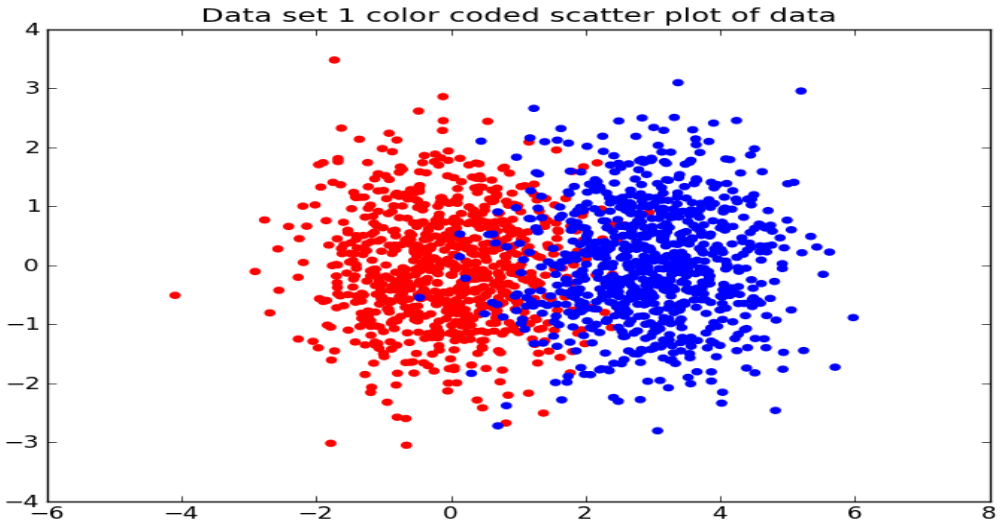
CSE 591 - Homework #5

Name: Anirudh Acharya
ASU ID: 1206293318

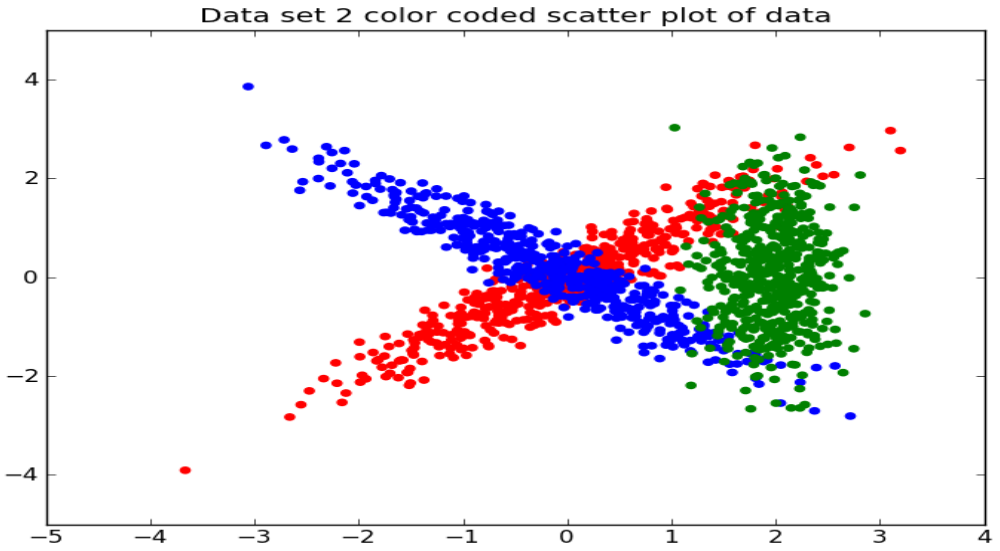
Data Visualization -

Following are the color coded scatter plots of the two datasets

Dataset 1



Dataset 2

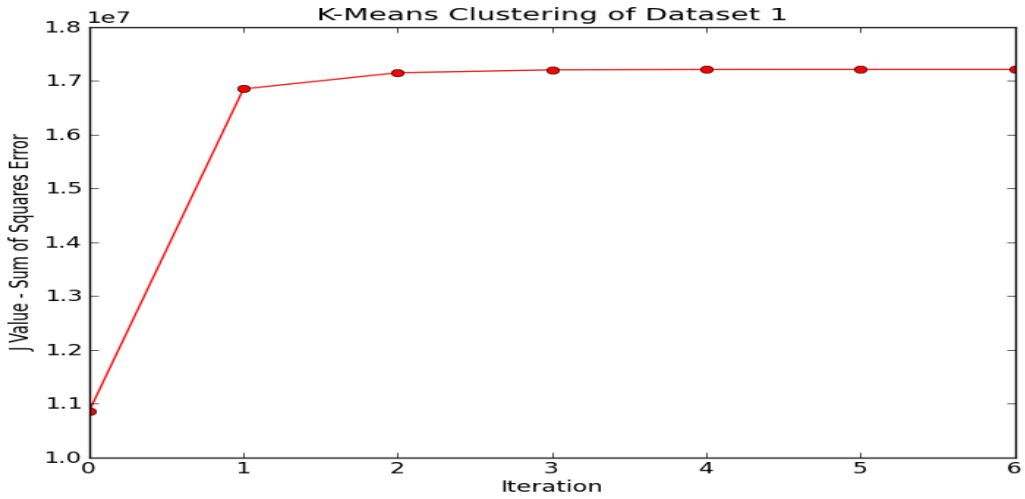


