CHAPTER 1

# Clustering using k-Means

K-means is a popular unsupervised learning algorithm used for data clustering. The goal of k-means is to group data points into distinct non-overlapping subgroups, or clusters, based on their features.

## 1.1 The K-Means Algorithm

Given a dataset $X = \{x_1, x_2, \ldots, x_N\}$, where each $x_i$ is a d-dimensional vector, and an integer $k$, the k-means clustering algorithm seeks to find $k$ cluster centroids $C = \{c_1, c_2, \ldots, c_k\}$ such that the distance from each data point to its nearest centroid is minimized.

The k-means algorithm works as follows:

1. Initialize $k$ centroids randomly.
2. Assign each data point to the nearest centroid. This forms $k$ clusters.
3. For each cluster, update its centroid by computing the mean of all points in the cluster.
4. Repeat steps 2 and 3 until the centroids do not change significantly or a maximum number of iterations is reached.

The measure of distance typically used in k-means is the Euclidean distance. For two d-dimensional vectors $x = (x_1, \ldots, x_d)$ and $y = (y_1, \ldots, y_d)$, the Euclidean distance is defined as:

$$d(x, y) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \ldots + (x_d - y_d)^2} \tag{1.1}$$

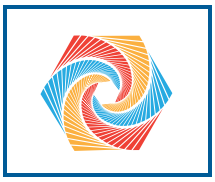
## 1.2 Choosing K

Choosing an appropriate value for $k$ is a significant aspect of the k-means algorithm. One common method is the Elbow Method, which involves plotting the explained variation as a function of the number of clusters, and picking the elbow of the curve as the number of clusters to use.

*This is a draft chapter from the Kontinua Project. Please see our website (https://kontinua.org/) for more details.*

Appendix A

# Answers to Exercises



# INDEX

k-Means Clustering, 1