School of Computer Science Engineering and Technology

Course- BTech
Course Code- 301
Year- 2022
Date- 10-04-2022

Type- Core
Course Name-AIML
Semester- Even
Batch- IV Sem Spl

Lab Assignment No. 11.1

Objective: To understand K-means clustering by implementing it for the task of news articles clustering.

1. Dataset: 20 Newsgroups is a widely used benchmark dataset that consists of around 20,000 news articles grouped into 20 categories. Load and Pre-process the dataset as follows: (20)

    a. **Load**: Import and Use **fetch_20newsgroups()** function from **sklearn.datasets** package to load the data. Use only following categories of documents: "alt.atheism", "talk.religion.misc", "comp.graphics", "sci.space".
    b. **Vectorize:** Use **TfidfVectorizer** class to vectorize all the documents using TFIDF vector space model.

2. Define a function **get_initial_centroids(data, k)** that returns randomly chosen 'K' (=3, say) initial centroids. (10)

3. After initialization, the k-means algorithm iterates between the following two steps: (30)
    a. Assign each data point to the closest centroid. For this define a function **assign_clusters(data, centroids)** that assigns each data point to the nearest centroid.
    b. Revise centroids as the mean of the assigned data points. For this define a function **revise_centroids(data, k, cluster_assignment)** that computes and returns new centroids by taking current clusters.

4. Define a function **compute_heterogeneity(data, k, centroids, cluster_assignment)** that computes heterogeniety as the performance assessment measure that checks the heterogeneity at each iteration of K-means. (10)

5. Finally, combine all the above functions into a single function **kmeans(data, k, initial_centroids, maxiter, record_heterogeneity=None, verbose=False)** that returns final centroids and cluster assignment. (10)

6. Plotting the heterogeneity: Define a function **plot_heterogeneity(heterogeneity, k)** that plots heterogeneity with respect to number of iterations. (5)

7. Finally, apply k-means on the vectorized version of news articles dataset. You can choose K=3 to start with. (5)

Additional Fun: How to choose K? You can use Elbow method for the same. Just plot Heterogeniety for different K values and detect the elbow point in the plot.