K-Means Clustering -

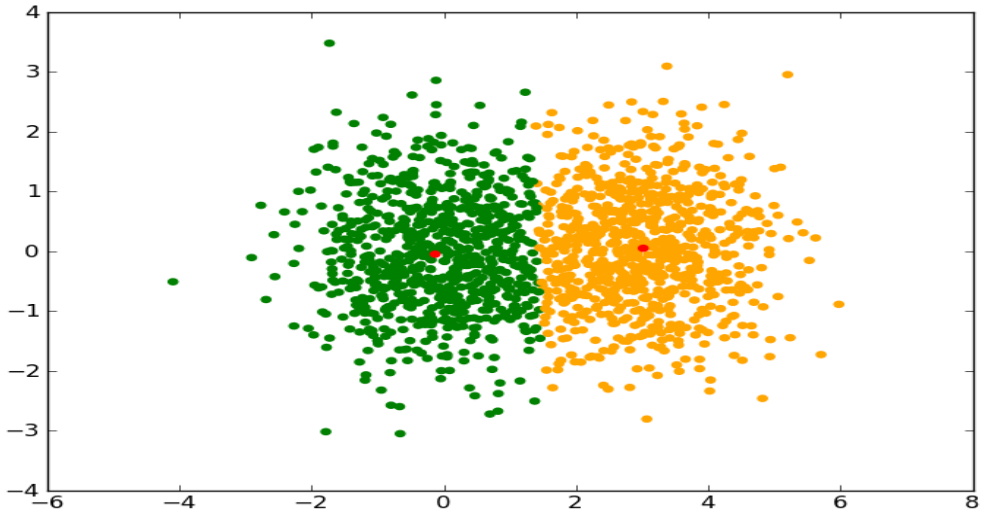
Dataset 1

Initial Cluster Center -
Cluster Center 1 - (0, 0)
Cluster Center 2 - (0, 2)
Value of K = 2

The graph of Sum of Squares Error vs Iteration



The scatter plot of the clustered data with the red dots denoting the cluster centers -

