**Chelsea Murray – 113481968 – Homework 4 – Data Mining Fall 2024**

1. The chat transcript can be found here: <https://chatgpt.com/share/6729a718-d024-800f-a2d2-42bd08fe6cf0>

|  |  |
| --- | --- |
| User: | Assume I am looking for information at the intelligence level of a professional in the field of data mining. Tell me about three real-world companies/organizations from three distinct sectors—software technology, telecommunications, and transportation—that have implemented data clustering algorithms in their businesses. Include detailed descriptions of how those companies/organizations have applied data clustering algorithms in their operations. |
| ChatGPT: | Certainly! Here’s an overview of how companies in software technology, telecommunications, and transportation apply data clustering algorithms to enhance their operations.  **1. Software Technology: Google**  **Application of Clustering:** Google uses data clustering extensively across its services, particularly in organizing search results, personalizing recommendations, and improving user experience. One significant area is **image recognition and search**. Google's algorithms cluster images with similar visual content, allowing the company to categorize vast amounts of images, videos, and other visual data. This clustering supports Google Photos in organizing user photos by people, places, or activities.  **How It Works:** By applying clustering algorithms (such as k-means and hierarchical clustering), Google groups similar visual data based on pixel patterns, color distributions, or shape similarities. The clustered data helps the search algorithm quickly identify relevant images for users' search queries, improving the speed and accuracy of visual search results. Moreover, clustering aids in refining Google’s recommendation systems for YouTube and Play Store by grouping users based on shared preferences, enabling the platform to suggest more personalized content.  **2. Telecommunications: AT&T**  **Application of Clustering:** AT&T uses data clustering to optimize its network performance and improve customer experience. Clustering algorithms help AT&T analyze patterns in user data across different geographic regions and identify areas experiencing high demand or recurring issues. By grouping these regions based on customer usage patterns and signal strength, AT&T can allocate network resources more effectively.  **How It Works:** Clustering algorithms such as DBSCAN (Density-Based Spatial Clustering of Applications with Noise) and k-means are employed to analyze location-based data, call drop frequencies, and network traffic. These clusters reveal zones of high usage or network congestion, helping AT&T to prioritize infrastructure upgrades or reconfigure network settings dynamically. For instance, AT&T might detect clusters of high data traffic in metropolitan areas during peak hours, prompting a temporary reallocation of bandwidth to prevent network slowdowns.  **3. Transportation: Uber**  **Application of Clustering:** Uber applies clustering algorithms for **demand forecasting** and **dynamic pricing**. By clustering locations based on demand patterns, Uber can identify “hot zones” where ride requests are likely to spike, especially during events, rush hours, or holidays. Clustering is also used to optimize driver availability in high-demand areas.  **How It Works:** Uber leverages k-means clustering and density-based clustering to group similar trip data by factors like pickup locations, drop-off points, and times of high demand. This data helps Uber forecast demand patterns in specific clusters, adjusting pricing through “surge pricing” when demand in these areas exceeds driver availability. Clustering also assists Uber in mapping out ideal “standby” zones for drivers to minimize wait times and optimize routes, enhancing both driver efficiency and customer satisfaction. |

[response once u are not sleep deprived]

1. Given the table:



2.1) Here is the application of bisecting K-means to cluster the data using K=4, number of trials=2, and choosing the cluster with the largest SSE for splitting at each step.

**Cluster Step 1: Initialize cluster containing all points.**

The first cluster contains every point in the table.

|  |  |
| --- | --- |
| **Cluster** | **Contents** |
| A | {(4.38, -15.96), (0.69, 2.5), (4.69, 36.75), (4.7, -7.83), (4.38, 34.43), (0.47, -2.5), (4.07, 33.8), (5.01, -16.59)} |

**Cluster Step 2: Split into two on the “all points” cluster using K-means and find the bisection with the smallest SSE.**

At the second step, the cluster is bisected twice (according to the number of trials specified). The first trial will use the first two points as centroids; the second trial will use the last two points as centroids.

