CS145 HW3 Report
Shunji Zhan
405030387

Problem1:

	Cluster1	Cluster2	Cluster3	Cluster4	Sum (O)
Label 1	0	1	4	0	5
Label 2	5	0	0	0	5
Label 3	0	5	0	0	5
Label 4	0	0	1	4	5
Sum (C)	5	6	5	4	20

	Same Cluster	Different Clusters
Same Class	TP = 32	FN = 8
Different Class	FP = 9	TN = 141

Purity = (5 + 5 + 4 + 4) / 20 = **0.9**
Precision = TP / (TP + FP) = **0.78**
Recall = TP / (TP + FN) = **0.8**
F-measure = 2P * R / (P + R) = **0.79**

I(C, O) = 5 / 20 * log2(20 * 5 / (5 * 5)) +
1 / 20 * log2(20 * 1 / (5 * 6)) +
5 / 20 * log2(20 * 5 / (5 * 6)) +
4 / 20 * log2(20 * 4 / (5 * 5)) +
1 / 20 * log2(20 * 1 / (5 * 5)) +
4 / 20 * log2(20 * 4 / (5 * 4)) +
= 0.5 – 0.029 + 0.4342 + 0.3356 – 0.016 + 0.4
= 1.6248
H(O) = [(-5 / 20) * log2(5 / 20)] * 4 = 2
H(C) = -[5 / 20 * log2(5 / 20) * 2 + 6 / 20 * log2(6 / 20) + 4 / 20 * log2(4 / 20)] = 1.99
NMI(C,O) = I(C, O) / sqrt(H(C) * H(O)) = **0.814**

This result can also be checked by:
python evaluate.py
purity: 0.9
TP: 32.0
TN: 141.0
FP: 9.0
FN: 8.0
precision: 0.780487804878
recall: 0.8
F-measure: 0.79012345679

Problem2 (K-means):

Strength: logic is clear and thus easy to implement. If not a lot of noise can do pretty well.

Weakness: when the clusters have different sizes and variances, it doesn't give good result. It also doesn't do well in distinguish spherical shaped clusters. Easily affected by distant outliers.

Output:

