National University of Computer and Emerging Sciences



Lab Manual 12

AL2002-Artificial Intelligence Lab

|  |  |
| --- | --- |
| Course Instructor | Dr. Hajra Waheed |
| Lab Instructor (s) | Muhammad Ahmed  Hamza Ayub |
| Section | B1 – B2 |
| Semester | Spring 2023 |

Department of Computer Science

FAST-NU, Lahore, Pakistan

Table of Contents

[1](#_1fob9te) Objectives 3

[2](#_3znysh7) Task Distribution 3

[3](#_2et92p0) K-Means Clustering 3

[3.1](#_tyjcwt) Introduction - Clustering 3

[3.2](#_3dy6vkm) Type of Learning Problem 3

[3.3](#_1t3h5sf) Clustering Applications 3

[3.4](#_4d34og8) Clustering Evaluation Metric – Inertia 4

[3.5](#_2s8eyo1) How K-Means Clustering Works 4

[3.5.1](#_17dp8vu) Algorithm Steps 4

[3.5.2](#_3rdcrjn) Stopping Criteria for K-Means Clustering 4

[4](#_26in1rg) K-Medoids Clustering 5

[4.1](#_lnxbz9) Introduction 5

[4.2](#_35nkun2) How K-Medoid Clustering Works 5

[4.3](#_1ksv4uv) How K-Medoid is Better Than K-Means 5

[5](#_44sinio) Exercise (50 Marks) 5

[5.1](#_2jxsxqh) Steps to perform 5

[6](#_z337ya) Submission Instructions 6

# Objectives

After performing this lab, students shall be able to understand the following:

* K-Means clustering
* K-Medoids clustering
* Application of K-Means & K-Medoids clustering with scikit-learn

# Task Distribution

|  |  |
| --- | --- |
| **Total Time** | **170 Minutes** |
| K-Means Clustering | 20 Minutes |
| K-Medoids clustering | 20 Minutes |
| Exercise | 120 Minutes |
| Online Submission | 10 Minutes |

# K-Means Clustering

## Introduction - Clustering

*Clustering is the process of dividing the entire data into groups (also known as clusters) based on the patterns in the data.* All the data points in a cluster should be similar to each other. The data points from different clusters should be as different as possible.

For example, a bank wants to give credit card offers to its millions of customers. Does it make sense to look at the details of each customer separately and then make a decision? Certainly not! It is a manual process and will take a huge amount of time.

One option is to segment its customers into different groups. For instance, the bank can group the customers based on their income: High Income, Average Income, & Low Income.

Instead of creating different strategies for individual customers, the bank can now make three different strategies or offers, one for each group.

## Type of Learning Problem

Clustering is an unsupervised learning problem. In clustering, we do not have a target to predict. We look at the data and then try to club similar observations and form different groups. Hence it is an unsupervised learning problem.

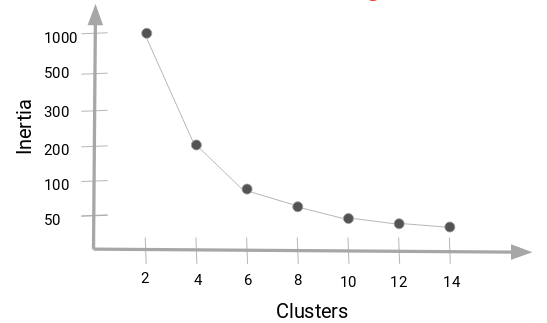
## Clustering Applications

Following are some of the applications of clustering:

* Customer Segmentation: which includes telecom, e-commerce, sports, advertising, sales, etc.
* Document Clustering: cluster similar documents together.
* Image Segmentation: to create clusters having similar pixels in the same group
* Recommendation Engines: Recommend similar products to user based on user’s history

## Clustering Evaluation Metric – Inertia

Inertia calculates the sum of distances of all the points within a cluster from the center (centroid) of that cluster. It tells us how far the points are within a cluster. We calculate this for all the clusters and the final inertial value is the sum of all these distances. The distance between the points in the clusters should be as low as possible. If the distance between the centroid of a cluster and the points in that cluster is small, it means that the points are closer to each other. We can say that the lesser the inertia value, the better our clusters are.