*Trial 1: K-Means Clustering around (4.38, -15.96) [Centroid 1] and (0.69, 2.5) [Centroid 2]*

*Trial 1 Round 1: (4.38, -15.96) [Centroid 1] and (0.69, 2.5) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | 0 | sqrt((0.69-4.38)2+(2.5-(-15.96))2) =sqrt(13.6161+340.7716) =18.8252 | Centroid 1 |
| (0.69, 2.5) | sqrt((4.38-0.69)2+(-15.96-2.5)2) =sqrt(13.6161+340.7716) =18.8252 | 0 | Centroid 2 |
| (4.69, 36.75) | sqrt((4.38-4.69)2+(-15.96-36.75)2) =sqrt(0.0961+2778.3441) = 52.7109 | sqrt((0.69-4.69)2+(2.5-36.75)2) =sqrt(16+1173.0625) = 34.4828 | Centroid 2 |
| (4.7, -7.83) | sqrt((4.38-4.7)2+(-15.96-(-7.83))2) =sqrt(0.1024+66.0969) = 8.1363 | sqrt((0.69-4.7)2+(2.5-(-7.83))2) =sqrt(16.0801+106.7089) = 11.0810 | Centroid 1 |
| (4.38, 34.43) | sqrt((4.38-4.38)2+(-15.96-34.43)2) =sqrt(0+2539.1521) = 50.39 | sqrt((0.69-4.38)2+(2.5-34.43)2) =sqrt(13.6161+1019.5249) = 32.1425 | Centroid 2 |
| (0.47, -2.5) | sqrt((4.38-0.47)2+(-15.96-(-2.5))2) =sqrt(15.2881+181.1716) = 14.0164 | sqrt((0.69-0.47)2+(2.5-(-2.5))2) =sqrt(0.0484+25) = 5.0048 | Centroid 2 |
| (4.07, 33.8) | sqrt((4.38-4.07)2+(-15.96-33.8)2) =sqrt(0.0961+2476.0576) = 49.7610 | sqrt((0.69-4.07)2+(2.5-33.8)2) =sqrt(11.4244+979.69) = 31.4820 | Centroid 2 |
| (5.01, -16.59) | sqrt((4.38-5.01)2+(-15.96-(-16.59))2) =sqrt(0.3969+0.3969) = 0.8910 | sqrt((0.69-5.01)2+(2.5-(-16.59))2) =sqrt(18.6624+364.4281) = 19.5727 | Centroid 1 |

New centroids = (mean(x), mean(y)) for each set: (4.6967, -13.46) and (2.86, 20.996)

*Trial 1 Round 2: (4.6967, -13.46) [Centroid 1] and (2.86, 20.996) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((4.6967-4.38)2+(-13.46-(-15.96))2) =sqrt(0.1003+6.25) = 2.5200 | sqrt((2.86-4.38)2+(20.996-(-15.96))2) =sqrt(2.3104+ 1365.7459) = 36.9872 | Centroid 1 (unchanged) |
| (0.69, 2.5) | sqrt((4.6967-0.69)2+(-13.46-2.5)2) =sqrt(16.0536+254.7216) = 16.4552 | sqrt((2.86-0.69)2+(20.996-2.5)2) =sqrt(4.7089+342.1020) = 18.6229 | Centroid 1 (changed) |
| (4.69, 36.75) | sqrt((4.6967-4.69)2+(-13.46-36.75)2) =sqrt(0.00004+2521.0441) = 50.2100 | sqrt((2.86-4.69)2+(20.996-36.75)2) =sqrt(3.3489+248.1885) = 15.8599 | Centroid 2 (unchanged) |
| (4.7, -7.83) | sqrt((4.6967-4.7)2+(-13.46-(-7.83))2) =sqrt(0.00001+31.6969) = 5.6300 | sqrt((2.86-4.7)2+(20.996-(-7.83))2) =sqrt(3.3856+830.9383) = 28.8847 | Centroid 1 (unchanged) |
| (4.38, 34.43) | sqrt((4.6967-4.38)2+(-13.46-34.43)2) =sqrt(0.1003+2293.4521) = 47.8910 | sqrt((2.86-4.38)2+(20.996-34.43)2) =sqrt(2.3104+180.4724) = 13.5197 | Centroid 2 (unchanged) |
| (0.47, -2.5) | sqrt((4.6967-0.47)2+(-13.46-(-2.5))2) =sqrt(17.8650+120.1216) = 11.7468 | sqrt((2.86-0.47)2+(20.996-(-2.5))2) =sqrt(5.7121+552.0620) = 23.6172 | Centroid 1 (changed) |
| (4.07, 33.8) | sqrt((4.6967-4.07)2+(-13.46-33.8)2) =sqrt(0.3928+2233.5076) = 47.2642 | sqrt((2.86-4.07)2+(20.996-33.8)2) =sqrt(1.4641+163.9424) = 12.8610 | Centroid 2 (unchanged) |
| (5.01, -16.59) | sqrt((4.6967-5.01)2+(-13.46-(-16.59))2) =sqrt(0.0982+9.7969) = 3.1456 | sqrt((2.86-5.01)2+(20.996-(-16.59))2) =sqrt(4.6225+1412.7074) = 37.6474 | Centroid 1 (unchanged) |

New centroids = (mean(x), mean(y)) for each set: (3.05, -8.076) and (4.38, 34.9933)

