COMP/INDR 421/521 HW07: Expectation-Maximization Clustering

Deadline: December 22, 2017, 11:59 PM

In this homework, you will implement an expectation-maximization (EM) clustering algorithm in R, Matlab, or Python. Here are the steps you need to follow:

1. Generate random data points from five bivariate Gaussian densities with the following parameters:

$$\mu_1 = \begin{bmatrix} +2.5 \\ +2.5 \end{bmatrix}, \quad \Sigma_1 = \begin{bmatrix} +0.8 & -0.6 \\ -0.6 & +0.8 \end{bmatrix}, \quad N_1 = 50$$

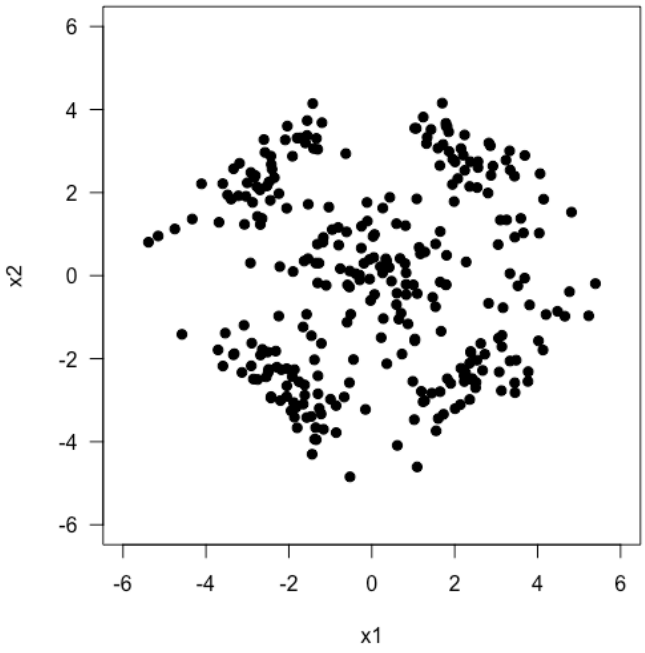
$$\mu_2 = \begin{bmatrix} -2.5 \\ +2.5 \end{bmatrix}, \quad \Sigma_2 = \begin{bmatrix} +0.8 & +0.6 \\ +0.6 & +0.8 \end{bmatrix}, \quad N_2 = 50$$

$$\mu_3 = \begin{bmatrix} -2.5 \\ -2.5 \end{bmatrix}, \quad \Sigma_3 = \begin{bmatrix} +0.8 & -0.6 \\ -0.6 & +0.8 \end{bmatrix}, \quad N_3 = 50$$

$$\mu_4 = \begin{bmatrix} +2.5 \\ -2.5 \end{bmatrix}, \quad \Sigma_4 = \begin{bmatrix} +0.8 & +0.6 \\ +0.6 & +0.8 \end{bmatrix}, \quad N_4 = 50$$

$$\mu_5 = \begin{bmatrix} +0.0 \\ +0.0 \end{bmatrix}, \quad \Sigma_5 = \begin{bmatrix} +1.6 & +0.0 \\ +0.0 & +1.6 \end{bmatrix}, \quad N_5 = 100$$

   Your data points should be similar to the following figure.



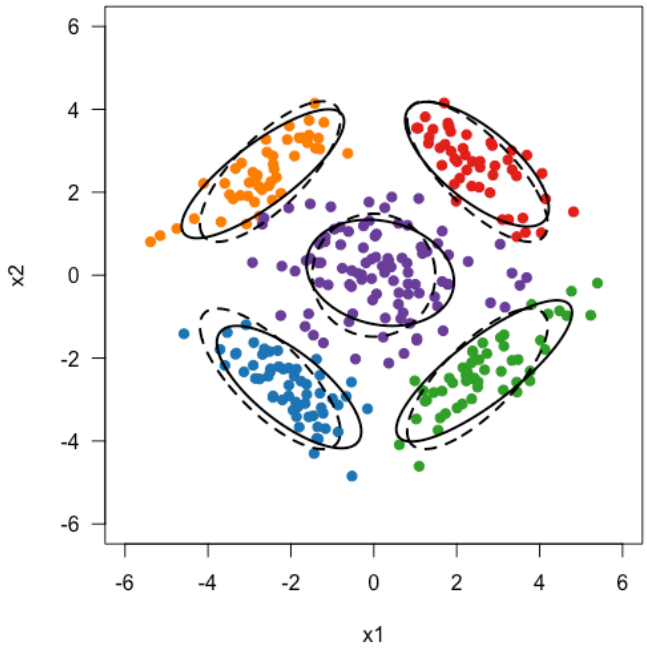
2. To initialize your EM algorithm from a good initial solution, run k-means clustering algorithm with $k = 5$ just for two iterations.

3. After running k-means clustering algorithm for two iterations, take centroids as the initial values for the mean vectors in your EM algorithm. Using the data points covered by each center, estimate the initial covariance matrices and prior probabilities in your EM algorithm.

4. After the initialization step, run your EM algorithm for 100 iterations. Report the mean vectors your EM algorithm finds. Your results should be similar to the following matrix.

```
##              [,1]        [,2]
##  [1,] -2.0441920 -2.69776844
##  [2,]  2.6622246 -2.30911081
##  [3,]  2.4887435  2.67687075
##  [4,] -2.6759195  2.44658904
##  [5,]  0.1553517  0.05773829
```

5. Draw the clustering result obtained by your EM algorithm by coloring each cluster with a different color. You should also draw the original Gaussian densities you use to generate data points and the Gaussian densities your EM algorithm finds with dashed and solid lines, respectively. Draw these Gaussian densities where their values are equal to 0.05. Your figure should be similar to the following figure.



What to submit: You need to submit your source code in a single file (.R file if you are using R, .m file if you are using Matlab, or .py file if you are using Python) and a short report explaining your approach (.doc, .docx, or .pdf file). You will put these two files in a single zip file named as ***STUDENTID.zip***, where ***STUDENTID*** should be replaced with your 7-digit student number.

How to submit: E-mail the zip file you created to mehmetgonen@ku.edu.tr with the subject line ***Intro2MachineLearningHW07***. Please follow the exact style mentioned for the subject line and do not send a zip file named as ***STUDENTID.zip***. Submissions that do not follow these guidelines will not be graded.

Late submission policy: Late submissions will not be graded.

Cheating policy: Very similar submissions will not be graded.