CS 6301 - Machine Learning Lab - Week 8

Date: 06.10.2023

TITLE

IMPLEMENTATION OF K-MEANS CLUSTERING

TASK

1. You are given the following data structures for K-means clustering:

   data: A list of data point three data points, where each data point is four-dimensional
   Example: data = [[0.34, -0.2, 1.13, 4.3], [5.1, -12.6, -7.0, 1.9], [-15.7, 0.06, -7.1, 11.2]]
   centroids: A dictionary of centroids, where the keys are strings (centroid names) and the values are lists (centroid locations)
   Example: centroids = {"centroid1": [1.1, 0.2, -3.1, -0.4], "centroid2": [9.3, 6.1, -4.7, 0.18]}

   You should **NOT** change the names of the centroids when you later update their locations.

   Write python code to implement K-means clustering and verify the results with manual calculation.

2. Write a program to cluster a set of points using K-means. Consider, K=3, clusters. Consider Euclidean distance as the distance measure. Randomly initialize a cluster mean as one of the data points. Iterate for 10 iterations. After iterations are over, print the final cluster means for each of the clusters. Use the ground truth cluster label present in the data set to compute and print the Jacquard distance of the obtained clusters with the ground truth clusters for each of the three clusters.

   Data Set Description: Data Filename: data4_19.csv The data set contains 150 data points, there are three clusters where each cluster refers to a type of iris plant. The first four columns represent the attributes listed below. Note that only the first four columns should be used as attributes. The last column is the ground truth cluster name and is to be used for evaluating the cluster quality. 1. sepal length in cm 2. sepal width in cm 3. petal length in cm 4. petal width in cm 5. Ground truth cluster name: -- Iris Setosa -- Iris Versicolour -- Iris Virginica