Mouhammadou Dabo

April 8, 2021

CS 1675: Intro to Machine Learning

Professor Milos Hauskrecht

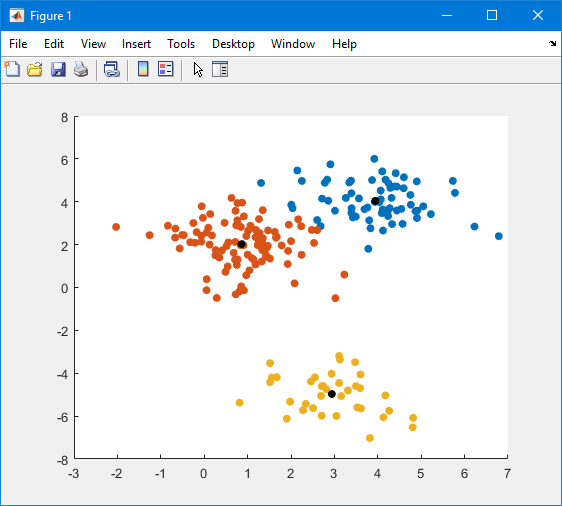
Problem Assignment 9

**Problem 1.** K-means clustering

**Part a.**

**Sizes of groups:** 66, 98, 36

**Distance:** 76.4997



**Plot 1**: k-means clustering – 3 clusters

**Part b.**

**Sizes of groups:** 49, 36, 63, 52

**Distance**: 84.2883

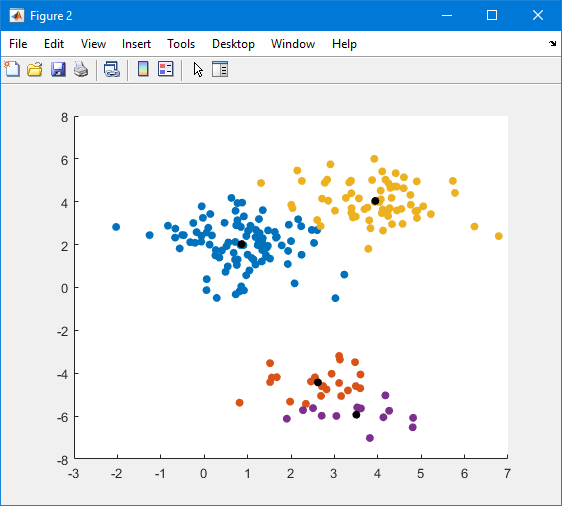
****

**Plot 2**: k-means clustering – 4 clusters

**Part c.**

**Sizes of groups:** 98, 23, 66, 13

**Distance**: 96.7493

****

**Plot 3:** k-means clustering – 4 clusters (rerun)

**Part d.**

Since the k-means clustering algorithm converges to centers minimizing the sum of squared center-point distances, the run with the smallest distance would be the best one.

A math expression that would help me compare these different clusterings and pick the best is:

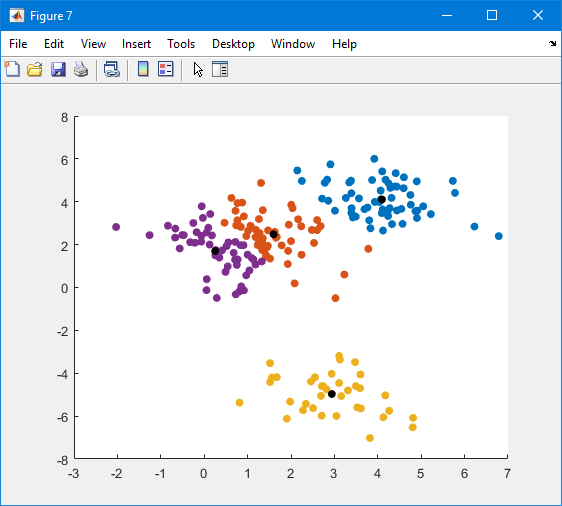
where

**Part e.**

|  |  |  |
| --- | --- | --- |
| **Run #** | **Cluster Sizes** | **Distance** |
| **1** | 65 36 39 60 | 85.2187 |
| **2** | 98 10 66 26 | 96.5767 |
| **3** | 36 63 51 50 | 84.2718 |
| **4** | 39 36 65 60 | 85.2187 |
| **5** | 52 63 49 36 | 84.2883 |
| **6** | 18 66 18 98 | 96.0650 |
| **7** | 60 54 36 50 | 81.6453 |
| **8** | 96 36 40 28 | 83.3128 |
| **9** | 98 10 66 26 | 96.5767 |
| **10** | 36 63 69 32 | 82.2638 |
| **11** | 96 36 28 40 | 83.3128 |
| **12** | 23 98 66 13 | 96.7493 |
| **13** | 61 36 40 63 | 83.1815 |
| **14** | 51 36 52 61 | 81.7123 |
| **15** | 98 23 66 13 | 96.7493 |
| **16** | 63 36 49 52 | 84.2883 |
| **17** | 52 36 61 51 | 81.7123 |
| **18** | 69 63 32 36 | 82.2638 |
| **19** | 63 38 63 36 | 83.1438 |
| **20** | 38 36 98 28 | 83.1076 |
| **21** | 61 36 42 61 | 82.2850 |
| **22** | 60 36 54 50 | 81.6453 |
| **23** | 46 89 36 29 | 82.7662 |
| **24** | 31 36 92 41 | 82.7317 |
| **25** | 66 10 98 26 | 96.5767 |
| **26** | 66 10 26 98 | 96.5767 |
| **27** | 32 36 69 63 | 82.2638 |
| **28** | 36 89 46 29 | 82.7662 |
| **29** | 98 26 66 10 | 96.5767 |
| **30** | 38 36 29 97 | 83.2772 |