*Trial 1 Round 3: (3.05, -8.076) [Centroid 1] and (4.38, 34.9933) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((3.05-4.38)2+(-8.076-(-15.96))2) =sqrt(1.7689+62.1575) = 7.9954 | sqrt((4.38-4.38)2+(34.9933-(-15.96))2) =sqrt(0+2596.2388) = 50.9533 | Centroid 1 (unchanged) |
| (0.69, 2.5) | sqrt((3.05-0.69)2+(-8.076-2.5)2) =sqrt(5.5696+111.8518) = 10.8361 | sqrt((4.38-0.69)2+(34.9933-2.5)2) =sqrt(13.6161+1055.8145) = 32.7022 | Centroid 1 (unchanged) |
| (4.69, 36.75) | sqrt((3.05-4.69)2+(-8.076-36.75)2) =sqrt(2.6896+2009.3703) = 44.8560 | sqrt((4.38-4.69)2+(34.9933-36.75)2) =sqrt(0.0961+3.0860) = 1.7838 | Centroid 2 (unchanged) |
| (4.7, -7.83) | sqrt((3.05-4.7)2+(-8.076-(-7.83))2) =sqrt(2.7225+0.0605) = 1.6682 | sqrt((4.38-4.7)2+(34.9933-(-7.83))2) =sqrt(0.1024+1833.8350) = 42.8245 | Centroid 1 (unchanged) |
| (4.38, 34.43) | sqrt((3.05-4.38)2+(-8.076-34.43)2) =sqrt(1.7689+1806.7600) = 42.5268 | sqrt((4.38-4.38)2+(34.9933-34.43)2) =sqrt(0+0.3173) = 0.5633 | Centroid 2 (unchanged) |
| (0.47, -2.5) | sqrt((3.05-0.47)2+(-8.076-(-2.5))2) =sqrt(6.6564+31.0918) = 6.1440 | sqrt((4.38-0.47)2+(34.9933-(-2.5))2) =sqrt(15.2881+1405.7475) = 37.6966 | Centroid 1 (unchanged) |
| (4.07, 33.8) | sqrt((3.05-4.07)2+(-8.076-33.8)2) =sqrt(1.0404+1753.5994) = 41.8884 | sqrt((4.38-4.07)2+(34.9933-33.8)2) =sqrt(0.0961+1.4240) = 1.5201 | Centroid 2 (unchanged) |
| (5.01, -16.59) | sqrt((3.05-5.01)2+(-8.076-(-16.59))2) =sqrt(3.8416+72.4882) = 8.7367 | sqrt((4.38-5.01)2+(34.9933-(-16.59))2) =sqrt(0.3969+2660.8368) = 51.5871 | Centroid 1 (unchanged) |

Since none of the points changed clusters, the centroids do not adjust and this is the final split. SSE can be found by squaring the Euclidean distances from each point to the respective centroid and then adding them together, resulting in the following bisection:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AA | (3.05, -8.076) | {(4.38, -15.96), (0.69, 2.5), (4.7, -7.83), (0.47, -2.5), (5.01, -16.59)} | 298.2090 |
| AB | (4.38, 34.9933) | {(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)} | 5.8100 |
| Total SSE | | | 304.0190 |

*SSE calculations*

Cluster AA: 7.99542+10.83612+1.66822+6.14402+8.73672=298.2090  
Cluster AB: 1.78382+0.56332+1.52012=5.8100

*Trial 2: K-Means Clustering around (4.07, 33.8) [Centroid 1] and (5.01, -16.59) [Centroid 2]*

