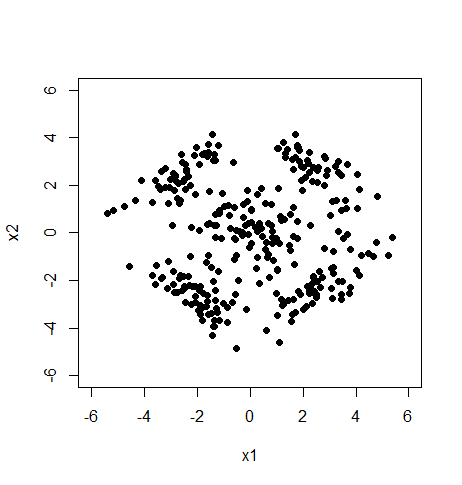
Eda Gür

50488

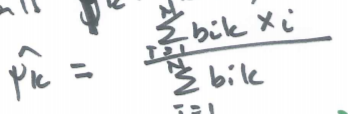
Comp421 HW6 Report

I started the project by defining the means, covariances and the class sizes in the given\_means, given\_covs and class\_sizes variables. I also defined a colors matrix and an empty data matrix with 2 coulumns. I also defined N as the total number of samples which is 300 and K which is 5 clusters.

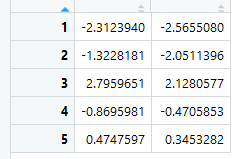
Then, I started generating the (x,y) pairs from the bivariate gaussian distributions with the mvrnorm given my means, covariances and class sizes and populated my data matrix X withthese values. After that I plotted X and the result is below:



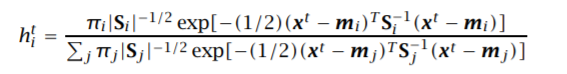
Then, I initialized my 5 centroids with X[sample(1:N, K),]. After that, I created a matrix D of dimensions K\*N. I populated this matrix with the Euclidian distances between each centroid and data point. Then, I picked the closest centroid to each data point. From this, I got an assignments vectors of size N which holds which data point maps to which centroid. I recalculated the centroids with the formula we have seen in class:



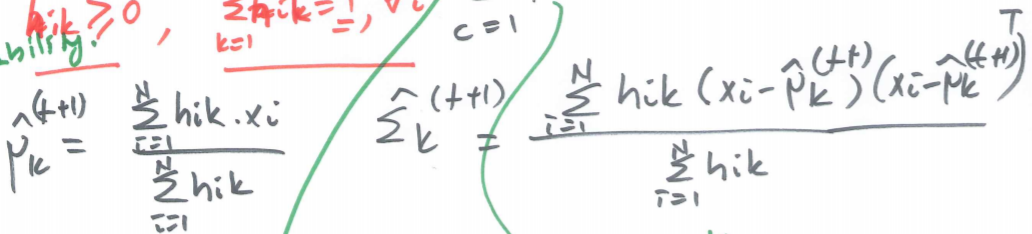
This is my first iteration of k-clusters algorithm. I do the same for a second time and I have the centroids I can use in my EM algorithm. My initial centroids in the EM after the k-means algorithm are below:

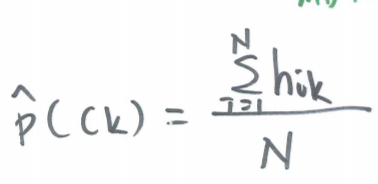


After that, for the gaussian component, I use the following formula:

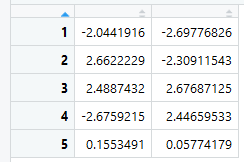


I calculate it for each cluster and data point. Then, I recalculate the prior probabilities, means and covariances with the formulas we have seen in class:





I repeat this procedure 100 times and I end up with the mean vectors:



I classify the points as to which centroid’s cluster they belong to. After that, I color them accordingly and draw the points, given Gaussian distributions and the calculated distributions and below is my result:

