**National University of Computer and Emerging Sciences**



**Lab Manual 10**

**CL461-Artificial Intelligence Lab**

|  |  |
| --- | --- |
| Course Instructor | Dr. Mubasher Baig |
| Lab Instructor (s) | Nimra  samiya |
| Section | A |
| Semester | Spring 2021 |

Department of Computer Science

FAST-NU, Lahore, Pakistan

Table of Contents

* **Objectives**

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

* **Task Distribution**

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

* **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.

* **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

* **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.
* **3.3 Types of unsupervised learning**

Unsupervised learning problems further grouped into clustering and association problems.

* **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.

* **3.5 Types of Clustering:**
* **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

* **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

* **4- Clustering Types**
* Hierarchical clustering
* K-means clustering
* K-NN (k nearest neighbors)
* Principal Component Analysis
* Singular Value Decomposition
* Independent Component Analysis
* **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\_)

* **5. Hierarchical Clustering:**

Hierarchical clustering is an algorithm which builds a hierarchy of clusters. It begins with all the data which is assigned to a cluster of their own. Here, two close cluster are going to be in the same cluster. This algorithm ends when there is only one cluster left.

Code Sample:

from sklearn.cluster

import AgglomerativeClustering

cluster = AgglomerativeClustering(n\_clusters=2, affinity='euclidean', linkage='ward')

cluster.fit\_predict(data\_scaled)

For plotting:

import matplotlib.pyplot as plt

%matplotlib inline

plt.figure(figsize=(10, 7))

plt.scatter(df['var1'], df['var2'], c=cluster.labels\_)

* **10 Exercise: (25)**
* **10.1 Problem 1: (Kmeans) (15)**

Online Retail Dataset is attached in the file. You need to load the dataset and preprocess it for missing values and outliers. You need to scale the data for better clusters. You are then required to perform Kmeans clustering on this dataset and try with different number of clusters and visualize it.

* **10.2 Problem 2: (Agglomerative Clustering) (10)**

Online Retail Dataset is attached in the file. You need to load the dataset and preprocess it for missing values and outliers. You need to scale the data for better clusters. You are then required to perform Hierarchical clustering on this dataset and try with different number of clusters and visualize it.

* **11. Submission Instructions:**
* A data file is attached. For Practice Exercise, One has to use this file.
* To make the submission, Create a Jupyter Notebook File (lab10\_rollno.ipynb), create a .zip file along with the data file and submit on the Portal.