For dataset1

Iteration :3

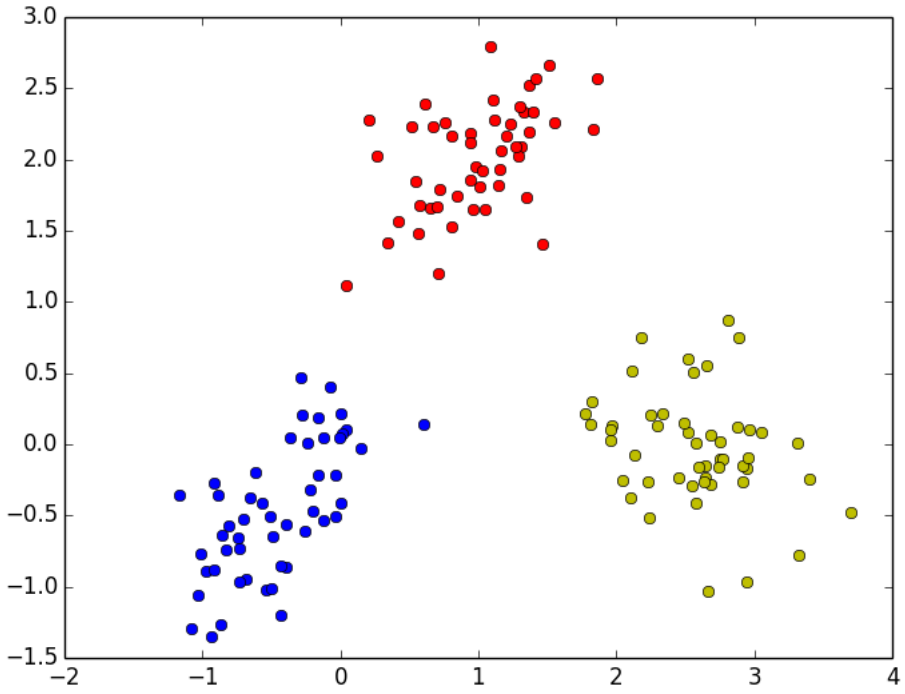
Purity is :1.0

NMI :1.0

Cluster 0 size :50

Cluster 1 size :50

Cluster 2 size :50



For dataset2

Iteration :9

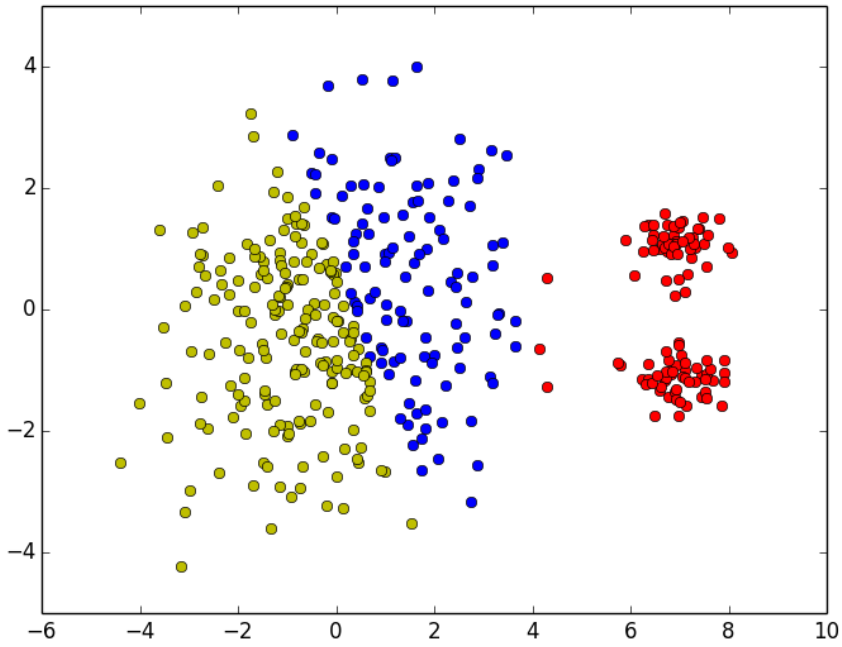
Purity is :0.8675

NMI :0.464256439333

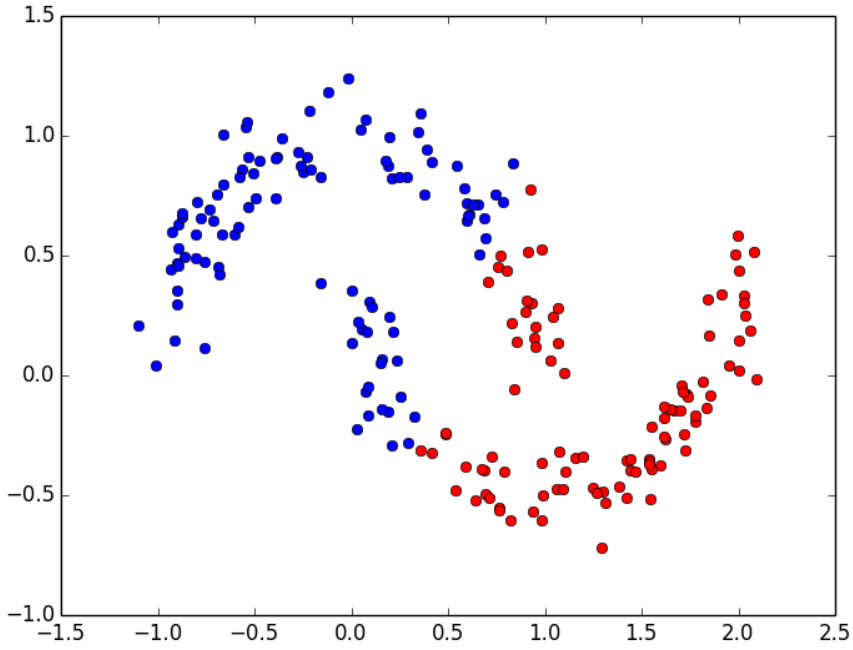
Cluster 0 size :103

Cluster 1 size :112

Cluster 2 size :185



For dataset3
Iteration :6
Purity is :0.78
NMI :0.169704955284
Cluster 0 size :98
Cluster 1 size :102



Problem3 (DBSCAN):

Strength: if parameters are chosen well, it can output reasonable result, even for spherical shaped clusters. It can also detect noises and thus minimize their negative influence on the result.

