CS 274A Homework 5

Zachary DeStefano, 15247592

Due Date: Wednesday March 5th

1 Results for Dataset1

1.1 k-Means Plots

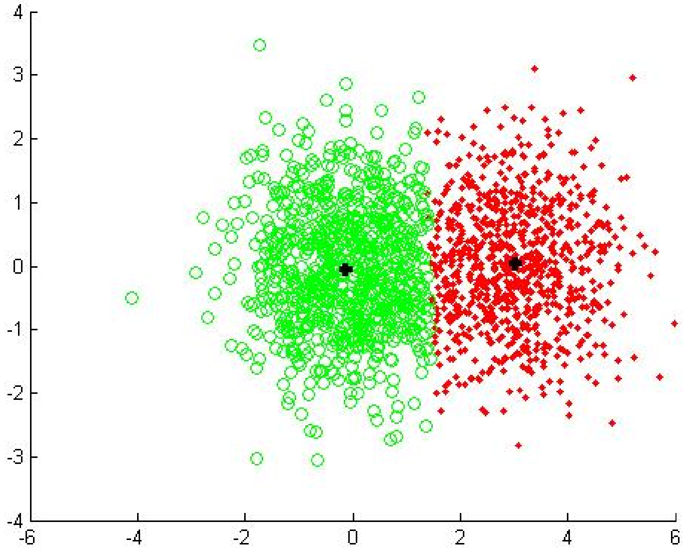

Figure 1: The k-Means plot for dataset1, K=2

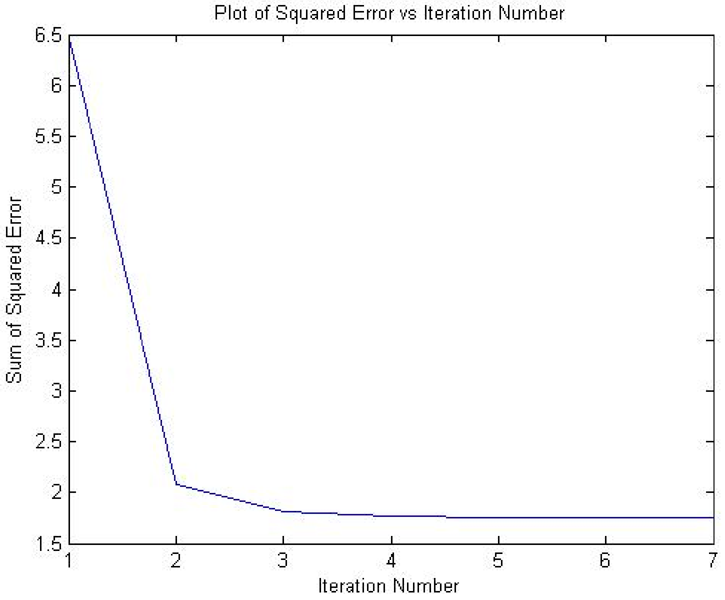

Figure 2: The Sum of Squared Error plot for dataset1, K=2

1.2 EM Plots

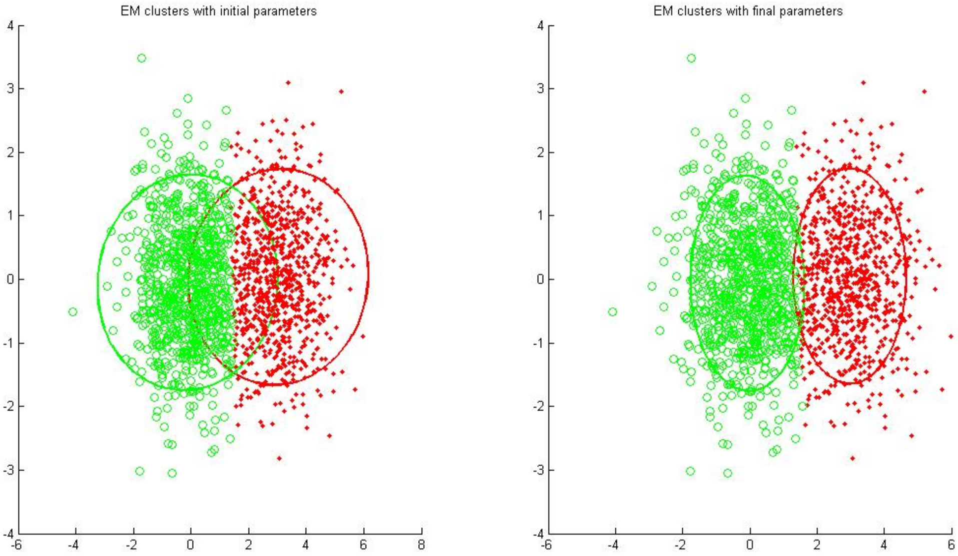

Figure 3: The EM cluster plots for dataset1, K=2, using initialization method 3

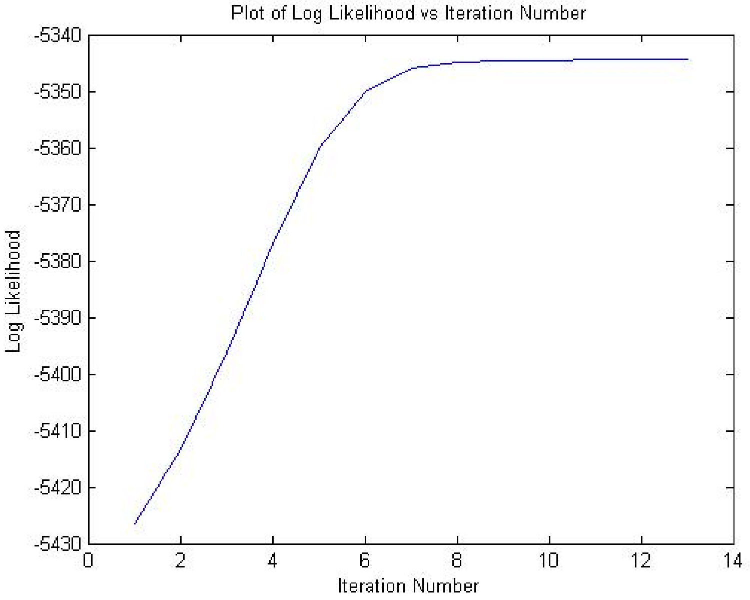

Figure 4: The likelihood plot for dataset1, K=2, using initialization method 3

1.3 Comments on k-Means and EM Results

For dataset1, there was not much of a difference between the K-means result and the EM algorithm result. This was likely due to the fact that the two Gaussians did not overlap so K-means put the points in the correct cluster easily. When I computed the best initialization method, k-Means proved to be the best one for this data set. This makes sense because the data was easily able to be grouped into two sets using k-Means.

1.4 Comments on BIC test results

Dataset 1		
K	log likelihood	BIC value
1	-5459.40	-5481.53
2	-5344.47	-5388.73
3	-5343.42	-5409.82
4	-5341.18	-5429.72
5	-5337.96	-5448.63

For dataset1, as expected, the log likelihood increases when you increase K, the number of clusters. When you use BIC to get the ideal number of clusters, we find that the BIC value is maximized at K=2. This is the solution we want since the data should be grouped into 2 clusters.

