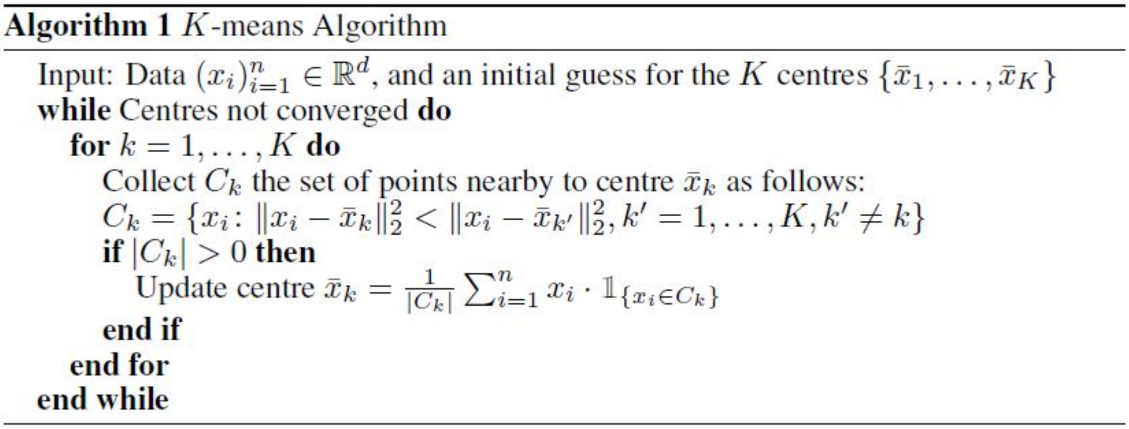
Lab 4: Clustering



Algorithm 1 K-means Algorithm

```
Input: Data $(x_i)_{i=1}^n \in \mathbb{R}^d$, and an initial guess for the $K$ centres $\{\bar{x}_1, \ldots, \bar{x}_K\}$
while Centres not converged do
    for $k = 1, \ldots, K$ do
        Collect $C_k$ the set of points nearby to centre $\bar{x}_k$ as follows:
        $C_k = \{x_i \colon \|x_i - \bar{x}_k\|_2^2 < \|x_i - \bar{x}_{k'}\|_2^2, k' = 1, \ldots, K, k' \neq k\}$
        if $|C_k| > 0$ then
            Update centre $\bar{x}_k = \frac{1}{|C_k|} \sum_{i=1}^n x_i \cdot \mathbb{1}_{\{x_i \in C_k\}}$
        end if
    end for
end while
```

1. Write a function to perform k-Means clustering of a given dataset. The function should take the following arguments: **(30 marks)**

    a) the dataset for clustering
    b) the number of clusters, **k**
    c) the initial centroids (optional)

    If the initial centroids are not provided, **k random points** are chosen as initial centroids.

    The function should return:
    a) the final cluster centroids
    b) cluster label associated with each datapoint
    c) sum of squared errors

    The **sum of squared errors(SSE)** is computed as follows, where k is the number of clusters and μ_j is the centroid of the j^{th} cluster:

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} ||x_i - \mu_j||_2^2$$

2. Generate a dataset by sampling **20** points **each** from uniform([-1,1]) and uniform([-0.5,1.5]). **(10+10 marks)**

    a. Run k-means with the following initial centroids.

        i. $\mu_1^{initial}$ = -0.1 and $\mu_2^{initial}$ = 0.1
        ii. $\mu_1^{initial}$ = 0 and $\mu_2^{initial}$ = 3.5

    After each iteration, generate a scatterplot of the dataset such that points belonging to the same cluster are given the same colour. Use different colours for different clusters. Display the cluster centroids with * (asterisk symbol)

b. Now add a random point generated from uniform([3,4]) to the dataset. Perform k-means clustering with k=2 for different sets of initial centroids. What do you observe? Are the clusters always found correctly?

3. There are three groups of people, say, **Kids, Adults and Aliens**. Each person has two features: **height** and **weight**, i.e., the data point x_i is represented as ($x_i(1)$, $x_i(2)$) where $x_i(1)$ represents height and $x_i(2)$ represents weight. The features for each group are distributed as follows:

Group	Height	Weight	No: of samples
Kids	Normal(5,1.1)	Normal(60,7)	100
Adults	Normal(3,1)	Normal(30,5)	100
Aliens	Normal(7,1)	Normal(40,2)	50

Run k-means on this dataset with different sets of initial centroids. Display the clusters after each iteration, as mentioned in the previous question. **(10+10+5+10 marks)**

a. Generate a plot of the sum of squared errors (SSE) against iteration number. Against each iteration number (x-axis), plot the SSE obtained (y-axis) in that iteration

b. Are you able to obtain distinct sets of final clusters when starting with different initial centroids? If yes, show at least two of such clusterings. In each case, show the initial cluster centroids in the scatterplot using a + (plus symbol).

c. If you were to select one clustering result from among the different clusterings obtained, how would you make a choice?

d. How can you modify your k-means algorithm such that for a given value of k, different results are not obtained on successive runs over the same dataset?

4. Plot the data in “Dataset.csv”. Visually identify the clusters. Let k be the number of clusters identified. **(5+10 marks)**

a. Run k-means on this dataset with the value of k identified above. Do you get the expected clusters? Why or why not?

b. Now run k-means for k = 2 to 10 on this dataset and for each value of k (x-axis), plot the **best** SSE obtained for that value of **k** (y-axis). How will you select a good value of k from this plot?