**Table 1**: K-means with k =4

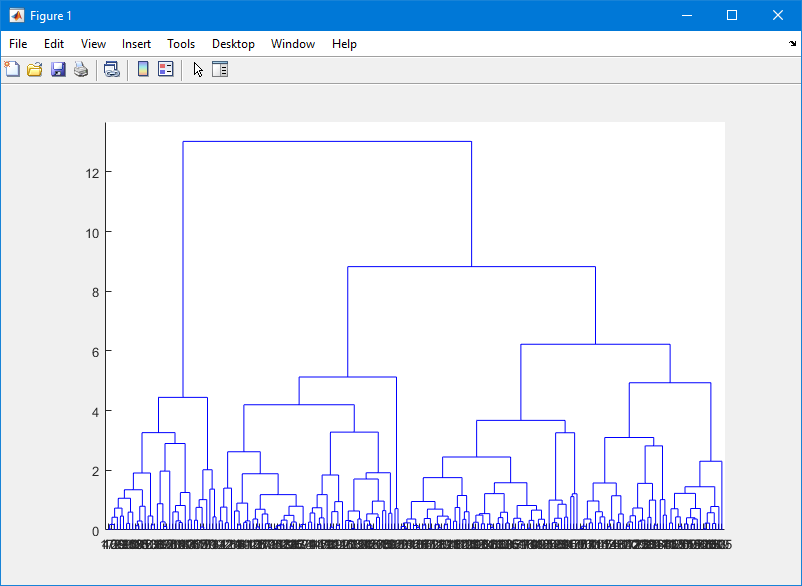
* By using the formula from Part d, the clustering that is the best is:
  + Run #7 with a cluster size of [60 54 36 50] and a distance of 81.6453

****

**Plot 4:** Run #7 – k-means clustering

**Problem 2. Hierarchical clustering**

**Part a.**

****

**Graph 1:** Dendrogram of Full Cluster

**Part b.**

****

**Plot 5:** Hierarchical Clustering Scatter Plot

* Comparing these results to the ones of Problem 2. Part e., the two clusters are different. One of the groupings match between the two (the purple from plot 5 and the yellow from plot 4).

**Problem 3.** Feature/Input ranking

**Part a.**

|  |  |  |
| --- | --- | --- |
| **Top** | **Dimension** | **Fisher score** |
| **1** | 48 | 0.3192 |
| **2** | 25 | 0.2140 |
| **3** | 21 | 0.1910 |
| **4** | 70 | 0.1892 |
| **5** | 65 | 0.1693 |
| **6** | 40 | 0.1673 |
| **7** | 29 | 0.1650 |
| **8** | 19 | 0.1402 |
| **9** | 57 | 0.1255 |
| **10** | 20 | 0.1212 |
| **11** | 24 | 0.0995 |
| **12** | 30 | 0.0950 |
| **13** | 12 | 0.0858 |
| **14** | 47 | 0.0846 |
| **15** | 61 | 0.0607 |
| **16** | 10 | 0.0579 |
| **17** | 34 | 0.0527 |
| **18** | 27 | 0.0462 |
| **19** | 39 | 0.0461 |
| **20** | 41 | 0.0422 |

**Table 2:** Top 20 Fisher scores

**Part b.**

|  |  |  |
| --- | --- | --- |
| **Top** | **Dimension** | **AUROC score** |
| **1** | 25 | 0.7340 |
| **2** | 29 | 0.6837 |
| **3** | 11 | 0.6695 |
| **4** | 47 | 0.6661 |
| **5** | 19 | 0.6315 |
| **6** | 34 | 0.6174 |
| **7** | 32 | 0.6021 |
| **8** | 30 | 0.6021 |
| **9** | 9 | 0.6000 |
| **10** | 56 | 0.5971 |
| **11** | 27 | 0.5953 |
| **12** | 60 | 0.5929 |
| **13** | 51 | 0.5881 |
| **14** | 26 | 0.5874 |
| **15** | 53 | 0.5845 |
| **16** | 7 | 0.5797 |
| **17** | 10 | 0.5709 |
| **18** | 61 | 0.5686 |
| **19** | 43 | 0.5567 |
| **20** | 44 | 0.5422 |

**Table 3:** Top 20 AUROC scores

* Comparing the results from part a to the ones from part b, the ordered lists are different. The two lists share only a small number of similar dimensions, but even so they are not ranked the same. Generally you would not find these two to be the same since the Fisher score is used to solve maximum likelihood, while AUROC is a performance metric for discrimination.