**CIDM 6355 Data Mining Methods HW3**

(60 points; Due 11:59 PM Central Time, March 18, 2024)

Requirements: This homework is open book, open slides, and open notes, but no collaboration or discussion is permitted before the due time. Any questions about the homework should be directed to the instructor. You must adhere to the instructions, completing all questions and deliverables. This is an individual assignment, so sharing your processes, scripts, screenshots, or answers with others constitutes cheating and will be reported. Additionally, ensure your answers meet the required format to avoid point deductions. Screenshots without date and time will receive a maximum of 50% of points. Please acknowledge your understanding and agreement to these requirements by typing your name below.

Type Your name: Mehnaz Afrose

Instruction: Please compile all the deliverables with the required format as below.

1. Deliverable 1 (Step 1): Please write down the average for all the five attributes (round them the third decimal place). All these numbers below are the overall centroid for all 325 cities. [5 points]

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Attributes | Cost\_living | Jobs | Climate | Health\_Care | Recreation |
| Average | 51.910 | 51.023 | 52.035 | 47.865 | 50.227 |

1. Deliverable 2 (Step 4.5): Take a screenshot of your Exampleset (Screenshot 1) [3 points]

A screenshot of a computer

Description automatically generated

1. Deliverable 3 (Step 4.8): based on the results in 4.5-4.8, please discuss the characteristics in each cluster and find an appropriate name for each cluster. For example, Cluster 0 includes 128 cities such as New Orleans, LA and Long Island, NY have highest scores in job opportunities, climate, healthcare, and recreation. However, this group of cities have quite high living cost. We can name this group of cities …….. [9 points: 3 points for each cluster]

Cluster 0 includes 97 cities such as Panama City, FL and Lafayette, LA have highest scores in cost of living and in climate and lowest score in recreation. Medium scores in health care and job opportunities. We can name this group of cities "Premium Urban Havens", which highlights the upscale nature of the cities within the cluster, emphasizing their higher cost of living.

Cluster 1 includes 101 cities such as Jamestown, NY and Benton Harbor, MI have lowest scores in climate, job opportunities, and health care facilities and have medium scores in living cost and recreation. We can name this group of cities "Rustic Enclaves" which reflect the lower scores in various aspects such as climate and job opportunities while acknowledging the affordability and potential recreational opportunities that the cities offer.

Cluster 2 includes 127 cities such as New Orleans, LA and Long Island, NY have highest scores in job opportunities, health care, and recreation. Also, these cities have the lowest score in living cost and medium score in climate. We can name this group of cities "Dynamic Metropolitan Centers." which emphasizes the bustling and lively nature of the cities within the cluster, highlighting their excellent job opportunities, healthcare facilities, and recreational options.

A screenshot of a computer

Description automatically generated

1. Deliverable 4 (Step 6.2): Take a screenshot of your Result History page (Screenshot 2) [3 points]

A screenshot of a computer

Description automatically generated

1. Deliverable 5 (Step 6.2): Please answer all the question in this deliverable [8 points]

* Based on the table above, when k increases, what happen to Avg. within centroid distance (increasing or decreasing)? [2 points]

Answer: decreasing.

* What about Davies Bouldin Index when k increases? [2 points]

Answer: When the value of k increases from 2 to 3, the value of Davies Bouldin Index increases. Then from k=4 to k=7, this value decreases. After that, at k=8, it increases, at k=9 the value decreases, at k=10 the value increases, from k=11 to k=12, the value decreases, at k=13 the value increases, at k=14 decreases, k=15 increases.

So, the value of Davies Bouldin Index is inconsistent with the value of k.

* Imagine an extreme case, when k=325, what would Avg. within centroid distance be? [2 points]

Answer: 0.000

* What potential problem will we encounter if we only use Avg. within centroid distance as the main criterion for evaluating clustering models? [2 points]

Answer: Relying solely on average within centroid distance for evaluating clustering models may overlook important factors like outliers, unequal cluster sizes, and cluster structure, leading to inaccurate assessments of clustering quality.

1. Deliverable 6 (Step 7.1): Draw an elbow chart using either average within centroid distance or DBI for k=2-15. Take a screenshot of your elbow chart with date and time (Screenshot 3). Observe your elbow chart and discuss which k is the best and why. [5 points: 3 points for screenshot and 2 points for your discussion]

A screenshot of a computer

Description automatically generated

Among all the k values, 5 appears to be the best choice. This is because, from the elbow chart, we can see that as the value of k increases, the distance decreases. While the decrease in distance from k=2 to k=3 is notable, at k=5, the rate of decrease slows down significantly. After that, we observe a consistent slower decrease in distance as k increases. Therefore, we can conclude that k=5 is the best value.