*Trial 2 Round 1: (4.07, 33.8) [Centroid 1] and (5.01, -16.59) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((4.07-4.38)2+(33.8-(-15.96))2) =sqrt(0.0961+2476.0576) = 49.7610 | sqrt((5.01-4.38)2+(-16.59-(-15.96))2) =sqrt(0.3969+0.3969) = 0.8910 | Centroid 2 |
| (0.69, 2.5) | sqrt((4.07-0.69)2+(33.8-2.5)2) =sqrt(11.4244+979.69) = 31.4820 | sqrt((5.01-0.69)2+(-16.59-2.5)2) =sqrt(18.6624+364.4281) = 19.5727 | Centroid 2 |
| (4.69, 36.75) | sqrt((4.07-4.69)2+(33.8-36.75)2) =sqrt(0.3844+8.7025) = 3.0144 | sqrt((5.01-4.69)2+(-16.59-36.75)2) =sqrt(0.1024+2879.3956) = 53.6610 | Centroid 1 |
| (4.7, -7.83) | sqrt((4.07-4.7)2+(33.8-(-7.83))2) =sqrt(0.3969+1733.0569) = 41.6348 | sqrt((5.01-4.7)2+(-16.59-(-7.83))2) =sqrt(0.0961+596.3364) = 24.4220 | Centroid 2 |
| (4.38, 34.43) | sqrt((4.07-4.38)2+(33.8-34.43)2) =sqrt(0.0961+0.3969) = 0.7021 | sqrt((5.01-4.38)2+(-16.59-34.43)2) =sqrt(0.3969+2603.0404) = 51.0239 | Centroid 1 |
| (0.47, -2.5) | sqrt((4.07-0.47)2+(33.8-(-2.5))2) =sqrt(12.96+1317.69) = 36.4781 | sqrt((5.01-0.47)2+(-16.59-(-2.5))2) =sqrt(20.6116+198.5281) = 14.8033 | Centroid 2 |
| (4.07, 33.8) | 0 | sqrt((5.01-4.07)2+(-16.59-33.8)2) =sqrt(0.8836+2539.1521) = 50.3988 | Centroid 1 |
| (5.01, -16.59) | sqrt((4.07-5.01)2+(33.8-(-16.59))2) =sqrt(0.8836+2539.1521) = 50.3988 | 0 | Centroid 2 |

In Trial 2, the same cluster pattern that Trial 1 ended on surfaces immediately. Therefore, the new centroids and resulting clusters and SSE are already known, as stated above, although Centroids 1 and 2 are swapped. For the sake of showing work, I will include this conclusion:

New centroids = (mean(x), mean(y)) for each set: (4.38, 34.9933) and (3.05, -8.076)

*Trial 2 Round 2: (4.38, 34.9933) [Centroid 1] and (3.05, -8.076) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((4.38-4.38)2+(34.9933-(-15.96))2) =sqrt(0+2596.2388) = 50.9533 | sqrt((3.05-4.38)2+(-8.076-(-15.96))2) =sqrt(1.7689+62.1575) = 7.9954 | Centroid 2 (unchanged) |
| (0.69, 2.5) | sqrt((4.38-0.69)2+(34.9933-2.5)2) =sqrt(13.6161+1055.8145) = 32.7022 | sqrt((3.05-0.69)2+(-8.076-2.5)2) =sqrt(5.5696+111.8518) = 10.8361 | Centroid 2 (unchanged) |
| (4.69, 36.75) | sqrt((4.38-4.69)2+(34.9933-36.75)2) =sqrt(0.0961+3.0860) = 1.7838 | sqrt((3.05-4.69)2+(-8.076-36.75)2) =sqrt(2.6896+2009.3703) = 44.8560 | Centroid 1 (unchanged) |
| (4.7, -7.83) | sqrt((4.38-4.7)2+(34.9933-(-7.83))2) =sqrt(0.1024+1833.8350) = 42.8245 | sqrt((3.05-4.7)2+(-8.076-(-7.83))2) =sqrt(2.7225+0.0605) = 1.6682 | Centroid 2 (unchanged) |
| (4.38, 34.43) | sqrt((4.38-4.38)2+(34.9933-34.43)2) =sqrt(0+0.3173) = 0.5633 | sqrt((3.05-4.38)2+(-8.076-34.43)2) =sqrt(1.7689+1806.7600) = 42.5268 | Centroid 1 (unchanged) |
| (0.47, -2.5) | sqrt((4.38-0.47)2+(34.9933-(-2.5))2) =sqrt(15.2881+1405.7475) = 37.6966 | sqrt((3.05-0.47)2+(-8.076-(-2.5))2) =sqrt(6.6564+31.0918) = 6.1440 | Centroid 2 (unchanged) |
| (4.07, 33.8) | sqrt((4.38-4.07)2+(34.9933-33.8)2) =sqrt(0.0961+1.4240) = 1.5201 | sqrt((3.05-4.07)2+(-8.076-33.8)2) =sqrt(1.0404+1753.5994) = 41.8884 | Centroid 1 (unchanged) |
| (5.01, -16.59) | sqrt((4.38-5.01)2+(34.9933-(-16.59))2) =sqrt(0.3969+2660.8368) = 51.5871 | sqrt((3.05-5.01)2+(-8.076-(-16.59))2) =sqrt(3.8416+72.4882) = 8.7367 | Centroid 2 (unchanged) |

Since none of the points changed clusters, this bisection concludes in the following state:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AC | (4.38, 34.9933) | {(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)} | 5.8100 |
| AD | (3.05, -8.076) | {(4.38, -15.96), (0.69, 2.5), (4.7, -7.83), (0.47, -2.5), (5.01, -16.59)} | 298.2090 |
| Total SSE | | | 304.0190 |

*SSE calculations*

Cluster AC: 1.78382+0.56332+1.52012=5.8100   
Cluster AD: 7.99542+10.83612+1.66822+6.14402+8.73672=298.2090

*Selecting a bisection*

With identical SSEs and clusters, the difference between each set of clusters generated is non-existent apart from name. Arbitrarily, we’ll pick the first one generated.

