National University of Computer and Emerging Sciences

Lab Manual Kmeans
Artificial Intelligence Lab

Department of Computer Science
FAST-NU, Lahore, Pakistan

Table of Contents

1 Objectives

After performing this lab, students shall be able to understand unsupervised clustering using state of the art machine learning model K-Means Clustering.

2 Task Distribution

Total Time	**170 Minutes**
Unsupervised Learning Concepts	30 Minutes
Clustering	20 Minutes
Kmeans	20 Minutes
Exercise	90 Minutes
Online Submission	10 Minutes

1. 3. Machine Learning Concepts:

Unsupervised Learning is a machine learning technique in which the users do not need to supervise the model. Instead, it allows the model to work on its own to discover patterns and information that was previously undetected. It mainly deals with the unlabelled data.

a. 3.1 Algorithms:

Unsupervised Learning Algorithms allow users to perform more complex processing tasks compared to supervised learning. Although, unsupervised learning can be more unpredictable compared with other natural learning methods. Unsupervised learning algorithms include clustering, anomaly detection, neural networks, etc

b. 3.2 Why Unsupervised learning?:

Here, are prime reasons for using Unsupervised Learning:

- Unsupervised machine learning finds all kind of unknown patterns in data.
- Unsupervised methods help you to find features which can be useful for categorization.
- It is taken place in real time, so all the input data to be analyzed and labeled in the presence of learners.
- It is easier to get unlabeled data from a computer than labeled data, which needs manual intervention.

c. 3.3 Types of unsupervised learning

Unsupervised learning problems further grouped into clustering and association problems.

d. 3.4 Clustering:

Clustering is an important concept when it comes to unsupervised learning. It mainly deals with finding a structure or pattern in a collection of uncategorized data. Clustering algorithms will process your data and find natural clusters(groups) if they exist in the data. You can also modify how many clusters your algorithms should identify. It allows you to adjust the granularity of these groups.

e. 3.5 Types of Clustering:

i. Exclusive (partitioning)

In this clustering method, Data are grouped in such a way that one data can belong to one cluster only.

Example: K-means

ii. Agglomerative

In this clustering technique, every data is a cluster. The iterative unions between the two nearest clusters reduce the number of clusters.

Example: Hierarchical clustering

2. 4- Clustering Types

- Hierarchical clustering
- K-means clustering
- K-NN (k nearest neighbors)
- Principal Component Analysis
- Singular Value Decomposition
- Independent Component Analysis

a. 4.1 KMeans Clustering

K means it is an iterative clustering algorithm which helps you to find the highest value for every iteration. Initially, the desired number of clusters are selected. In this clustering method, you need to

cluster the data points into k groups. A larger k means smaller groups with more granularity in the same way. A lower k means larger groups with less granularity.

The output of the algorithm is a group of "labels." It assigns data point to one of the k groups. In k-means clustering, each group is defined by creating a centroid for each group. The centroids are like the heart of the cluster, which captures the points closest to them and adds them to the cluster.

Code Example:

from sklearn.cluster import KMeans

kmeans = KMeans(n_clusters=4, max_iter=50)

kmeans.fit(dataframe)

```
For plotting:

import matplotlib.pyplot as plt
%matplotlib inline

plt.figure(figsize=(10, 7))
plt.scatter(df['var1'], df['var2'], c=cluster.labels_)
```

3. EXERCISE (10)

You own a supermarket mall and through membership cards, you have some basic data about your customers like Customer ID, age, gender, annual income and spending score. You want to

understand the customers like who are the target customers so that the sense can be given to the marketing team and plan the strategy accordingly. Solve the given problem by using the dataset given on the GCR.

1. What are the features used in this dataset for customer segmentation?
2. What are the total no of missing values in the dataset?
3. Replace the missing values with mean in case of numerical feature and mode in case of categorical feature.
4. What is the distribution of the 'Age' feature in the dataset?
5. Which feature has the highest correlation with the 'Spending Score (1-100)' feature?
6. What is the optimal number of clusters for customer segmentation according to the Elbow Method?
7. What is the average annual income of customers in the dataset?
8. What is the average spending score of male customers in the dataset?
9. Which cluster has the highest average income and spending score?
10. What is the percentage of customers in Cluster 1?
11. What is the most frequent age group in Cluster 2?
12. What is the average income of customers in Cluster 3?
13. Show the count of value in each cluster.
14. Show customers from each cluster.
15. Make a visualization of the clusters.