Weakness: the result is extremely dependent to parameters. Chosen different Eps and Minpts can result in totally different outcomes, and some of them are bad clustering.

Output:

For dataset1

Esp :0.477548092264

Number of clusters formed :3

Noise points :4

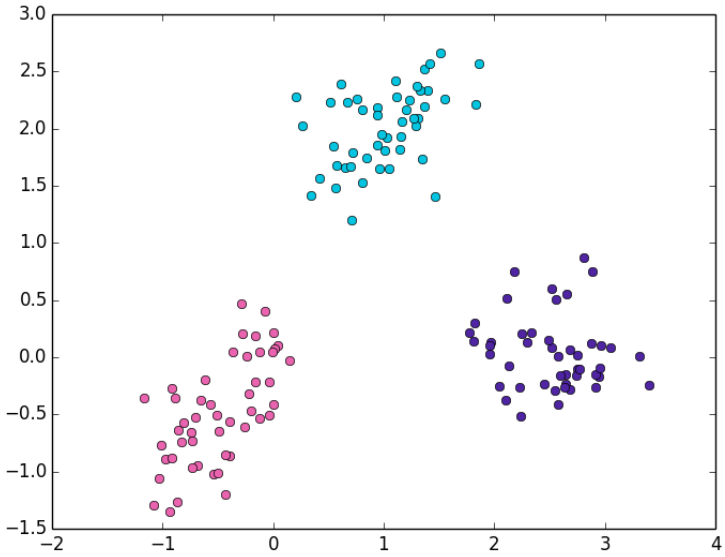
Purity is :0.973333333333

NMI :1.0

Cluster 0 size :50

Cluster 1 size :47

Cluster 2 size :49



For dataset2

Esp :0.7405776933

Number of clusters formed :3

Noise points :11

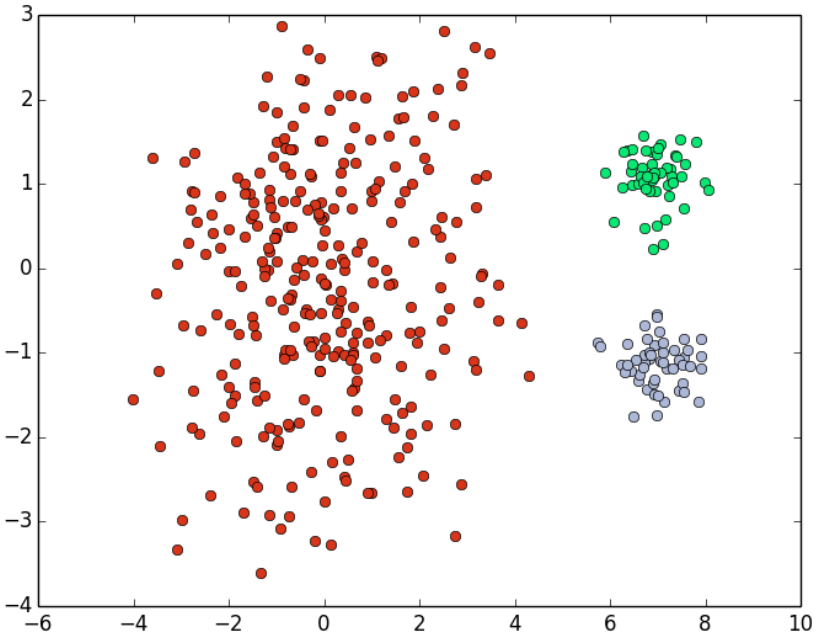
Purity is :0.9725

NMI :1.0

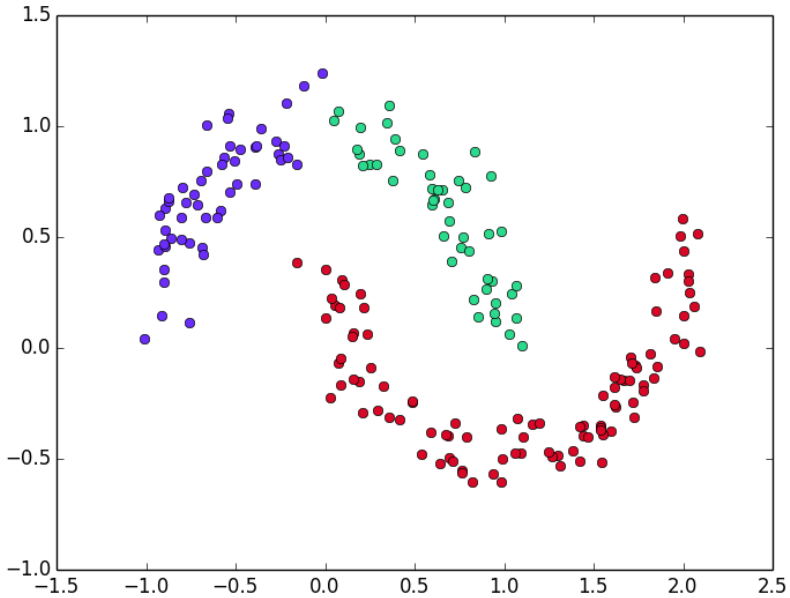
Cluster 0 size :50

Cluster 1 size :289

Cluster 2 size :50



For dataset3
Esp :0.188103671919
Number of clusters formed :3
Noise points :4