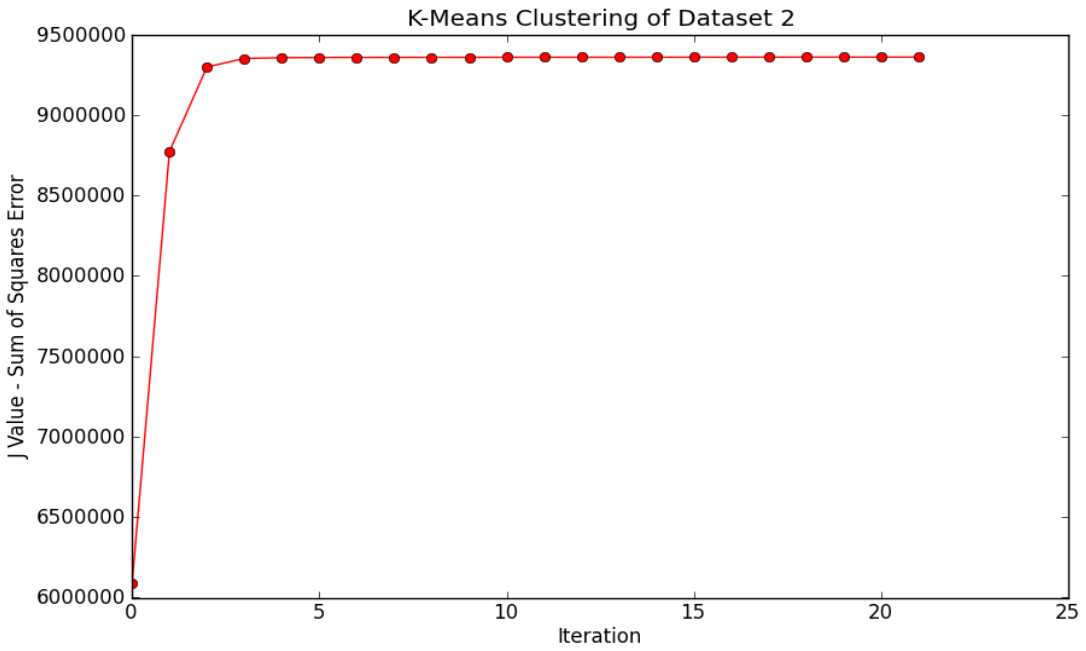


Final Cluster Centers -
Cluster Center 1 - (3.01698571, 0.04558442)
Cluster Center 2 - (-0.13084763, -0.05305959)

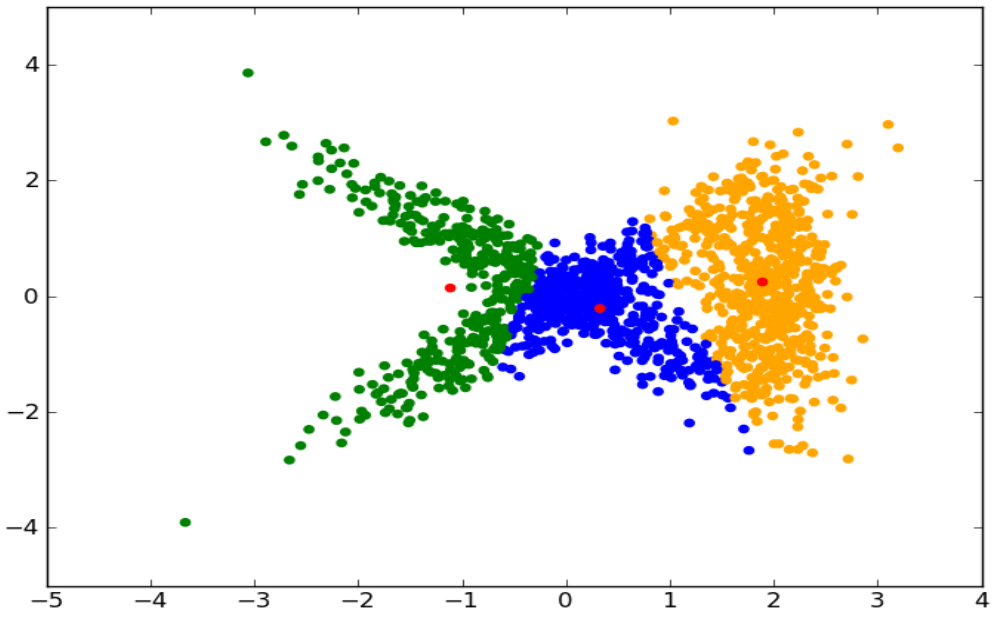
Dataset2

Initial Cluster Center -
Cluster Center 1 : (-1.862807, -1.526727)
Cluster Center 2 : (1.181673, -1.492543)
Cluster Center 3 : (-0.947083, 0.899998)
Value of K = 3

The graph of Sum of Square Errors vs Iteration



Scatter Plot of the clustered data with the red dots denoting the cluster centers.