2 Plots for Dataset2

2.1 k-Means Plots

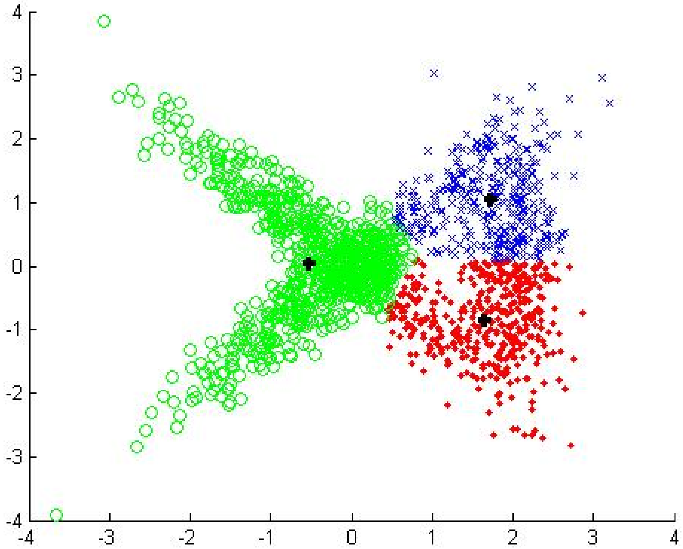

Figure 5: The k-Means plot for dataset2, K=3

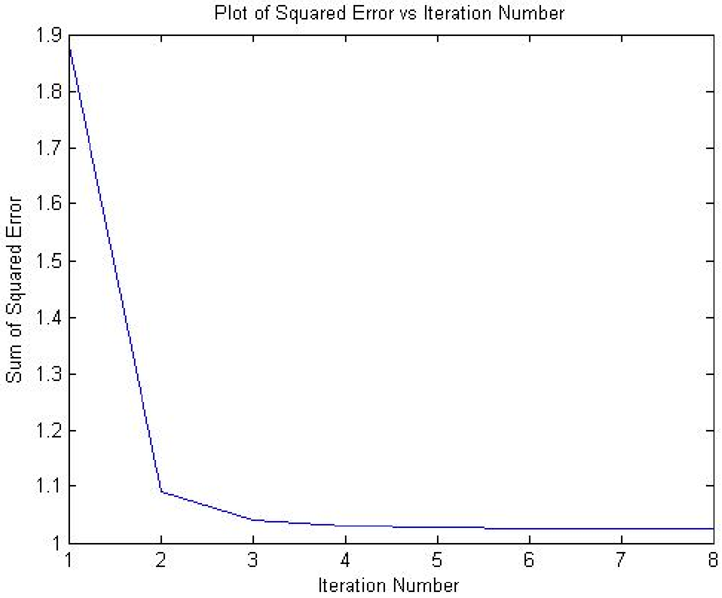

Figure 6: The Sum of Squared Error plot for dataset2, K=3

2.2 EM Plots

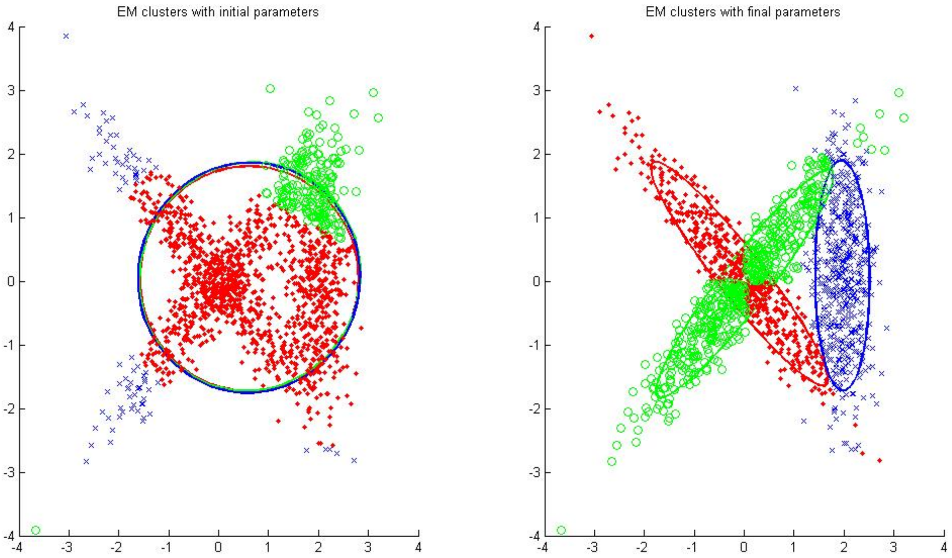

Figure 7: The EM cluster plots for dataset2, K=3, using initialization method 1

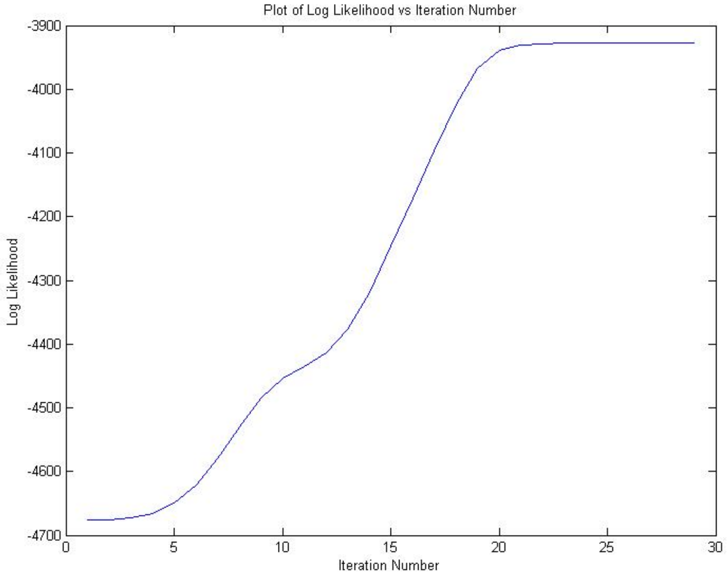

Figure 8: The likelihood plot for dataset2, K=3, using initialization method 1

2.3 Comments on k-Means and EM Results

With dataset2, there was a substantial difference between the two methods. The EM algorithm seemed to capture the clusters much better than K-means. This is likely due to the fact that two of the Gaussians used to generate the data had an overlapping mean. Since K-means just clusters them into groups, it would struggle to group those two sets. However, EM is specifically fitting the data to Gaussians. It was thus much better to use the EM algorithm to cluster this data set. For the initialization method, assigning random member weights proved to be the best one when I compared the methods. It is to be expected that there would not be much of a difference between random member weights and using k-means when initializing the data due to the fact that k-means did not give us representative groups in the first place.

2.4 Comments on BIC test results

Dataset 2		
K	log likelihood	BIC value
1	-4676.77	-4698.72
2	-4449.92	-4493.80
3	-3993.20	-3993.20
4	-3922.58	-4010.34
5	-3921.62	-4031.31

For dataset2 as with the other data sets, the log likelihood increases when we increase K, the number of clusters. When computing the BIC value, we find that it is maximized at K=3. This is what we want since the data was generated from 3 Gaussians so we wanted the best result to occur when we put the data into 3 clusters.

