**Department of Electrical Engineering**

**Faculty Member:** LE Munadi Sial **Date:** 11-Oct-2023

**Semester:** 7th **Group:**

# CS471 Machine Learning

**Lab 5: K-Means Clustering**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | **PLO4 - CLO4** | **PLO4 -CLO4** | **PLO5 -CLO5** | **PLO8 -CLO6** | **PLO9 -CLO7** |
| **Name** | **Reg. No** | **Viva /Quiz / Lab Performance** | **Analysis of data in Lab Report** | **Modern Tool Usage** | **Ethics** | **Individual and Team Work** |
|  |  | **5 Marks** | **5 Marks** | **5 Marks** | **5 Marks** | **5 Marks** |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

## Introduction

This laboratory exercise is focused on K-means clustering which is a widely used unsupervised learning technique. Clustering is used on unlabeled data to look for interesting groups and patterns.

## Objectives

The following are the main objectives of this lab:

* Implement data structures such as lists and dictionaries in python
* Create, alter and loop through lists
* Use slicing to access range of items in a list
* Utilize various list methods such as append, insert, extend, remove, pop etc
* Create and implement a dictionary
* Create Numpy arrays and perform matrix operations and broadcasting
* Use Scipy for minimization, scarce matrices and iterpolation

## Lab Conduct

* Respect faculty and peers through speech and actions
* The lab faculty will be available to assist the students. In case some aspect of the lab experiment is not understood, the students are advised to seek help from the faculty.
* In the tasks, there are commented lines such as #YOUR CODE STARTS HERE# where you have to provide the code. You must put the code between the #START and #END parts of these commented lines. Do NOT remove the commented lines.
* Use the tab key to provide the indentation in python.
* When you provide the code in the report, keep the font size at 12
* Upon completing the lab, you must delete the manual from the lab computer

**Theory**

K-means clustering is an unsupervised learning technique that is used to find groups, clusters or patterns in unlabeled. As the dataset is not labelled, only the arrangement of the inputs on the feature space are available. In K-means clustering, K number of clusters are set and then the examples are compared to the cluster centroids. The distance of each feature is used as a metric to define which cluster it belongs to. The cluster centroids are iteratively shifted and the examples belonging to them also change. After enough iterations, useful groups in the feature space are obtained. To determine the best number for clusters, a cost function can be calculated for each K number.

A brief summary of the keywords and functions in python is provided below:

**print()** output text on console

**input()** get input from user on console

**range()**  create a sequence of numbers

**len()** gives the number of characters in a string

**if** contains code that executes depending on a logical condition

**else** connects with **if** and **elif**, executes when conditions are not met

**elif** equivalent to **else if**

**while** loops code as long as a condition is true

**for** loops code through a sequence of items in an iterable object

**break** exit loop immediately

**continue**  jump to the next iteration of the loop

**def** used to define a function

**append(I)** append item I to the end of the list

**insert(i, I)** insert item I at i position of the list

**extend(L)** extend/concatenate a second list L

**remove(I)** remove a specified item I from a list

**pop(i)** remove item at specific index i in the list

**count(I)** return total number of a specific item I from a list

**index(I)** return index of first occurrence of a specific item I

**reverse** reverse the items of the list

For this lab, you will be provided with some dataset files in .csv format which you will need for the tasks. Additionally, for the final task, you will need to arrange your own dataset by downloading it from the internet. You will need to make use of numpy, pandas and matplotlib libraries for the given tasks.

**Lab Task 1 – 2-Means Clustering \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

In this task, use the provided dataset. Write the code which performs clustering of the dataset into 2 clusters. The pseudocode for the clustering algorithm is provided as follows:

**specify K number of centroids**

**randomly initialize K number of centroids u**

**for j = 1:epochs**

**for i = 1:m**

**c(i) = index of closest cluster to training example**

**for k = 1:K**

**u(k) = mean of all training examples indexed to k**

**plot of x1 and x2 clusters**

To determine the index c(i), you will need to write a function that calculates the Euclidean distance between the points in the feature space. This function will be used to find the closest centroid from each training example. After determining the indexes, the cluster centroids themselves are updated by taking the average of the x values. For k-th cluster, the training examples with index k will be averaged. This completes one iteration of clustering after which a scatter plot is made. The iterations are repeated until interesting groups are obtained in the plots.

Due to the initial randomization of cluster centroids, you may have to repeat the clustering a few times. Also, ensure the random centroids are from within the domain of the feature space.

Your code must generate scatter plots showing the clusters at each iteration. The input values must be colored and marked according to the cluster to which they belong at each iteration. The cluster centroids must also be shown Provide all of the codes and screenshots of the final output. You must include plots of at 3 iterations showing the progress of your clustering algorithm.

***### TASK 1 CODE STARTS HERE ###***

*### TASK 1 CODE ENDS HERE ###*

***### TASK 1 OUTPUT SCREENSHOT STARTS HERE ###***

*### TASK 1 OUTPUT SCREENSHOT ENDS HERE ###*

**Lab Task 2 – K-Means Clustering \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Repeat task 1, however, set value of k = 3, 4 and 5. For every k value, generate at least three plots.

***### TASK 2, 3 CLUSTER PLOTS START HERE ###***

*### TASK 2, 3 CLUSTER PLOTS END HERE ###*

***### TASK 2, 4 CLUSTER PLOTS START HERE ###***

*### TASK 2, 4 CLUSTER PLOTS END HERE ###*

***### TASK 2, 5 CLUSTER PLOTS START HERE ###***

*### TASK 2, 5 CLUSTER PLOTS END HERE ###*

**Lab Task 3 – Cost Function \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Load the given dataset into the python program for this task. In this task, you will modify your code so that it performs clustering from

k = 2, 3, 4… 10.

For each k value, perform 20 iterations (epochs) of centroid update before moving to the next value of k. Additionally, at the last iteration, determine the cost for that K-value:

**for K = 2:10**

**randomly initialize K number of centroids u**

**for j = 1:epochs**

**for i = 1:m**

**compute c(i)**

**for k = 1:K**

**compute u(k)**

**plot of x1 and x2 clusters**

**compute cost for current K value**

**plot of cost and K**

Store the costs for each k in a list. After the last iteration of the last cluster, make a plot of k vs. cost.

***### TASK 3 CODE STARTS HERE ###***

*### TASK 3 CODE ENDS HERE ###*

***### TASK 3 CLUSTERS START HERE ###***

*### TASK 3 CLUSTERS END HERE ###*

***### TASK 3 PLOT(k, cost) STARTS HERE ###***

*### TASK 3 PLOT(k, cost) ENDS HERE ###*

**Lab Task 4 – Your Own Dataset \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_**

Download your own CSV dataset from the internet (e.g. Kaggle). Your dataset must have at least 500 rows and at least 4 feature columns. Perform clustering of your dataset, showcase the plots and provide explanation.

***### TASK 4 CODE STARTS HERE ###***

*### TASK 4 CODE ENDS HERE ###*

***### TASK 4 PLOTS START HERE ###***

*### TASK 4 PLOTS END HERE ###*