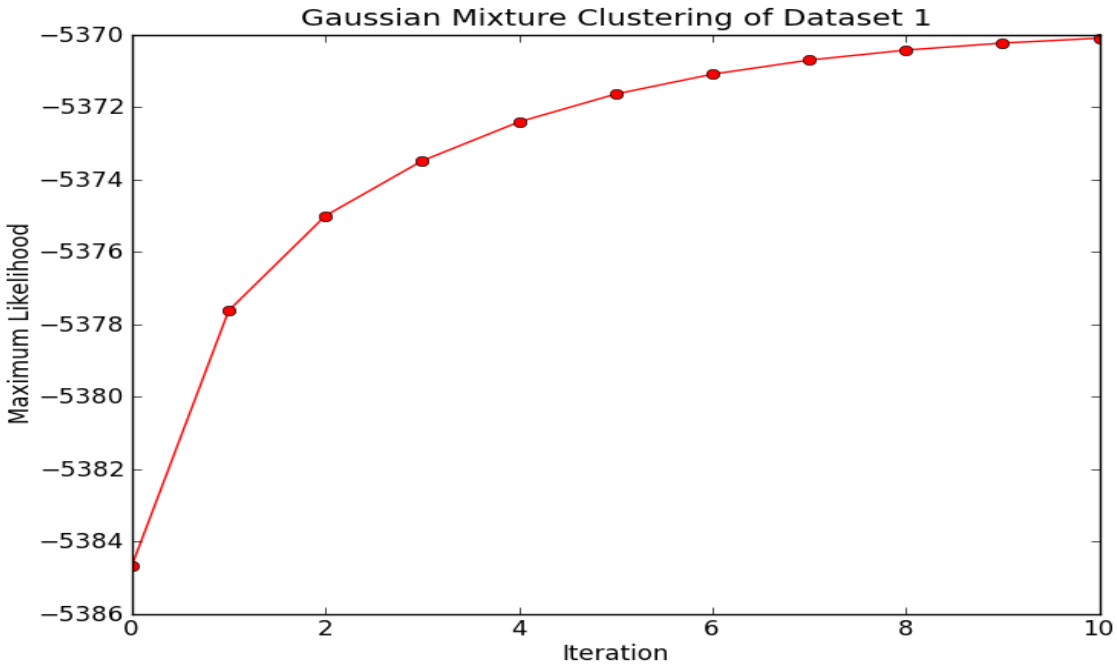
Final Cluster Centers -
Cluster Center 1 : (1.89165417, 0.24267632)
Cluster Center 2 : (-1.11387973, 0.13775981)
Cluster Center 3 : (0.32698275, -0.21492636)

Gaussian Mixture Model -

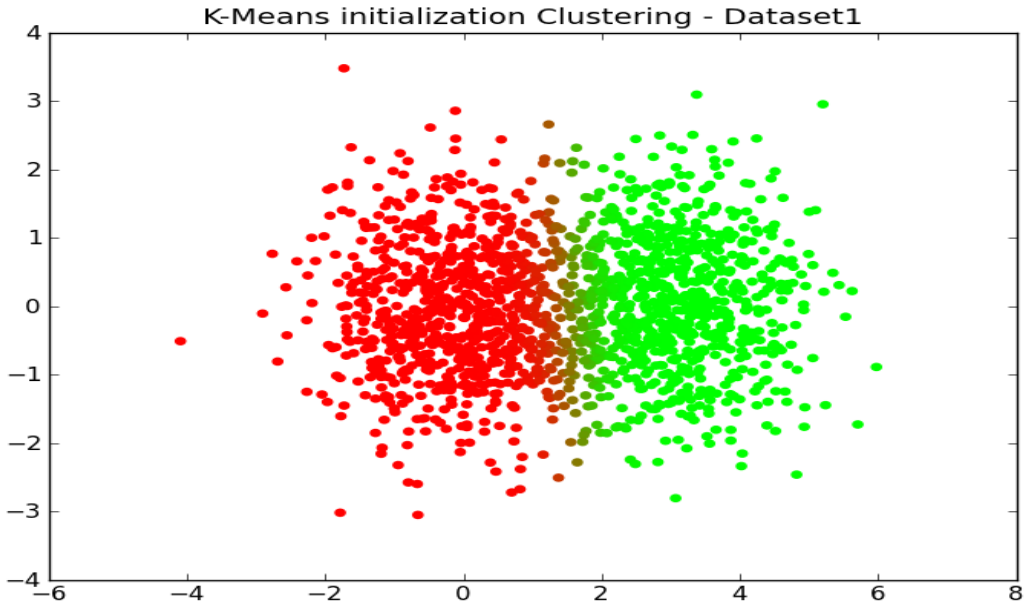
Dataset 1

Random Initial Cluster Center
Cluster Center 1 : (0.68582, -0.6702557)
Cluster Center 2 : (3.346991, -0.079277)
Value of K = 2