**Cluster Step 3: Split into two on the highest-SSE cluster, using two trials to find the lowest-SSE bisection.**

The current set of clusters is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AA | (3.05, -8.076) | {(4.38, -15.96), (0.69, 2.5), (4.7, -7.83), (0.47, -2.5), (5.01, -16.59)} | 298.2090 |
| AB | (4.38, 34.9933) | {(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)} | 5.8100 |
| Total SSE | | | 304.0190 |

Since Cluster AA has the highest SSE, it will be bisected using 2 trials of K-means to generate the next pair of clusters. These trials will again use the first 2 and last 2 points as centroids.

*Trial 1: K-Means Clustering around (4.38, -15.96) [Centroid 1] and (0.69, 2.5) [Centroid 2]*

*Trial 1 Round 1: (4.38, -15.96) [Centroid 1] and (0.69, 2.5) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | 0 | sqrt((0.69-4.38)2+(2.5-(-15.96))2) =sqrt(13.6161+340.7716) =18.8252 | Centroid 1 |
| (0.69, 2.5) | sqrt((4.38-0.69)2+(-15.96-2.5)2) =sqrt(13.6161+340.7716) =18.8252 | 0 | Centroid 2 |
| (4.7, -7.83) | sqrt((4.38-4.7)2+(-15.96-(-7.83))2) =sqrt(0.1024+66.0969) = 8.1363 | sqrt((0.69-4.7)2+(2.5-(-7.83))2) =sqrt(16.0801+106.7089) = 11.0810 | Centroid 1 |
| (0.47, -2.5) | sqrt((4.38-0.47)2+(-15.96-(-2.5))2) =sqrt(15.2881+181.1716) = 14.0164 | sqrt((0.69-0.47)2+(2.5-(-2.5))2) =sqrt(0.0484+25) = 5.0048 | Centroid 2 |
| (5.01, -16.59) | sqrt((4.38-5.01)2+(-15.96-(-16.59))2) =sqrt(0.3969+0.3969) = 0.8910 | sqrt((0.69-5.01)2+(2.5-(-16.59))2) =sqrt(18.6624+364.4281) = 19.5727 | Centroid 1 |

New centroids = (mean(x), mean(y)) for each set: (4.6967, -13.46) and (0.58, 0)

*Trial 1 Round 2: (4.6967, -13.46) [Centroid 1] and (0.58, 0) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((4.6967-4.38)2+(-13.46-(-15.96))2) =sqrt(0.1003+6.25) = 2.5200 | sqrt((0.58-4.38)2+(0-(-15.96))2) =sqrt(14.44+254.7216) = 16.4061 | Centroid 1 (unchanged) |
| (0.69, 2.5) | sqrt((4.6967-0.69)2+(-13.46-2.5)2) =sqrt(16.0536+254.7216) = 16.4552 | sqrt((0.58-0.69)2+(0-2.5)2) =sqrt(0.0121+6.25) = 2.5024 | Centroid 2 (unchanged) |
| (4.7, -7.83) | sqrt((4.6967-4.7)2+(-13.46-(-7.83))2) =sqrt(0.00001+31.6969) = 5.6300 | sqrt((0.58-4.7)2+(0-(-7.83))2) =sqrt(16.9744+61.3089) = 8.8478 | Centroid 1 (unchanged) |
| (0.47, -2.5) | sqrt((4.6967-0.47)2+(-13.46-(-2.5))2) =sqrt(17.8650+120.1216) = 11.7468 | sqrt((0.58-0.47)2+(0-(-2.5))2) =sqrt(0.0121+6.25) = 2.5024 | Centroid 2 (unchanged) |
| (5.01, -16.59) | sqrt((4.6967-5.01)2+(-13.46-(-16.59))2) =sqrt(0.0982+9.7969) = 3.1456 | sqrt((0.58-5.01)2+(0-(-16.59))2) =sqrt(19.6249+275.2281) = 17.1713 | Centroid 1 (unchanged) |

Since none of the points changed clusters, the centroids do not adjust and this is the final split.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AAA | (4.6967, -13.46) | {(4.38, -15.96), (4.7, -7.83), (5.01, -16.59)} | 47.9421 |
| AAB | (0.58, 0) | {(0.69, 2.5), (0.47, -2.5)} | 12.5240 |
| Total SSE | | | 60.4661 |