In the above image, when we changed the cluster value from 2 to 4, the inertia value reduced very sharply. This decrease in the inertia value reduces and eventually becomes constant as we increase the number of clusters further. So, the cluster value where this decrease in inertia value becomes constant can be chosen as the right cluster value for our data.

## How K-Means Clustering Works

K-means is a centroid-based algorithm, or a distance-based algorithm, where we calculate the distances to assign a point to a cluster. In K-Means, each cluster is associated with a centroid. The main objective of the K-Means algorithm is to minimize the sum of distances between the points and their respective cluster centroid.

### Algorithm Steps

1. Choose the number of clusters *k*
2. Select k random points from the data as centroids
3. Assign all the points to the closest cluster centroid
4. Recompute the centroids of newly formed clusters
5. Repeat steps 3 and 4

### Stopping Criteria for K-Means Clustering

There are essentially three stopping criteria that can be adopted to stop the K-means algorithm:

1. Centroids of newly formed clusters do not change.
2. Points remain in the same cluster.
3. Maximum number of iterations are reached.

# K-Medoids Clustering

## Introduction

K-Medoids is a clustering algorithm resembling the K-Means clustering technique. It falls under the category of [unsupervised machine learning](https://en.wikipedia.org/wiki/Unsupervised_learning). It majorly differs from the K-Means algorithm in terms of the way it selects the clusters’ centres. The former selects the average of a cluster’s points as its centre (which may or may not be one of the data points) while the latter always picks the actual data points from the clusters as their centres (‘medoids’).

## How K-Medoid Clustering Works

The steps followed by the K-Medoids algorithm for clustering are as follows:

1. Randomly choose ‘k’ points from the input data (‘k’ is the number of clusters to be formed).
2. Each data point gets assigned to the cluster to which its nearest medoid belongs.
3. For each data point of cluster i, its distance from all other data points is computed and added. The point of ith cluster for which the computed sum of distances from other points is minimal is assigned as the medoid for that cluster.
4. Steps (2) and (3) are repeated until convergence is reached i.e. the medoids stop moving.

## How K-Medoid is Better Than K-Means

Mean of the data points is a measure that gets highly affected by the extreme points. So in K-Means algorithm, the centroid may get shifted to a wrong position and hence result in incorrect clustering if the data has outliers. On the contrary, a medoid in the K-Medoids algorithm is the most central element of the cluster, such that its distance from other points is minimum. Since medoids do not get influenced by extremities, the K-Medoids algorithm is more robust to outliers and noise than K-Means algorithm.

# Exercise (50 Marks)

Perform K-Means & K-Medoid clustering on the following dataset: **‘credit\_card.csv’.** Available on Google Classroom

## Steps to perform

Following steps should be performed:

1. Explore the dataset
2. Import the required libraries
3. scikit-learn for K-Means: <https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>
4. [scikit-learn-extra](https://scikit-learn-extra.readthedocs.io/en/latest/index.html" \l ":~:text=scikit%2Dlearn%2Dextra%20is%20a,novelty%20or%20lower%20citation%20number.) for K-Medoids: <https://scikit-learn-extra.readthedocs.io/en/latest/generated/sklearn_extra.cluster.KMedoids.html>
5. Standardize the data using sklearn StandardScaler: <https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.StandardScaler.html>
6. Fit the dataset on K-Means & K-Medoid models.
7. Display the cluster centers and inertia for both techniques
8. Find the best k value by plotting the inertia vs number of clusters graph.
9. Using the best k value, perform predictions.

# Submission Instructions

Always read the submission instructions carefully.

* Rename your Jupyter notebook to your roll number and download the notebook as **.ipynb** extension.
* To download the required file, go to **File->Download .ipynb**
* Only submit the **.ipynb** file. DO NOT **zip** or **rar** your submission file.
* Submit this file on Google Classroom under the relevant assignment.
* Late submissions will not be accepted.