The graph of Maximum Likelihood vs Iteration



Scatter Plot of Clustered Data



<u>Final Cluster Center</u>

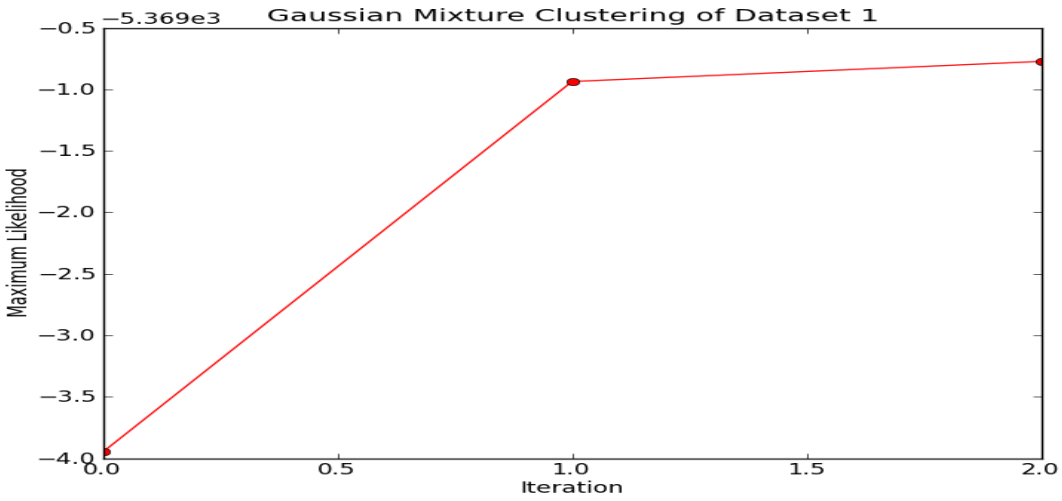
Cluster Center 1 : (-0.06410551, -0.06218497)

Cluster Center 2 : (3.02896249 , 0.05761322)

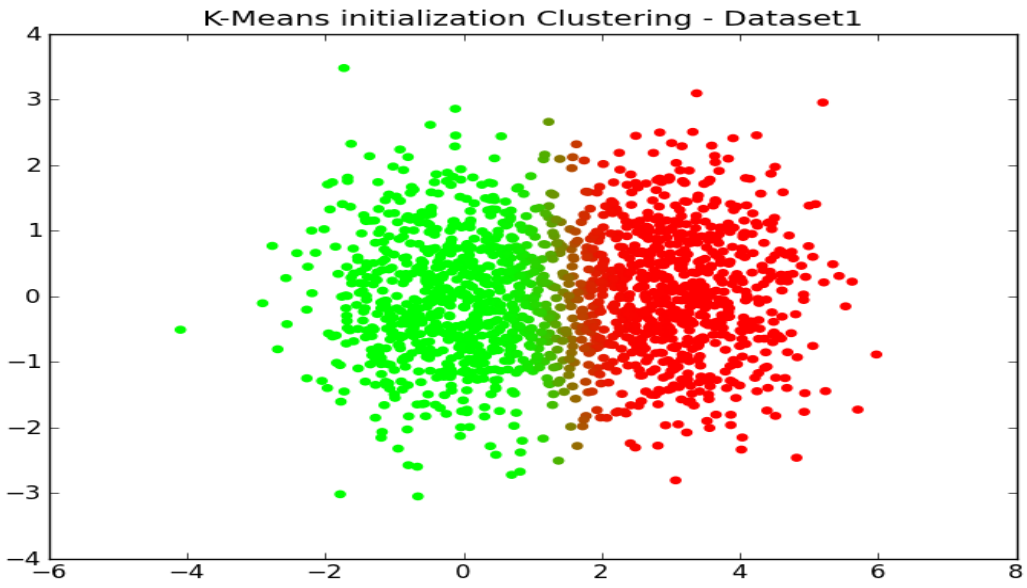
Cluster Centers Assumed from K-Means -

Initial Cluster Centers -
Cluster Center 1 : (3.01698571, 0.04558442)
Cluster Center 2 : (-0.13084763, -0.05305959)