*SSE calculations*

Cluster AAA: 2.52002+5.63002+3.14562=47.9421  
Cluster AAB: 2.50242+2.50242=12.5240

*Trial 2: K-Means Clustering around (0.47, -2.5) [Centroid 1] and (5.01, -16.59) [Centroid 2]*

*Trial 2 Round 1: (0.47, -2.5) [Centroid 1] and (5.01, -16.59) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((0.47-4.38)2+(-2.5-(-15.96))2) =sqrt(15.2881+181.1716) = 14.0164 | sqrt((5.01-4.38)2+(-16.59-(-15.96))2) =sqrt(0.3969+0.3969) = 0.8910 | Centroid 2 |
| (0.69, 2.5) | sqrt((0.47-0.69)2+(-2.5-2.5)2) =sqrt(0.0484+25) = 5.0048 | sqrt((5.01-0.69)2+(-16.59-2.5)2) =sqrt(18.6624+364.4281) = 19.5727 | Centroid 1 |
| (4.7, -7.83) | sqrt((0.47-4.7)2+(-2.5-(-7.83))2) =sqrt(17.8929+28.4089) = 6.8045 | sqrt((5.01-4.7)2+(-16.59-(-7.83))2) =sqrt(0.0961+596.3364) = 24.4220 | Centroid 1 |
| (0.47, -2.5) | 0 | sqrt((5.01-0.47)2+(-16.59-(-2.5))2) =sqrt(20.6116+198.5281) = 14.8033 | Centroid 1 |
| (5.01, -16.59) | sqrt((0.47-5.01)2+(-2.5-(-16.59))2) =sqrt(20.6116+198.5281) = 14.8034 | 0 | Centroid 2 |

New centroids = (mean(x), mean(y)) for each set: (1.9533, -2.61) and (4.695, -16.275)

*Trial 2 Round 2: (1.9533, -2.61) [Centroid 1] and (4.695, -16.275) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((1.9533-4.38)2+(-2.61-(-15.96))2) =sqrt(5.8889+178.2225) = 13.5688 | sqrt((4.695-4.38)2+(-16.275-(-15.96))2) =sqrt(0.0992+0.0992) = 0.4454 | Centroid 2 (unchanged) |
| (0.69, 2.5) | sqrt((1.9533-0.69)2+(-2.61-2.5)2) =sqrt(1.5959+26.1121) = 5.2638 | sqrt((4.695-0.69)2+(-16.275-2.5)2) =sqrt(16.0400+352.5006) = 19.1974 | Centroid 1 (unchanged) |
| (4.7, -7.83) | sqrt((1.9533-4.7)2+(-2.61-(-7.83))2) =sqrt(7.5444+27.2484) = 5.8985 | sqrt((4.695-4.7)2+(-16.275-(-7.83))2) =sqrt(0.00003+71.3180) = 8.4450 | Centroid 1 (unchanged) |
| (0.47, -2.5) | sqrt((1.9533-0.47)2+(-2.61-(-2.5))2) =sqrt(2.2002+0.0121) =1.4874 | sqrt((4.695-0.47)2+(-16.275-(-2.5))2) =sqrt(17.8506+189.7506) = 14.4084 | Centroid 1 (unchanged) |
| (5.01, -16.59) | sqrt((1.9533-5.01)2+(-2.61-(-16.59))2) =sqrt(9.3434+195.4404) = 14.3103 | sqrt((4.695-5.01)2+(-16.275-(-16.59))2) =sqrt(0.0992+0.0992) = 0.4454 | Centroid 2 (unchanged) |

Since none of the points changed clusters, the centroids do not adjust and this is the final split.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AAC | (1.9533, -2.61) | {(0.69, 2.5), (4.7, -7.83), (0.47, -2.5)} | 64.7123 |
| AAD | (4.695, -16.275) | {(4.38, -15.96), (5.01, -16.59)} | 0.3968 |
| Total SSE | | | 65.1091 |

*SSE calculations*

Cluster AAC: 5.26382+5.89852+1.48742= 64.7123  
Cluster AAD: 0.44542+0.44542= 0.3968

*Selecting a bisection*

Comparing the two trial results:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AAA | (4.6967, -13.46) | {(4.38, -15.96), (4.7, -7.83), (5.01, -16.59)} | 47.9421 |
| AAB | (0.58, 0) | {(0.69, 2.5), (0.47, -2.5)} | 12.5240 |
| Total SSE | | | 60.4661 |
|  | | | |
| AAC | (1.9533, -2.61) | {(0.69, 2.5), (4.7, -7.83), (0.47, -2.5)} | 64.7123 |
| AAD | (4.695, -16.275) | {(4.38, -15.96), (5.01, -16.59)} | 0.3968 |
| Total SSE | | | 65.1091 |

With a lower SSE, the first trial is the more optimal bisection, and will be retained as the cluster set generated for this iteration.

