Christian Hower

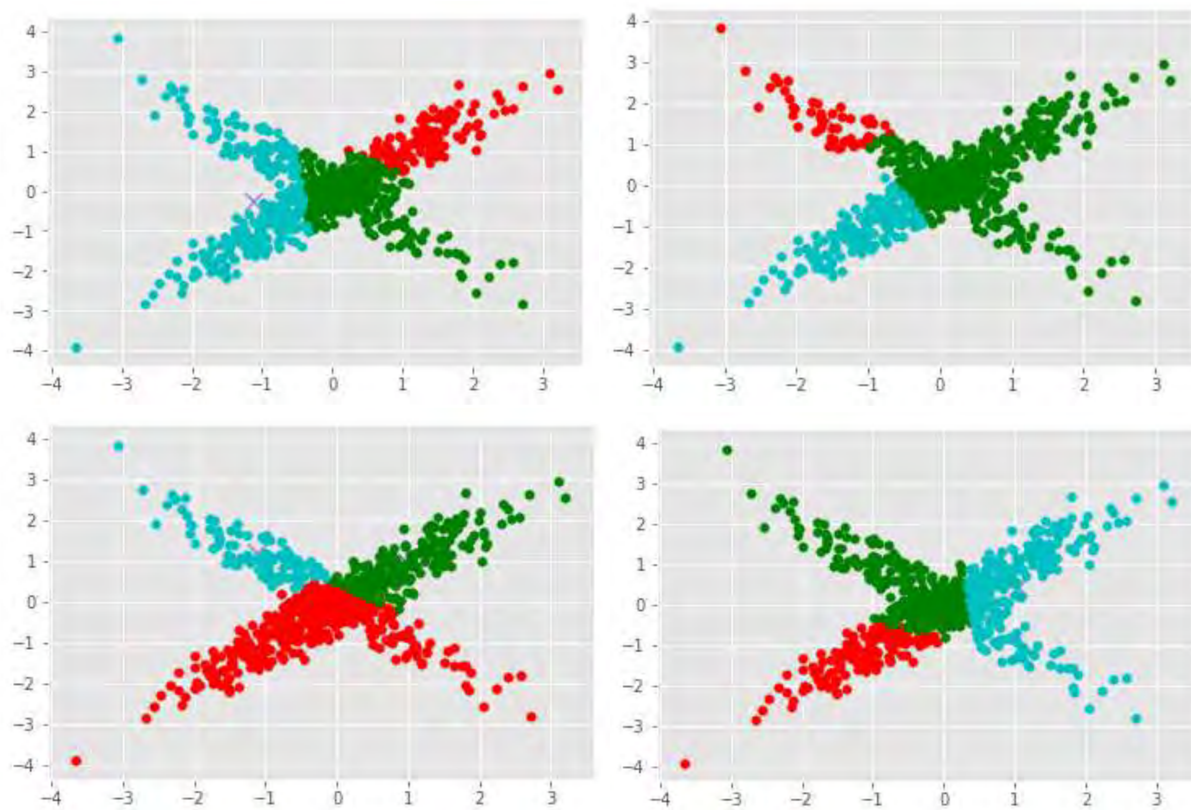
CS 546

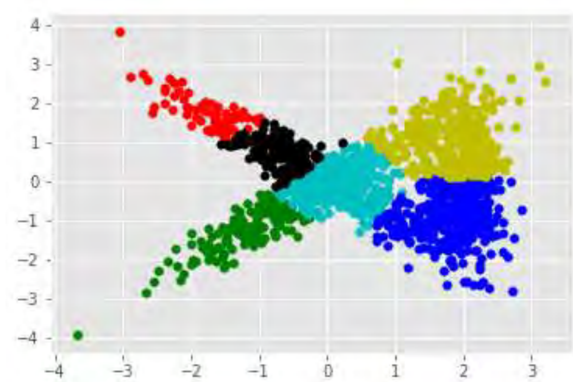
Programming 2

Kmeans & GMM

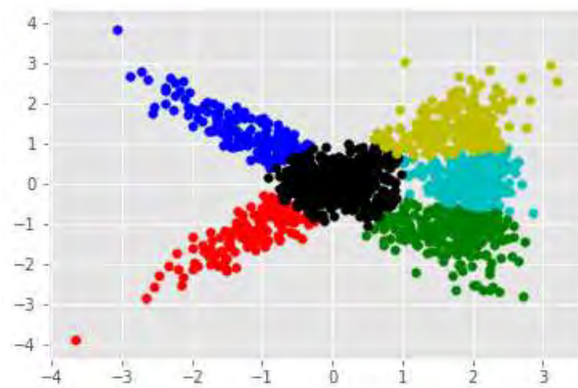
KMEANS

These plots show 5 different convergent K Means clusterings generated on these data. It is not stricly apparent what the best clustering should be. Since K means uses hard clustering areas where the groups overlap seem to be arbitrarily placed in one class or another based on the random initilization. Subjectively these goupings seem to miss some apparent continuity. Euclidean distance also plays a role in the regions with bad classification.





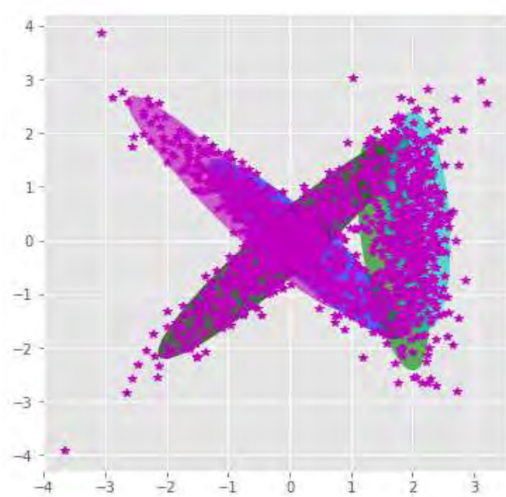
Error: 6550.32

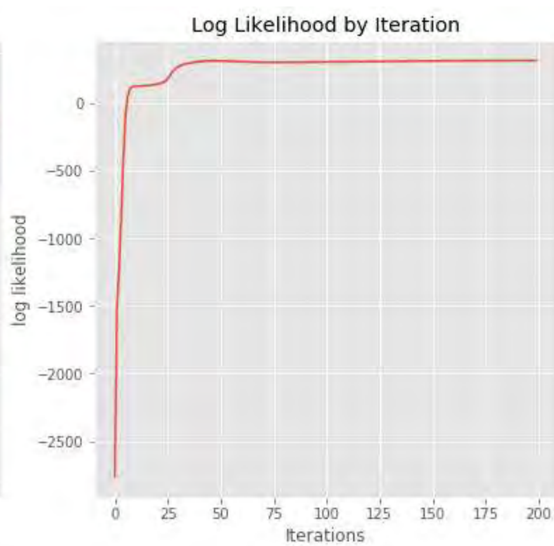


Error: 2882.22

GAUSSIAN MIXTURE MODEL

GMM offers a better approach for these data by utilizing soft clustering. Partial membership of groups allows for areas that are overlapping or otherwise unclear to be understood. The GMM method captures the structure of each distribution better in spite of the overlapping regions. The continuity that KMeans missed was preserved in this case. Another way GMM performed better at this task is that it converged to essentially the same solution each time. Which is better than K Means which had to run several times to compare which solution was best.





log likelihood: 300.09517049

mu: [[1.44886068 1.24866081]

 [1.98464409 -0.01424965]

 [-0.29603355 -0.3156568]

 [-0.77359476 0.62937279]

[0.34192933 -0.20981459]]

sigma: [array([[0.39926566, 0.33107313],

[0.33107313, 0.47213327]]), array([[0.09477024, -0.00836294],

[-0.00836294, 0.98852416]]), array([[0.77921667, 0.75334168],

[0.75334168, 0.83961981]]), array([[0.82282535, -0.85728359],

[-0.85728359, 0.98631687]]), array([[0.69224802, -0.66854288],

[-0.66854288, 0.75641709]])]