# HW3 Part 2

Hsin-Ping Hsu (G01167652)

**1. Cluster each dataset using K-means, EM, and DBScan. Try different values of k in the interval [2,5] for K-Means.**

In order to visualize the results, I employed PCA to reduce the dimension from 5 to 2. Please note the PCA was only employed for visualization, it was not used for any clustering.

**2. Determine, using SSE which is the best value of K. Compare with what DBScan finds**.

* Test data 1:

According to the SSE curve of test data 1 (Fig. 1), there is a distinct knee at the value of K equals 3; therefore, the best value of K for this data is 3. And the best value of eps for DBScan is 10 that was found via grid search.

Comparing the results via K-means and DBScan, I observed that their results are similar as shown in Fig. 2 and Fig.3. Through DBScan, most of the points are considered as core points; some of them are considered as boarder points; the rest of them are considered as noise points.

|  |
| --- |
| 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 |
| Fig. 1 SSE curve of test data 1 |

|  |  |
| --- | --- |
| 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 | 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 |
| Fig. 2 Three clusters found by K-means | Fig. 3 Three clusters found by DBScan |

* Test data 2:

From the SSE curve of test data 2 (Fig. 4), I observed that there is no distinct knee; the minimum SSE occurs when the value of K equals 5. In addition, the best value of eps for DBScan for this data is 0.3 that was found via grid search. However, as shown in Fig. 5 and Fig. 6, the data cannot be well-separated, which is probably because of the original data is unstructured.

|  |
| --- |
| 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 |
| Fig. 4 SSE curve of test data 2 |

|  |  |
| --- | --- |
| 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 | 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 |
| Fig. 5 Five clusters found by K-means | Fig. 6 Three clusters found by DBScan |

**3. For your best clustering, in all cases (K-means, EM, and DBScan), and for both datasets, do the following**

a. Measure the clustering validity using correlation

|  |  |  |  |
| --- | --- | --- | --- |
| Correlation | K-means | EM | DBScan |
| Test Data 1 | -0.809 (K = 3) | -0.809 (K = 3) | -0.809 |
| Test Data 2 | -0.415 (K = 5) | -0.3756 (K = 5) | -0.1835 |

b. Compute the silhouette coefficient

|  |  |  |  |
| --- | --- | --- | --- |
| silhouette coefficient | K-means | EM | DBScan |
| Test Data 1 | 0.7979 (K = 3) | 0.7979 (K = 3) | 0.7979 |
| Test Data 2 | 0.1604 (K = 5) | 0.1239 (K = 5) | -0.0435 |

* Test Data 1:

|  |  |  |
| --- | --- | --- |
| 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 | 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 | 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 |
| Fig. 6 Three clusters found by  K-means | Fig. 7 Three clusters found by EM | Fig. 8 Three clusters found by DBScan |

* Test Data 2:

|  |  |  |
| --- | --- | --- |
| 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 | 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 | 一張含有 螢幕擷取畫面 的圖片  自動產生的描述 |
| Fig. 9 Five clusters found by  K-means | Fig. 10 Five clusters found by EM | Fig. 11 Three clusters found by DBScan |

**4. According to your results in 3, answer the following question: Which of the two datasets REALLY exhibits the clustering structure you found in 1.?**

According to the results found in 3, the first data can be well-separated via K-means, EM and DBScan. This data has a relatively higher correlation and silhouette coefficient compare with the second data set. On the other hand, the second data set has relatively low correlation and silhouette coefficient; one possible reason of this might because of the random generated data points.