Purity is :0.985
NMI :0.817348927469
Cluster 0 size :99
Cluster 1 size :47
Cluster 2 size :51



Problem4 (GMM):

Strength: It is robust and general, will not be affected too much by outliers. Mixture models are more general than portioning.

Weakness: compared to the previous two methods, it runs relatively slow and is much harder to implement. It also converges to a local optimum, not necessarily the global optimum.

Output:

For dataset1
Number of Iterations = 26

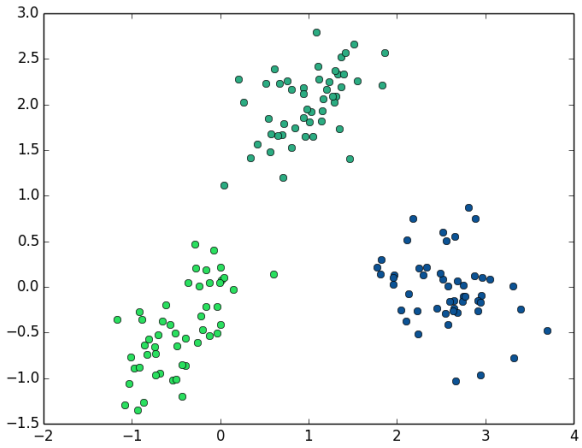
After Calculations
Final mean =
-0.462497803748 -0.463941570915
0.98985865093 2.01176299401
2.57343074153 -0.0271142473287

Final covariance =
For Cluster : 1
0.149178033089 0.117310900178
0.117310900178 0.215451381833

