WITS UNIVERSITY

School of Computer Science and Applied Mathematics
UNIVERSITY OF THE WITWATERSRAND, JOHANNESBURG

COMS4040A & COMS7045A Project – Report
K-Means and Fuzzy C-Means Clustering
Parallel Machine Learnig Algorithms

Shameel Nkosi, 1814731, Coms Hons
Siraj Motaung BDA Hons

July 3, 2021

1 Introduction

In machine learning, clustering is a technique of grouping objects of similar characteristics into the same groups called clusters. The similarity between any two objects is measured by their distance from each other.

In this project, we have implemented the K-Means clustering algorithm and the Fuzzy C-Means algorithm. Traditionally, we aim to classify objects into separate clusters. The K-Means algorithm allows us to achieve this type of clustering, where an object belongs to only one cluster. On the other hand, Fuzzy C-Means allows us to classify an object as a member of different clusters. The K and C in K-Means and Fuzzy C-Means respectively are hyperparameters that specify the number of clusters or groups we wish to have.

1.1 Problem Statement

In this project, we have taken a dataset from the UCI Repository. This data set describes the chemical composition of the wine. We want to use these chemical compositions to find the origins of the wine. These wines are from 3 different cultivars in the same region in Italy. We, therefore, aim to find out which of these wines belong to which cultivar.

2 Methodology

2.1 Algorithms

This subsection describes the algorithms we implemented in-depth.

2.1.1 K-Means Clustering

As mentioned above, K-Means partitions all objects into K clusters or subsets of the data set. Each object is an observation in our data set, which belongs to the nearest mean among the K means. Initially, we need to initialize the K mean as points in the same space as all the data points. The easiest way is to choose k random points as the initial means. Mathematically we can describe the problem as follows: Given a set of observations in a d-dimensional space, K-Means partitions these observations into $k \leq n$ sets $S = \{S_1, S_2, .., S_k\}$. We then want to find the minimum distance between each observation and the means. Formally, we to find:

$$\arg\min_S \sum_{i=1}^{k} \sum_{\mathbf{x} \in S} \|\mathbf{x} - \mu_i\|^2 \tag{1}$$

K-Means is an iterative algorithm. We iteratively update the means until there is stability i.e. the means aren't changing anymore. In our implementation, however, we set several epochs so that we can measure the time it takes for each implementation. The algorithm runs as follows:

- choose k random obeservations as your initial centroids or means
- Repeat for a specified number of epochs or until convergence:
  - Calculate the distance of each data point with all the means.
  - Assign each data point to a cluster with the least distance
  - recalculate the means by assigning the mean to the average means of the data points