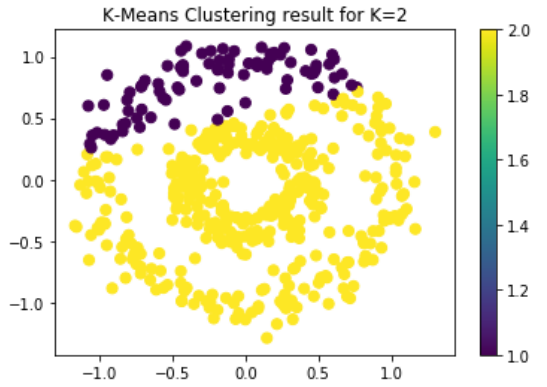
Comp4331 Assignment 3
Name: Li Chung Yat
Student ID: 20422979

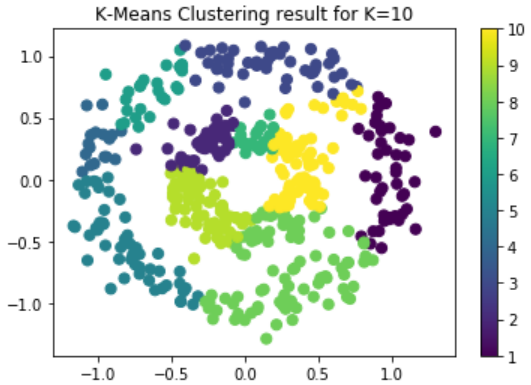
Data Mining Report

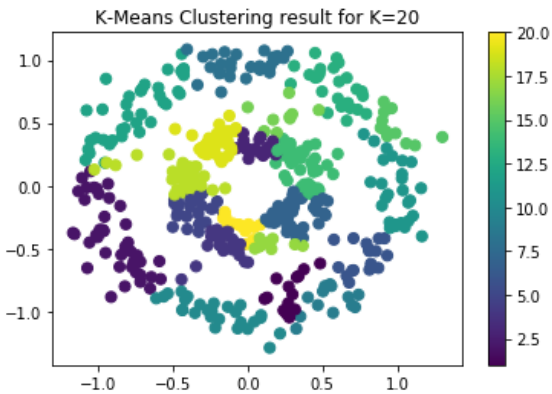
K-Mean:
Total Training time: 7.295394206000083 second

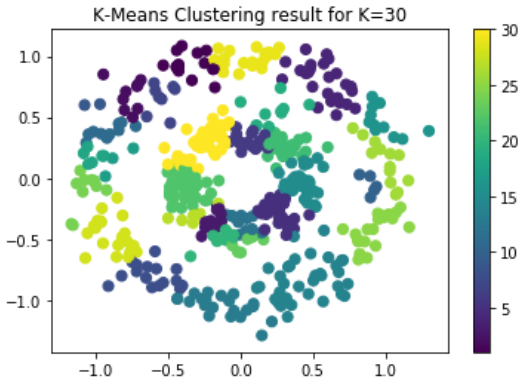
Testing with K ∈ {2,10,20,30}
K = 2,
SSE for K =2 is 330.129214086304
Silhouette Coefficient: 0.2847399620059492
Training time for K = 30 is 0.40459268099994006 second
K = 10,
SSE for K =10 is 53.58578095430631
Silhouette Coefficient: 0.297860556788313
Training time for K = 30 is 0.9922661299999618 second
K = 20,
SSE for K =20 is 26.570864476745164
Silhouette Coefficient: 0.31544861951857706
Training time for K = 30 is 2.926060317000065 second
K = 30,
SSE for K =30 is 17.10923202994908
Silhouette Coefficient: 0.3135956839692115
Training time for K = 30 is 2.784365353999874 second

Based on the clustering result above, in terms of efficiency K = 10 seems to have the best result. Although the Silhouette Coefficient is lower than K = 20 and K = 30 a bit, the training time only half of them as K = 10 needs less iterations shifting center to generate stabilize result. I used change of SSE as an approximation of centers shift as if SSE does not change, center will not be changing. If SSE stabilized in 3 consecutive iterations, I will try the next K. SSE also drop the most from K = 2 to K = 10 as well, implies that it possibly is around or on the elbow point.