For Cluster : 2
0.160309281048 0.0748980182681
0.0748980182681 0.139426814158

For Cluster : 3
0.180387888855 -0.0467201611333
-0.0467201611333 0.152057292033

Purity is :1.0
NMI :1.0
Cluster 0 size :50
Cluster 1 size :50
Cluster 2 size :50



For dataset2
Number of Iterations = 60

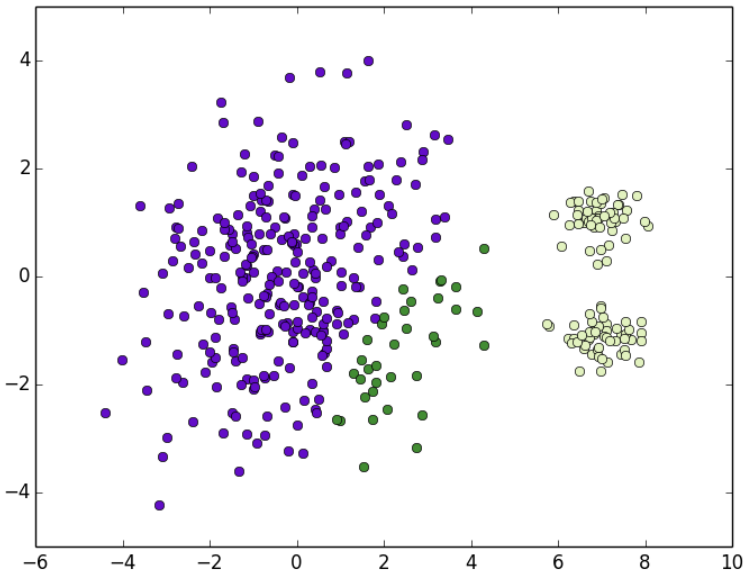
After Calculations
Final mean =
6.97649235198 -0.0259625658049
-0.3059097563 0.0290544014694
2.02079952275 -1.28868763161

Final covariance =
For Cluster : 1
0.227976706733 -0.0302421268225
-0.0302421268225 1.30257250042

For Cluster : 2
2.39135938542 0.651893810493
0.651893810493 2.28436205204

For Cluster : 3
1.57511651604 0.663164323351
0.663164323351 1.21555931019

Purity is :0.875
NMI :0.603803956121
Cluster 0 size :100
Cluster 1 size :265
Cluster 2 size :35



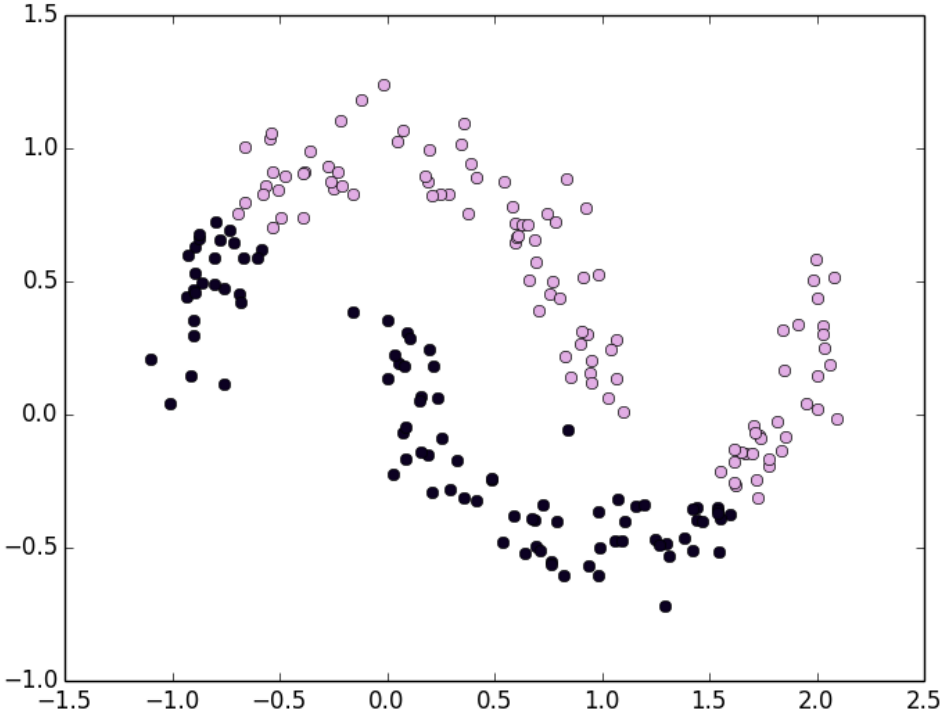
For dataset3
Number of Iterations = 110

After Calculations
Final mean =
0.746589652273 0.456362706313
0.282624134888 -0.0596588974576

Final covariance =
For Cluster : 1
0.769219007365 -0.287852063191
-0.287852063191 0.19019007583

For Cluster : 2
0.682736130041 -0.300607450603
-0.300607450603 0.175861444996

Purity is :0.69
NMI :0.075947839504
Cluster 0 size :106
Cluster 1 size :94


1.5
1.0
0.5
0.0
−0.5
−1.0
−1.5
−1.0
−0.5
0.0
0.5
1.0
1.5
2.0
2.5