CS 274A Homework 5

Zachary DeStefano, 15247592

Due Date: Wednesday March 5th

What to Submit

- Submit your MATLAB code to the Homework 5 dropbox in EEE, including a script that loads each data set and runs your code on each data set. Please upload all of your functions in a single compressed file (please use .gz or zip format). If you are using something other than MATLAB please also provide a README file to explain to us how to use your code.
- Submit a written hardcopy summary of your results in class with the information below.

Please try to put multiple plots on the same page using "subplots" in MATLAB (where different pages could correspond to different data sets and different algorithms), to avoid printing lots of pages. For each data set, using the true K value for each one, show the following

1. The K-means solution (scatter plot in two dimensions (any two dimensions for Data Set 4) illustrating the location of the solution (i.e., the cluster means), and plotting the data from different clusters with different symbols (and/or in color if you would like to use color).
2. A plot of the sum-squared-error (divided by n) as a function of iteration number in the K-means algorithm.
3. The initial parameter values and the final parameter values (2 plots, each showing means and covariances for each cluster) for the EM/Gaussian mixtures code for the highest-likelihood solution. You can use or modify the code from Homework 1 to do the plotting.
4. A plot of the log-likelihood (for one run of your algorithm) as a function of iteration number during EM.
5. Add some brief comments (1 paragraph) on the difference between K-means and EM for each data set.
6. Generate a table of log-likelihood and BIC scores for K going from $K = 1$ to some maximum value (e.g., $K = 5$). Comment briefly on the results.