3 Plots for Dataset3

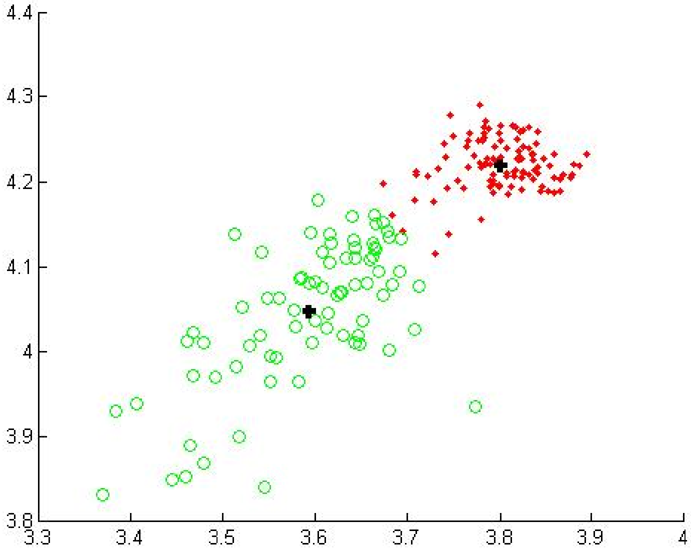

Figure 9: The k-Means plot for dataset3, K=2

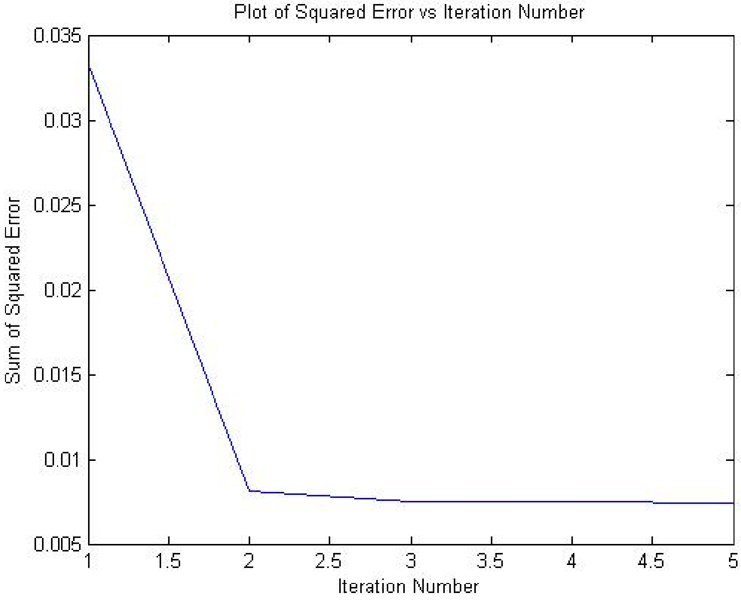

Figure 10: The Sum of Squared Error plot for dataset3, K=2

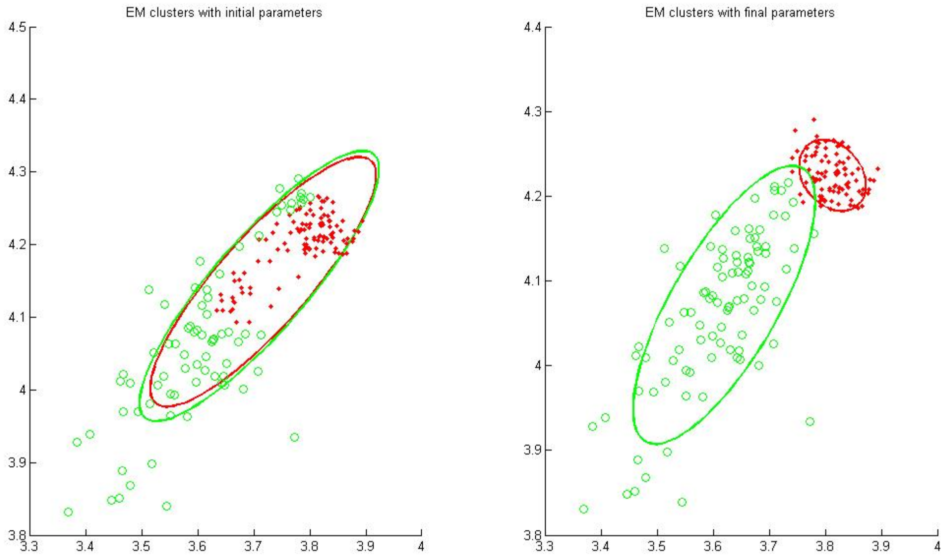

Figure 11: The EM cluster plots for dataset3, K=2, using initialization method 3

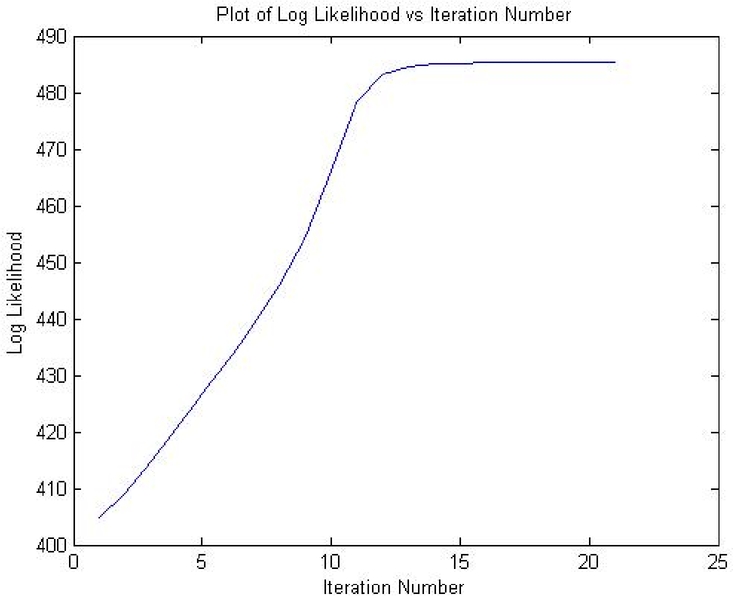

Figure 12: The likelihood plot for dataset3, K=2, using initialization method 3

3.1 Comments on k-Means and EM Results

For dataset3, we are trying to take real world patient data and cluster the patients into two groups. Both methods perform very well in this case likely due to the fact that there are no two overlapping cluster centers. I calculated the accuracy of the cluster results and found that the EM algorithm has an accuracy of 96.7%. The k-Means algorithm produced an accuracy of 96.15%. Thus, EM did end up performing slightly better but not by much at clustering our data. This could be because in general the two groups of patients have a normal distribution for those numbers so if we got more data, then EM would possibly perform even better. When I went to calculate the best initialization method for the EM run, K-means proved to be the best one. This is to be expected because it performed very well at clustering the data in the beginning.

3.2 Comments on BIC test results

Dataset 3		
K	log likelihood	BIC value
1	402.79	387.17
2	485.35	454.13
3	494.04	447.20
4	503.97	441.52
5	507.75	429.69

For dataset3, as with the other data sets, the likelihood increases as K increases. However, the BIC value is maximized at K=2. This is exactly what we want since we want to cluster patients into two groups.