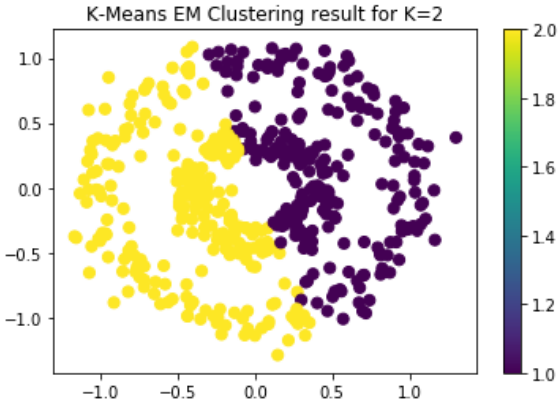






K-Mean Expectation and Maximization

K = 2,
iteration 1:
SSE for K = 2 is, 531.4376546291828
Center for cluster 1 is: 0.108910141746058 ,-0.33299521640192653
Center for cluster 2 is: -0.2804274433927719 ,0.2916622080450923
iteration 2:
SSE for K = 2 is, 329.54364762342254
Center for cluster 1 is: -0.2169750764950825 ,0.3187278401758657
Center for cluster 2 is: 0.18798323973648304 ,-0.3430193128422904
iteration 3:
SSE for K = 2 is, 344.06381292152224
Center for cluster 1 is: 0.204962768323458 ,-0.3560274898597506
Center for cluster 2 is: -0.20463643639315274 ,0.3242037008320859
iteration 4:
SSE for K = 2 is, 351.41846598124926
Center for cluster 1 is: -0.20029399436297782 ,0.326070146882678
Center for cluster 2 is: 0.20777896953899216 ,-0.3631386112433929
iteration 5:
SSE for K = 2 is, 354.27625207384904
Center for cluster 1 is: 0.20685043393892083 ,-0.367015635079573
Center for cluster 2 is: -0.19731255247723448 ,0.3274998268340583
iteration 6:
SSE for K = 2 is, 355.3668643781437
Center for cluster 1 is: -0.19458155962959134 ,0.3289588007578
Center for cluster 2 is: 0.20476669677775064 ,-0.3695110114411882
iteration 7:
SSE for K = 2 is, 355.8077579890804
Center for cluster 1 is: 0.20227269717712482 ,-0.37145597260044055
Center for cluster 2 is: -0.19189006851675558 ,0.3304728683786414
iteration 8:
SSE for K = 2 is, 356.0143561323834
Center for cluster 1 is: -0.18918738484766276 ,0.3320176033067984
Center for cluster 2 is: 0.19960868151841074 ,-0.373184632307868
iteration 9:
SSE for K = 2 is, 356.13729175724575
Center for cluster 1 is: 0.19685788266111356 ,-0.3748258855852218
Center for cluster 2 is: -0.18645752238809282 ,0.3335746284127884



