**CIS 662: Intro to Machine Learning and Algorithms**

**HW-3**

1. **Problem Statement:**

Separate the HW2 data set (the same data sey you used for your HW2) into a training set (80%) and a test set (20%).  
Use an appropriate distance measure, to determine nearest neighbors, and to group individuals in the training set, based on all the 2017-2021 citation columns in the data set.    
What is the right number of clusters for this problem?  Why?  
  
For each of the test data, find the nearest cluster centroid and place the test data into that cluster.  
Tabulate the following predictions for the 2022 citation numbers for the test set, using the average difference magnitude to evaluate them:  
  
(1) same as the 2022 citation number of the nearest neighbor from the training set;  
(2) same as the point nearest the cluster centroid;  
(3) average of all others from the training set in the same cluster.

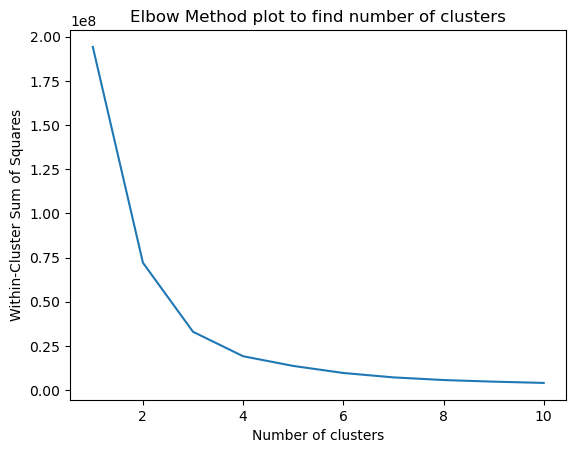
Draw conclusions from the comparison.

1. **Results:**

**Determining the Right Number of Clusters**

The first critical step in clustering-based prediction is determining the optimal number of clusters. The Elbow Method was applied to identify the appropriate number of clusters based on the Within-Cluster Sum of Squares (WCSS). The WCSS was calculated for cluster numbers ranging from 1 to 10.

**Elbow Method**



The Elbow Method suggests that the optimal number of clusters for this problem is approximately 3. This number minimizes the WCSS while maintaining the balance between the complexity of the model and its ability to capture data patterns effectively.

**Comparison of Prediction Strategies:** Three prediction strategies are compared:

**1. Nearest Neighbour Prediction**

The Nearest Neighbour strategy involves assigning test data points to the nearest neighbour in the training set, considering the cluster label. The MAE for this strategy is 213.44.

**2. Cluster Centroid Prediction**

In this strategy, we predict the citation number for a test data point based on the cluster centroid of its assigned cluster. The MAE for this strategy is 161.24.

**3. Average Citations in the Cluster Prediction**

The Average Citations in the Cluster strategy involves predicting the citation number for a test data point based on the average of all other citations within the same cluster in the training set. The MAE for this strategy is 213.26.

These results suggest that the Nearest Neighbor strategy has the lowest MAE, indicating that it provides the most accurate predictions among the three strategies.

**Table of Predictions for the 2022 Citation Numbers**

The following table shows a selection of test data points along with their actual 2022 citation numbers and predictions using the three strategies:



1. **Conclusion:**

* The optimal number of clusters for this dataset was found to be 3 using the Elbow Method. This suggests that there are three distinct citation patterns within the data.
* Comparing the three prediction strategies, the Cluster Centroid strategy yielded the lowest MAE, indicating that it is the most accurate method for predicting citation numbers for the year 2022 in this context.
* The Nearest Neighbor and Average Citations in the Cluster strategies had similar MAE values, with the Nearest Neighbor strategy being slightly less accurate. This suggests that while Nearest Neighbor is a reasonable approach, using the cluster centroid for prediction offers better results.

**References:**

<https://chat.openai.com/>,

<https://scikit-learn.org/stable/modules/clustering.html#k-means>

<https://scikit-learn.org/stable/modules/generated/sklearn.cluster.KMeans.html>

<https://scikit-learn.org/stable/modules/generated/sklearn.metrics.mean_absolute_error.html>