**Cluster Step 3: Split into two on the highest-SSE cluster, using two trials to find the lowest-SSE bisection.**

The current set of clusters is as follows:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AAA | (4.6967, -13.46) | {(4.38, -15.96), (4.7, -7.83), (5.01, -16.59)} | 47.9421 |
| AAB | (0.58, 0) | {(0.69, 2.5), (0.47, -2.5)} | 12.5240 |
| AB | (4.38, 34.9933) | {(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)} | 5.8100 |
| Total SSE | | | 66.2761 |

Since Cluster AAA has the highest SSE, it will be bisected using 2 trials of K-means to generate the next pair of clusters. These trials will again use the first 2 and last 2 points as centroids (since this set only contains 3 points, the trials will have 1 centroid in common).

*Trial 1: K-Means Clustering around (4.38, -15.96) [Centroid 1] and (4.7, -7.83) [Centroid 2]*

*Trial 1 Round 1: (4.38, -15.96) [Centroid 1] and (4.7, -7.83) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | 0 | sqrt((4.7-4.38)2+(-7.83-(-15.96))2) =sqrt(0.1024+66.0969) = 8.1363 | Centroid 1 |
| (4.7, -7.83) | sqrt((4.38-4.7)2+(-15.96-(-7.83))2) =sqrt(0.1024+66.0969) = 8.1363 | 0 | Centroid 2 |
| (5.01, -16.59) | sqrt((4.38-5.01)2+(-15.96-(-16.59))2) =sqrt(0.3969+0.3969) = 0.8910 | sqrt((4.7-5.01)2+(-7.83-(-16.59))2) =sqrt(0.0961+76.7376) = 8.7655 | Centroid 1 |

New centroids = (mean(x), mean(y)) for each set: (4.695, -16.275) and (4.7, -7.83)

*Trial 1 Round 2: (4.695, -16.275) [Centroid 1] and (4.7, -7.83) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((4.695-4.38)2+(-16.275-(-15.96))2) =sqrt(0.0992+0.0992) = 0.4454 | sqrt((4.7-4.38)2+(-7.83-(-15.96))2) =sqrt(0.1024+66.0969) = 8.1363 | Centroid 1 (unchanged) |
| (4.7, -7.83) | sqrt((4.695-4.7)2+(-16.275-(-7.83))2) =sqrt(0.00003+71.3180) = 8.4450 | 0 | Centroid 2 (unchanged) |
| (5.01, -16.59) | sqrt((4.695-5.01)2+(-16.275-(-16.59))2) =sqrt(0.0992+0.0992) = 0.4454 | sqrt((4.7-5.01)2+(-7.83-(-16.59))2) =sqrt(0.0961+76.7376) = 8.7655 | Centroid 1 (unchanged) |

Since none of the points changed clusters, the centroids do not adjust and this is the final split.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AAAA | (4.695, -16.275) | {(4.38, -15.96), (5.01, -16.59)} | 0.3968 |
| AAAB | (4.7, -7.83) | {(4.7, -7.83)} | 0 |
| Total SSE | | | 0.3968 |

*SSE calculations*

Cluster AAAA: 0.44542+0.44542= 0.3968  
Cluster AAAB: 02 = 0

*Trial 2: K-Means Clustering around (4.7, -7.83) [Centroid 1] and (5.01, -16.59) [Centroid 2]*

*Trial 2 Round 1: (4.7, -7.83) [Centroid 1] and (5.01, -16.59) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((4.7-4.38)2+(-7.83-(-15.96))2) =sqrt(0.1024+66.0969) = 8.1363 | sqrt((5.01-4.38)2+(-16.59-(-15.96))2) =sqrt(0.3969+0.3969) = 0.8910 | Centroid 2 |
| (4.7, -7.83) | 0 | sqrt((5.01-4.7)2+(-16.59-(-7.83))2) =sqrt(0.0961+596.3364) = 24.4220 | Centroid 1 |
| (5.01, -16.59) | sqrt((4.7-5.01)2+(-7.83-(-16.59))2) =sqrt(0.0961+76.7376) = 8.7655 | 0 | Centroid 2 |

Much like with the second trial for the first bisection, the centroid split is the same (just labeled differently), so the math is identical (but mirrored) for the final centroid outcome and SSE.

