HOMEWORK #1

Software Project (0368-2161)

Due Date: 11/05/2021

Introduction

The K-means algorithm is a popular clustering method for finding a partition of N unlabeled observations into K distinct clusters, where K is a parameter of the method. In this assignment you will implement this algorithm in both Python and C. The goals of the assignment are:

- Practice the material taught in class both for C and Python.
- Transform a known algorithm into working executable code.
- Read input stream and process it.
- Create an interface for programs.
- Experience the difference in the programming effort and the running time of both languages.

K-means

Given a set of N datapoints $x_1, x_2, \ldots, x_N \in R^d$, the goal is to group the data into K clusters, each datapoint is assigned to exactly one cluster and the number of clusters K is such that $K < N$. We will denote the group of clusters by $S_1, S_2, \ldots, S_K$, each cluster S_j is represented by it's centroid which is the mean $\mu_j \in R^d$ of the cluster's members.

Algorithm 1 k-means clustering algorithm

1: Initialize centroids $\mu_1, \mu_2, \ldots, \mu_K$ as first k datapoints $x_1, x_2, \ldots, x_K$
2: **repeat**
3: Assign x_i to the closest cluster S_j:

$$\underset{S_j}{\text{argmin}} \ (x_i - \mu_j)^2, \ \forall j \ 1 \leq j \leq K$$

4: update all centroids as follows:

$$\mu_i = \frac{\sum_{x_l \in S_i} x_l}{|S_i|}$$

5: **until** convergence: $(no\ change\ in\ \mu)\ OR\ (iteration\_number = max\_iter)$

Assignment Description

Implement the k-means algorithm as detailed in 1.1 both in C and Python.
The behavior of the program is as follows:

1. The program receives the arguments: K, filename and optional(max_iter):
   (a) K – the number of clusters required.
   (b) filename - *.txt file that contains datapoints separated by commas.
   (c) max_iter – the maximum number of iterations of the K-means algorithm, if not provided the default value is 200.
2. The final centroids are returned to the cmd.