1. Deliverable 7: Use k=3 to include all the four performance operators in your process. Take a screenshot of the description view of Cluster Density Performance and Item Distribution Performance (two screenshots in total: Screenshot 4 and Screenshot 5) and then briefly discuss each result. [7 points: 4 points for two screenshots and 3 points for your discussion]

Cluster Density Performance:

A screenshot of a computer

Description automatically generated

The result shows the average distances between data points within each cluster. Lower distances mean data points are closer together, indicating higher density. Cluster 0 has the lowest average distance, showing the highest density, followed by Cluster 1 with slightly higher density. Cluster 2 has the highest average distance, suggesting lower density. These results help understand how tightly clustered the data points are within each group.

Item Distribution Performance:

A screenshot of a computer

Description automatically generated

The Item Distribution Performance result of 0.338 suggests that the examples are reasonably distributed among the clusters in the flat clustering model. This value indicates the degree of balance in cluster sizes, with lower values indicating more even distribution and higher values suggesting potential imbalance or concentration of examples within certain clusters.

1. Deliverable R1: take a screenshot of the result after running the script in Line 19 with date and time (Screenshot 6) and time briefly interpret the result. [5 points: 3 points for screenshot and 2 points for your interpretation]

A screenshot of a computer

Description automatically generated

In the K-means clustering results in R, we have three clusters with sizes of 92, 128, and 105. Cluster means showing the average values for each feature within each cluster. The clustering vector indicates which cluster each data point belongs to. The within-cluster sum of squares measures the variability within each cluster, with higher values indicating greater dispersion of data points. The percentage of between-cluster sum of squares to total sum of squares is 40.6%, indicating that 40.6% of the total variance in the data is explained by the clustering.

1. Deliverable R2: take a screenshot of the result after running the script in Line 24 with date and time (Screenshot 7) and time briefly interpret the result. [5 points: 3 points for screenshot and 2 points for your interpretation]



This data frame presents the size of each cluster (CityCluster.size) alongside the average values for each feature within each cluster. For instance:

Cluster 1 has a size of 92 and average values of 41.04761 for Cost of Living, 33.73391 for Jobs, 23.47891 for Climate, 45.55793 for Health Care, and 50.02924 for Recreation.

Cluster 2 includes 128 data points with average values of 40.13391 for Cost of Living, 77.22789 for Jobs, 68.54938 for Climate, 61.29539 for Health Care, and 73.56758 for Recreation.

Cluster 3 has 105 data points, with average values of 75.78324 for Cost of Living, 34.22686 for Jobs, 56.92467 for Climate, 33.51457 for Health Care, and 21.94552 for Recreation.

These values offer insight into the distribution and characteristics of the data within each cluster.

1. Deliverable R3: take a screenshot of the result after running the script in Line 28 with date (Screenshot 8) and time briefly interpret the result. [5 points: 3 points for screenshot and 2 points for your interpretation]

A screenshot of a computer

Description automatically generated

This is a new data frame created to contain cluster ID and the five attributes of each observation.

1. Deliverable R4: Compare the clustering result for each observation in R (which is saved in CityRecords.csv) and that in RapidMiner (k=3 only). Compare the two clustering results and answer the question: Are the two clustering results in R and RM the same or not? Why? You may follow the instruction in the next slide and take a screenshot of your PivotTable with date and time to support your answer (Screenshot 9). Attention: you cannot just simply compare the cluster name because R and RM may label each cluster differently. For example, New Orleans, LA is labeled as cluster\_0 in RM, but Cluster 3 in R, but cluster\_0 in RM might be the same with Cluster 3 in R. [5 points: 3 points for screenshot and 2 points for your answer]

A screenshot of a computer

Description automatically generated

From the pivot table, we can observe that the two clustering results in R and RM are different. If they were the same, we would have obtained a rank zero diagonal matrix in the pivot table. This implies that all the cities of one cluster in RM would belong to a single cluster in R. Instead, we can see that out of a total of 92 cities in cluster 1 of R, there are 14 cities from cluster 0, 49 cities from cluster 1, and 29 cities from cluster 2 of RM. Dissimilarities also belong to other rows. So, we can conclude that the two clustering results in R and RM are not the same.