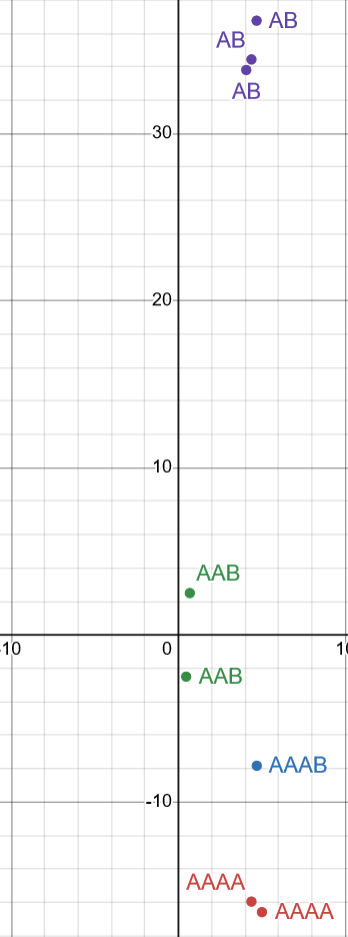
*Trial 1 Round 2: (4.695, -16.275) [Centroid 1] and (4.7, -7.83) [Centroid 2]*

|  |  |  |  |
| --- | --- | --- | --- |
| **Point** | **Distance to Centroid 1** | **Distance to Centroid 2** | **Closest Centroid** |
| (4.38, -15.96) | sqrt((4.7-4.38)2+(-7.83-(-15.96))2) =sqrt(0.1024+66.0969) = 8.1363 | sqrt((4.695-4.38)2+(-16.275-(-15.96))2) =sqrt(0.0992+0.0992) = 0.4454 | Centroid 2 (unchanged) |
| (4.7, -7.83) | 0 | sqrt((4.695-4.7)2+(-16.275-(-7.83))2) =sqrt(0.00003+71.3180) = 8.4450 | Centroid 1 (unchanged) |
| (5.01, -16.59) | sqrt((4.7-5.01)2+(-7.83-(-16.59))2) =sqrt(0.0961+76.7376) = 8.7655 | sqrt((4.695-5.01)2+(-16.275-(-16.59))2) =sqrt(0.0992+0.0992) = 0.4454 | Centroid 2 (unchanged) |

Since none of the points changed clusters, the centroids do not adjust and this is the final split.

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AAAC | (4.7, -7.83) | {(4.7, -7.83)} | 0 |
| AAAD | (4.695, -16.275) | {(4.38, -15.96), (5.01, -16.59)} | 0.3968 |
| Total SSE | | | 0.3968 |

Arbitrarily, with an identical SSE, the first bisection will be chosen and added to the cluster set.

****

**Final Result**

With 4 clusters generated, this is the final set of clusters:

|  |  |  |  |
| --- | --- | --- | --- |
| **Cluster** | **Centroid** | **Contents** | **SSE** |
| AAAA | (4.695, -16.275) | {(4.38, -15.96), (5.01, -16.59)} | 0.3968 |
| AAAB | (4.7, -7.83) | {(4.7, -7.83)} | 0 |
| AAB | (0.58, 0) | {(0.69, 2.5), (0.47, -2.5)} | 12.5240 |
| AB | (4.38, 34.9933) | {(4.69, 36.75), (4.38, 34.43), (4.07, 33.8)} | 5.8100 |
| Total SSE | | | 18.7308 |

On the right is the cluster set on a plot generated using the Desmos graphing calculator. Predictably, the cluster with the highest SSE has the greatest distance between points.

2.2) Here is the similarity matrix for the points in the set, using the formula 1/(1+dist(a, b)) as the measure of similarity with dist referring to Euclidean distance. The points themselves are in color coded cells to indicate which set they belong to:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **(4.38, -15.96)** | **(5.01, -16.59)** | **(4.7, -7.83)** | **(0.69, 2.5)** | **(0.47, -2.5)** | **(4.69, 36.75)** | **(4.38, 34.43)** | **(4.07, 33.8)** |
| **(4.38, -15.96)** | | 1 | 0.5288 | 0.1095 |  |  |  |  |  |
| **(5.01, -16.59)** | | 0.5288 | 1 |  |  |  |  |  |  |
| **(4.7, -7.83)** | | 0.1095 |  | 1 |  |  |  |  |  |
| **(0.69, 2.5)** | |  |  |  | 1 |  |  |  |  |
| **(0.47, -2.5)** | |  |  |  |  | 1 |  |  |  |
| **(4.69, 36.75)** | |  |  |  |  |  | 1 |  |  |
| **(4.38, 34.43)** | |  |  |  |  |  |  | 1 |  |
| **(4.07, 33.8)** | |  |  |  |  |  |  |  | 1 |