K-means

Anmol Anand
Department of Computer Science
Texas A&M University
aanand@tamu.edu

1 Introduction

K-means is a clustering algorithm aiming to partition a dataset into K distinct clusters, where each data point belongs to the cluster with the nearest mean (centroid). The quantitative goal is to minimize the sum of squared distances between data points and their respective cluster centroids, effectively grouping similar data points together while maximizing dissimilarity between clusters.

This project aims to compare the performance of Lloyd's algorithm when it uses two different centroid initialization methods - D2 Sampling versus Metropolis Hastings. When initializing with Metropolis Hastings, different values of Markov chain length have been used. Moreover, these experiments are run for different number of clusters while generating samples and while performing Lloyd's algorithm.

2 Code

The code can be found in this Github repo github.com/anmol-anand/k-means. The instructions to run the code are mentioned in the README.
Note: For each set of hyperparameters $\{K$, initialization method(D2 or Metropolis Hastings), Markov chain length$\}$, Lloyd's algorithm is run 20 times and the average performance of these 20 runs is considered as the performance for this set of hyperparameters.

3 Sample Generation

I have generated 1000 samples such that we already know the ground truth clusters. To achieve this, first, centroids are randomly generated under the constraint that the distance between any two centroid pairs is at least $2 \cdot RADIUS$ where $RADIUS$ is a predefined constant (defined as 10^6 in generate_samples.py). Then, the samples in each cluster are chosen such that their distance from the respective ground truth centroid is less than $RADIUS$. This ensures that they will belong to this cluster as their distance from other centroids will be greater than $RADIUS$ (because of triangle inequality).

4 Analysing the results

After generating the samples, Lloyd's algorithm is run either with D2 sampling initialization (green line plots) or Metropolis Hastings initialization (red line plots) of cluster centroids. This experiment is run for three values of $K = \{10, 100, 500\}$ While using Metropolis Hastings initialization, the experiment is performed with different lengths of Markov chain varying on the x-axis of the plots. Turns out, the D2 Sampling approach always performs better than Metropolis Hastings in terms of accuracy. The performance is measured on the y-axis in terms of the sum of squared distance of samples from respective cluster centroids. The smaller this value, the better the performance.

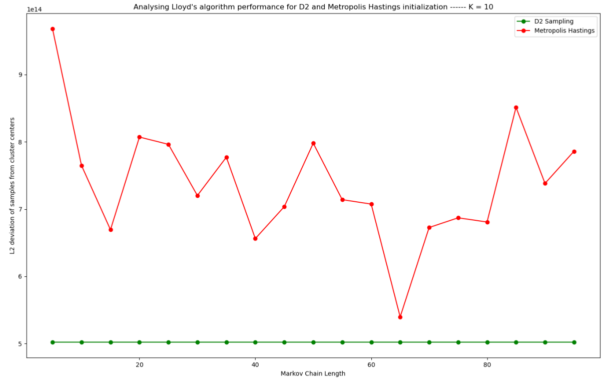

Figure 1: K = 10

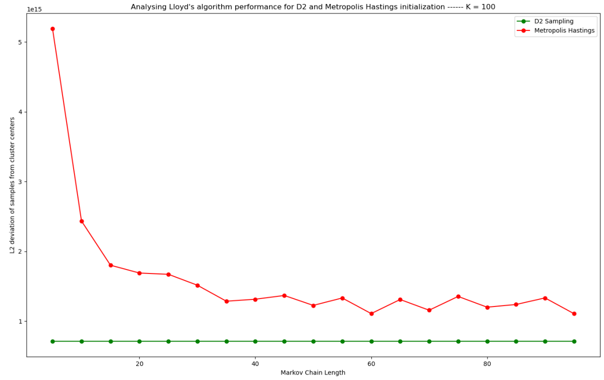

Figure 2: K = 100

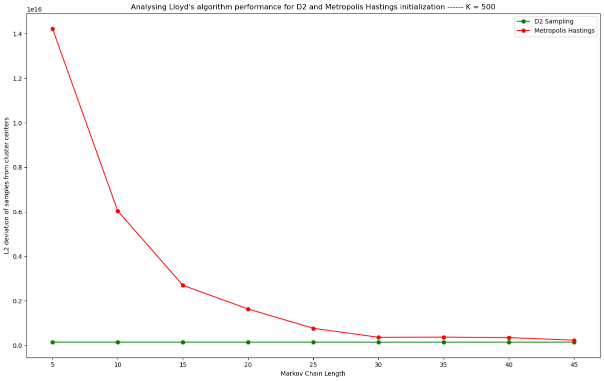

Figure 3: K = 500