Large Scale Computing 6D Data set Assignment\_Joon Jung

First, seeing how many clusters are there using Silhouette score. It would give me an error starting from 1, so it starts from 2. Intuitively, it makes sense why 1 wouldn’t work, with graph shape.

A screen shot of a computer

Description automatically generated

Seems like k=6 has the highest score. Then I was able to find the cluster centers for these 6 objects

A black background with white numbers

Description automatically generated

This was the data point count for cluster

A screenshot of a computer program

Description automatically generated

This is later outcome after writing more code to spit out explained variance using PCA. I was choosing random dimensions and trying to visualize each one of them and I wasn’t getting anywhere, until a classmate let me know of doing explained variance using PCA makes things quite clear, so with some help, I was able to do that.

A black and white text

Description automatically generated with medium confidence

This is the explained variance for the 6 clusters. I have rounded up for convenience and very low number would equal to 0. Looking at explained variance for each cluster, I was finally able to get a sense of what dimension they are.

Cluster 0

Explained Variance : [0.667532579484611,0.16840892427319903,0.16405646522964556,6.929762936557088e-07,6.789404767085611e-07,6.590957742331805e-07]

=

[0.67,0.17,0.17,0,0,0] (This means the shape is 3D shape)

Cluster 1

[0.17909312541701872,0.1687224314623146,0.16829984627464847,0.16407410268450642,0.16242286350966054,0.15738763065185113]

=

[0.18,0.17,0.17,0.16,0.16,0.16] (This means the shape is 6D shape)

Cluster 2

[0.8051414231803381,0.19485541000405773,8.360352638480423e-07,8.09783183448728e-07,7.874518208281083e-07,7.335453360552615e-07]

=

[0.805,0.195,0,0,0,0] (This means the shape is 2D shape)

Cluster 3

[0.2062124440163537,0.19883410326848905,0.1891570934422642,0.18217913230927205,0.17708422840739735,0.04653299855622356]

=

[0.21,0.20,0.19,0.18,0.18,0.05] (This means the shape is 6D shape)

Cluster 4

[0.5166967490795857,0.48329355208320635,2.582002320051601e-06,2.4169661102865635e-06,2.386097312316698e-06,2.313771465276198e-06]

=

[0.52,0.48,0,0,0,0] (This means the shape is 2D shape)

Cluster 5

[0.9999999999999999,7.341032373504719e-17,1.773140751195716e-48,1.3113388294429844e-79,2.1551332000955306e-111,0.0]

=

[1,0,0,0,0,0] (This means the shape is 1D shape)

After knowing the dimensions of them, it became much helpful. I would choose other PCA and other dimension combination for 3D plot I am using and try to visualize it in my head as I go through different angles and projections of the objects in 3D dimensional space. I have purposely used 2D for visualization for some of them to show the actual shape of them well.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Cluster | Object | Center location  (Rounded) | Dimension | Points |
| Cluster 0 | A blue square with lines and numbers  Description automatically generated with medium confidenceCube (but 3D) | [25, 25, 25, 75, 75, 75] | 3D | 3500 |
| Cluster 1 | A graph with a blue spot  Description automatically generated with medium confidenceCube (but 6D) | [15, 80, 15, 80, 15, 80] | 6D | 3995 |
| Cluster 2 | Square | [75, 75, 75, 75, 75, 75] | 2D | 3000 |
| Cluster 3 | A graph of a graph with a blue object  Description automatically generated with medium confidence  Tetrapod (things you see in piers) | [0.085, 0.082, 0.082, 0.080, 0.087, 0.088] | 6D | 2400 |
| Cluster 4 | A graph with blue dots  Description automatically generatedElipse or Circle | [70, 60, 50, 40, 30, 20] | 2D | 2500 |
| Cluster 5 | A graph of a line in a grid  Description automatically generatedStraight Line | [14.88875627, 14.88875627, 14.88875627, 14.88875627, 14.88875627, 14.88875627] | 1D | 1994 |