COMP 421 – INTRODUCTION TO MACHINE LEARNING – FALL 2017
HW07 – Expectation-Maximization Clustering
Zafer Çavdar, 49995

1- Generate data

Using *mvrnorm* function in R with given class means, class covariances and class sizes, I generated bivariate Gaussian data points.

2- Initialization with k-means

I sampled k=5 datapoints randomly as initial centroids and calculated the assignments according to distances. According to new assignments (cluster memberships), I calculated new centroids. One more time, I calculated assignments from centroids and new centroids from new assignments. Last assignments I found will be used in EM clustering as initial means in the first iteration.

3- Estimations before EM iteration

Before starting EM iterations, I calculated prior probabilities (number of assignments in class c divided bynumber of all datapoints) and covariance matrices according to initial clusters.

4- Implementing EM algorithm and running it 100 iterations

To implement EM algorithm, firstly I defined posterior probability that x^i is generated by G_c .

$$h^i{}_c = \frac{(c)|S_c|^{-1/2} exp(-0.5*(x^i-m_c)^T S_c{}^{-1}(x^i-m_c)}{\sum_j |S_j|^{-1/2} exp(-0.5*(x^i-m_j)^T S_j{}^{-1}(x^i-m_j)}$$

where ***Π***(c) is prior, S(c) is covariance and m(c) is mean for class c,

After calculating $h_c{}^i$, I updated priors, then means and then covariances respectively with following rules:

$$\Pi_c = \frac{\sum_i h_c{}^i}{N}$$

$$m_c = \frac{\sum_i h_c{}^i x^i}{\sum_i h_c{}^i}$$

$$S_c = \frac{\sum_i h_c{}^i (x^i-m_c)(x^i-m_c)^T}{\sum_i h_c{}^i}$$

With these update rules, I calculated these equations 100 times and got the means after 100 iterations.

5- Plotting clustering results and drawing ellipses

With the last calculated values of means, I calculated distances and assignments for all datapoints and plotted them with different color. Since we have class covarinces and class means provided in the homework description, I draw ellipses to show original Gaussian densities with dashed lines. Also gaussian densities that my EM algoritm found is represented on the plot with solid lined ellipses. Assignments and densities I found are very similar to expected output.