Python for Engineering Data Analysis Mini-Report

Short Dive into K-means Clustering

8th of January 2023

Katharina Julia Brenner
Chaitanya Chawla

Student IDs:	ge54xoj, ge78yof
Module:	[EI04024] Python for Engineering Data Analysis
Supervisors:	M.Sc. Felix Mayr

Professorship Simulation of Nanosystems for Energy Conversion
Technical University Munich
School of Computation, Information and Technology

Contents:

1 Introduction

1.1 Clustering

Clustering is a type of **Unsupervised Learning method** [1], i.e., it learns through datasets consisting of unlabeled input data. It is the task of dividing the data points into groups such that the data points in the same groups are similar to each other and dissimilar to those in the other groups.

The different types of methods for clustering include – Density-Based Methods [2], Hierarchical Based Methods [3], Partitioning Methods [4], and Grid-Based Methods [5].

1.2 K-means Clustering

K-means clustering is one of the simplest clustering algorithms. The "k" refers to the number of clusters in the solution. The given n observations are assigned a cluster with the nearest mean serving as a reference of the cluster.

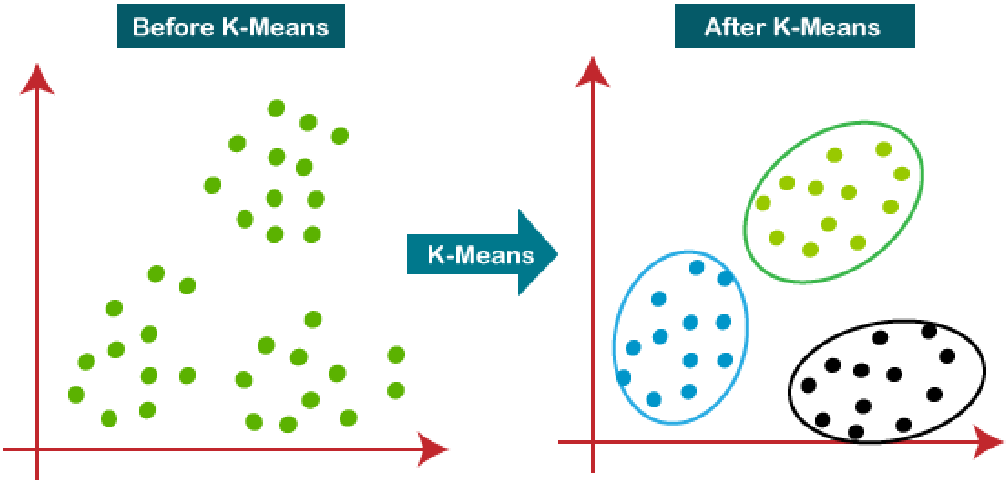

Figure 1: Figure showing the transition between unclassified Data into clusters with the help of K-means Clustering.

2 Algorithm

2.1 Pseudo - Code

The steps for the pseudo-code are as follows –

(i) Initialize k random points (called means or cluster centroids) within the boundaries of the data set.

(ii) Iterate through items
   a. Find the mean closest to the item by calculating the Euclidean distance [5] of the item with each of the mean

b. Assign the item to this specific mean
c. Update mean by shifting it to the average of the items in that new cluster

(iii) Step (ii) is repeated for a fixed number of iterations. If between two iterations, no item changes classification, then algorithm is supposed to find the optimal solution and further iterations don't need to be executed.

2.2 Initializing Means

Each mean has the same dimension as the number of features in each item.

We initialize each mean's value in the range of the feature values of all the items. For that we can find the minimum and maximum value of each feature, and randomly assign values for each feature of the means.

2.3 Determining the value of K

Elbow method [6] – The basic idea behind this method is that it plots the various values of cost with changing *k*. As the value of *K* increases, there will be fewer elements in the cluster. So average distortion will decrease. The point where this distortion declines the most is the **elbow point**, and this k is taken as the optimal k

2.4 Variations

Instead of the Euclidean Distance, other commonly used measures are –

(i) **Cosine Distance** - determines the cosine of the angle between the point vectors of the two points in the n-dimensional space

$$d = \frac{X.Y}{||X||*||Y||}$$

(ii) **Manhattan Distance** - computes the sum of the absolute differences between the coordinates of the two data points

$$d = \sum_n X_i - Y_i$$

(iii) **Minkowski Distance** - is a generalized version of the Euclidean distance and the Manhattan distance

$$d = (\sum_n |X_i - Y_i|^{\frac{1}{p}})^p$$

4 Bibliography

[1] Barlow, Horace B. "Unsupervised learning." *Neural computation* 1.3 (1989): 295-311.

[2] Kriegel, Hans-Peter, et al. "Density-based clustering." *Wiley interdisciplinary reviews: data mining and knowledge discovery* 1.3 (2011): 231-240.

[3] McInnes, Leland, John Healy, and Steve Astels. "hdbscan: Hierarchical density based clustering." *J. Open Source Softw.* 2.11 (2017): 205.

[4] Äyrämö, Sami, and Tommi Kärkkäinen. "Introduction to partitioning-based clustering methods with a robust example." *Reports of the Department of Mathematical Information Technology. Series C, Software engineering and computational intelligence* 1/2006 (2006).

[5] Cheng, Wei, Wei Wang, and Sandra Batista. "Grid-based clustering." *Data clustering*. Chapman and Hall/CRC, 2018. 128-148.

[6] Humaira, H., and R. Rasyidah. "Determining the appropiate cluster number using Elbow method for K-Means algorithm." *Proceedings of the 2nd Workshop on Multidisciplinary and Applications (WMA)*. 2020.