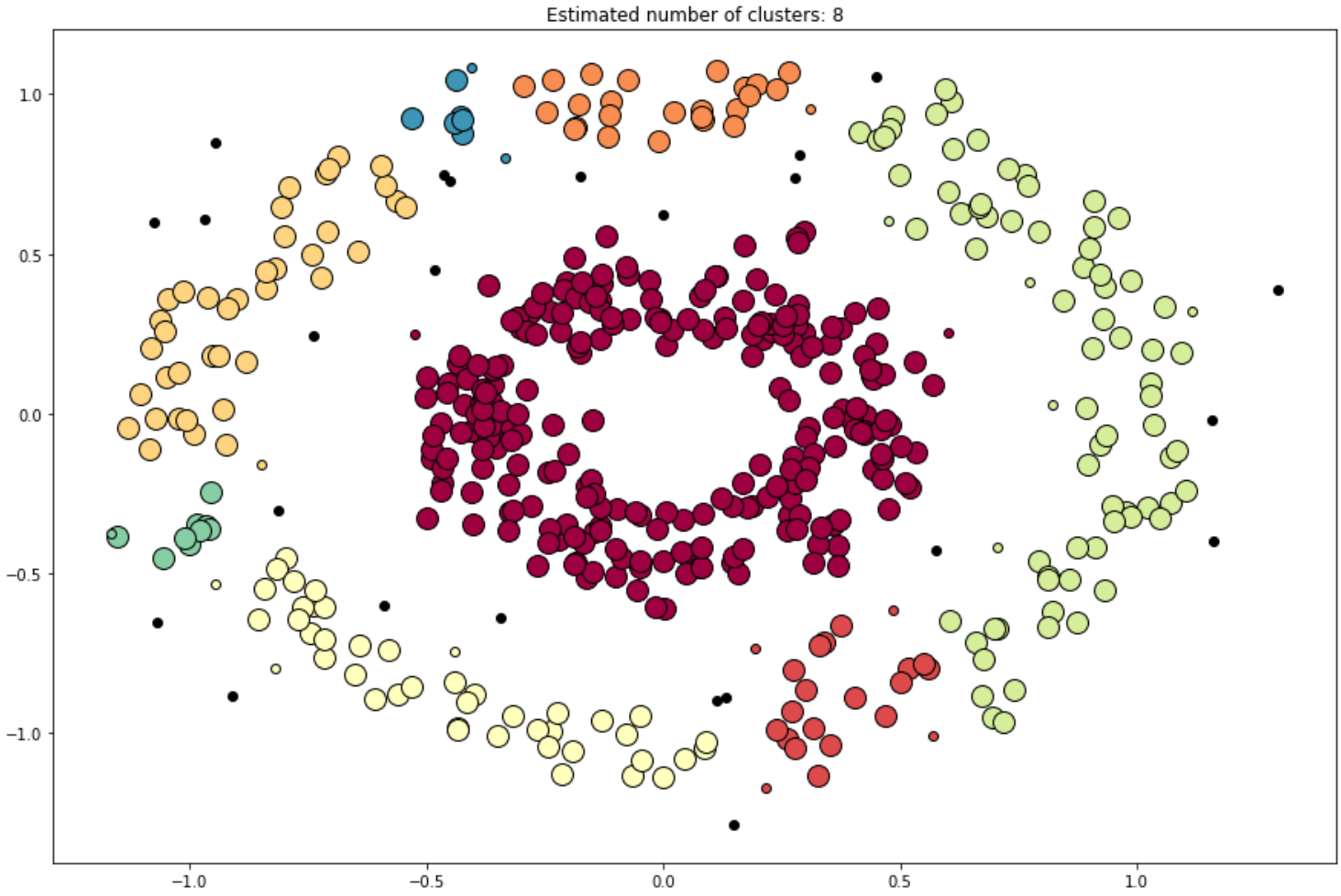
It took 9 iterations for K-Mean EM to obtain a stabilized SSE. We can observe that the center of cluster 1 and cluster 2 are shifting with each other and converge to a center point which fit the assumption of EM algorithm.

Take a look at the cluster result generated above by K-Mean EM. It has much better performance then simple K-Mean that the circular data points are approximately evenly divided into 2 halves. Although the clustering result is also bad, which should be 2 clusters in ring shape, comparing to simple K-Mean = 2, the result is much better already.

Training time takes 1.0505223079999269 which is much longer than simple K-Mean. It is because it needs 9 iterations to obtain optimal result but simple K-Mean in general takes on 3 or 4 iterations.

DBSCAN

Just to make a declaration here as well, in DBSCAN there are no copying or cheating if my codes are similar with some others RMBI students. These codes are sample codes from my RMBI3110 TA Mina, all works are originated by myself.



Using DBSCAN, the result is close enough to be described as “good”, probably if radius (eps) can be a little bit larger, the result can be two ring-like clusters.

Training time for DBSCAN is 0.20650327300018034. It is extremely fast and more accurate. DBSCAN is well-known for having good clustering result in 2 dimensions, it is believed that if we have higher dimension as dataset, DBSCAN might have poorer results compare to the other two algorithms.