The graph of Log Likelihood vs Iteration



Scatter Plot of the Clustered Data

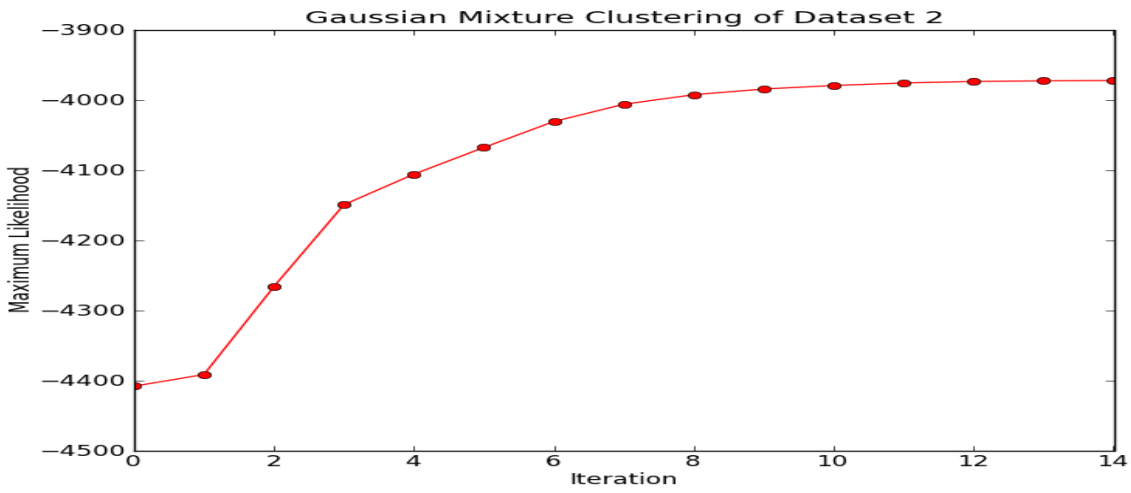


Final Cluster Center -
Cluster Center 1 : (2.99738613, 0.0528863963)
Cluster Center 2 : (-0.09726523, -0.059952553)

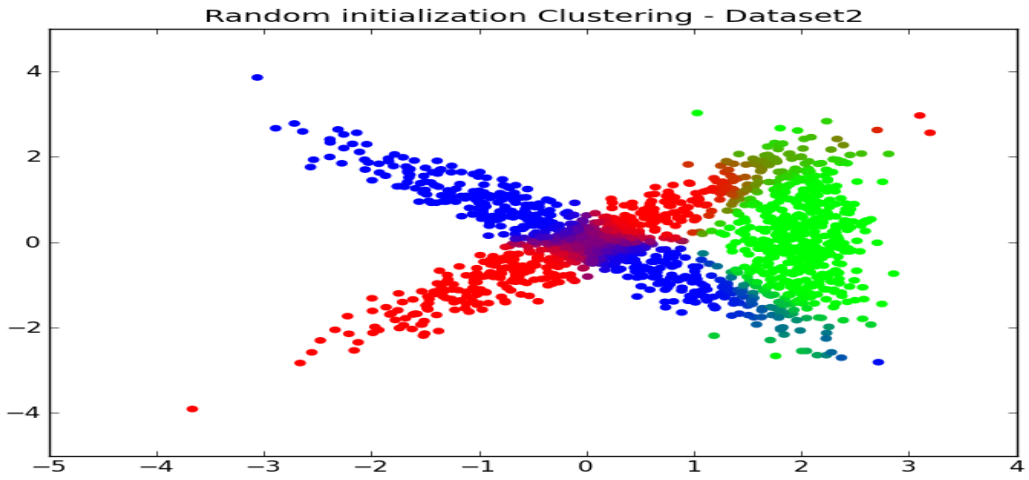
Dataset2

Random Cluster Center Assumed -
Cluster Center 1 : (-1.119466, 1.079995)
Cluster Center 2 : (1.836951, -0.067417)
Cluster Center 3 : (1.378642, -1.355791)

The graph of Log Likelihood vs Iteration



Scatter Plot of the Clustered Data -



Final Cluster Centers

Cluster Center 1 : (-0.05669648, -0.08577299)

Cluster Center 2 : (1.95924659, 0.13023854)

Cluster Center 3 : (-0.10411049, 0.11805673)

