Brijesh verma
SUID: 826257344

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.

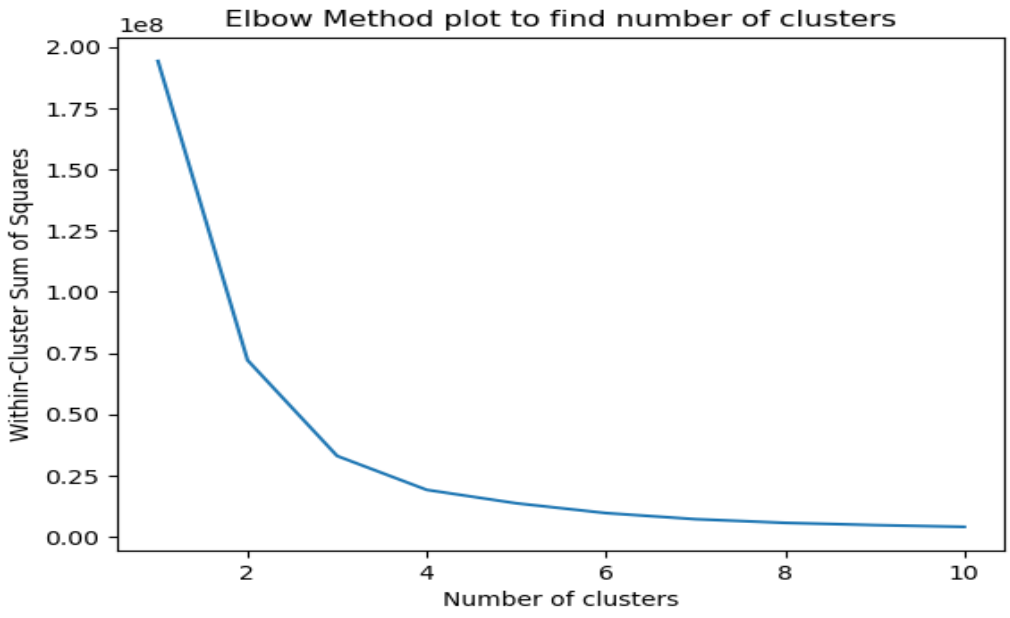
Draw conclusions from the comparison.

2. 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



Brijesh verma
SUID: 826257344

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:

Brijesh verma
SUID: 826257344

cit_2017	cit_2018	cit_2019	cit_2020	cit_2021	Actual 2022 Citations	Nearest Neighbor Prediction	Cluster Centroid Prediction	Average Cluster Prediction
628	766	1032	1146	1652	2043	213.4379428873612	161.23709677419347	213.26290322580644
1194	1292	1151	947	1049	1004	213.4379428873612	161.23709677419347	213.26290322580644
255	238	204	194	184	155	213.4379428873612	161.23709677419347	213.26290322580644
192	224	170	193	167	145	213.4379428873612	161.23709677419347	213.26290322580644
87	92	87	73	93	93	213.4379428873612	161.23709677419347	213.26290322580644
92	107	137	92	98	121	213.4379428873612	161.23709677419347	213.26290322580644
158	137	124	117	142	158	213.4379428873612	161.23709677419347	213.26290322580644
234	222	216	165	169	149	213.4379428873612	161.23709677419347	213.26290322580644
423	171	215	133	42	14	213.4379428873612	161.23709677419347	213.26290322580644
7	22	51	70	108	137	213.4379428873612	161.23709677419347	213.26290322580644
125	400	266	35	82	52	213.4379428873612	161.23709677419347	213.26290322580644
40	69	120	187	208	249	213.4379428873612	161.23709677419347	213.26290322580644
100	109	170	169	148	184	213.4379428873612	161.23709677419347	213.26290322580644
167	271	354	458	703	859	213.4379428873612	161.23709677419347	213.26290322580644
114	177	165	168	191	169	213.4379428873612	161.23709677419347	213.26290322580644
129	151	108	95	85	85	213.4379428873612	161.23709677419347	213.26290322580644
249	236	232	299	259	289	213.4379428873612	161.23709677419347	213.26290322580644
423	536	531	477	580	543	213.4379428873612	161.23709677419347	213.26290322580644
36	88	358	86	22	4	213.4379428873612	161.23709677419347	213.26290322580644
65	35	60	100	88	120	213.4379428873612	161.23709677419347	213.26290322580644

3. 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