
7. You will lose points if you use any of the mpi apis.

Delivery Notes:

- This is a group assignment of **3 or 4 members** . If you submit as a group of more less than of 4 members or less than 3 members, All the group members will get **zero**
- All students should work and fully understand everything in the submitted solution.
- No late submission is allowed.
- Submissions will be on the blackboard. It is your duty to ensure that your submission was properly uploaded to the blackboard after you finish submitting it. If your submission was not uploaded properly while marking, you will not receive a grade for the assignment.
- No submission through e-mails.
- You will put your code .c file and your data in a folder named **CS371_Assign2_firstStudentID_secondStudentID_thirdStudentID** and compress them to a .zip file with the same folder name. The compressed file would be the file to be delivered.
- Failing to abide by the naming convention of the file or failing to submit the files as per the requested extension, would result in a **zero** for all team members.
- In case of **cheating** you will get a **negative grade** whether you give your solution to someone, take the solution from someone/internet, or even send it to someone for any reason.
- If the team or any member didn't attend the discussion without formal excuse before the discussion time the whole team will get **zero grade.**
- Due Date **14/05/2022**