Cluster Centers Assumed from K-Means

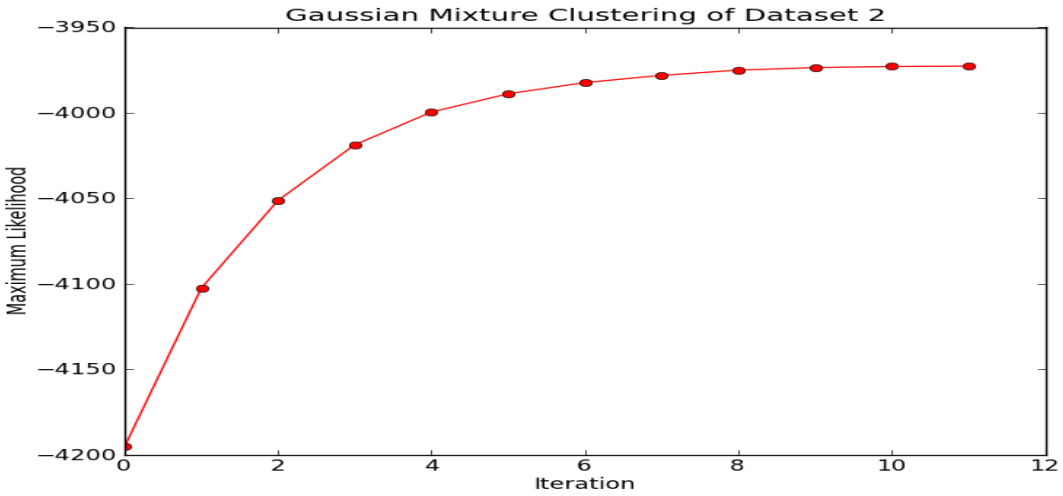
Initial Cluster Centers-

Cluster Center 1 : (1.89165417, 0.24267632)

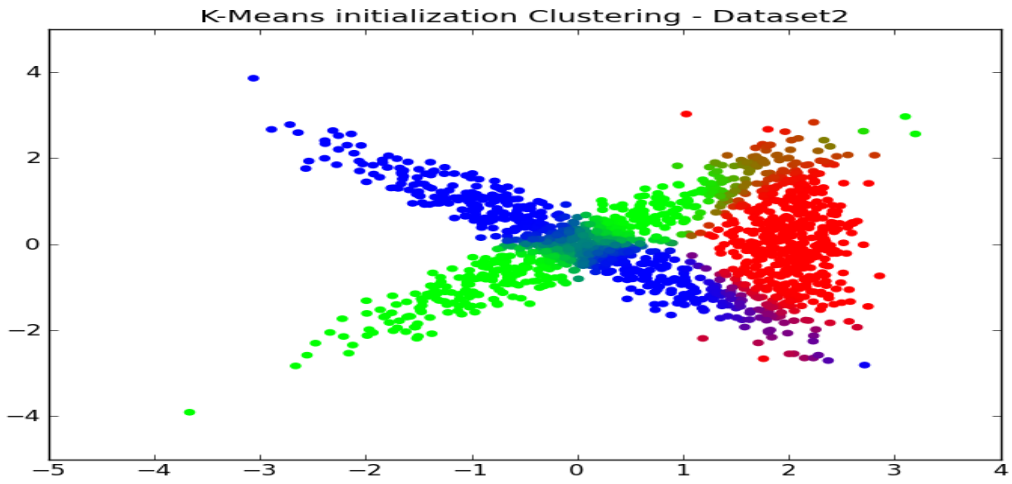
Cluster Center 2 : (-1.11387973, 0.13775981)

Cluster Center 3 : (0.32698275, -0.21492636)

The graph of Log Likelihood vs Iteration



Scatter Plot of Clustered Data



Final Cluster Centers -
Cluster Center 1: (1.95988682 0.12342337)
Cluster Center 2 : (-0.05080608, -0.08031782)
Cluster Center 3: (-0.1071640,2 0.12051856)

Comments about the initialization of the cluster centers for gaussian mixture model clustering -

The Gaussian Mixture Model Clustering requires us to initialize the values of the mean and the covariance of the component gaussians. The initialization is very important because a wrong initialization can lead the EM algorithm to converge to a local minima which will not give a very efficient clustering.

For example initializing all the covariance to the covariance of the whole dataset causes the EM algorithm to converge to a result that is not very different from the K-Means result. A good initialization of these parameters also contributes to a faster convergence of the EM algorithm. Initializing the mean and covariance values from the K-Means Clustering result gives a better result as seen from the results shown above.