In-class Assignment # 3

Due: Saturday, March 3, 2018, 11:59 p.m.
Total Points: 50

Question 1. [50 POINTS]

This exercise will give you some practice coding the EM algorithm, where we will apply the mixture of Gaussian technique to one-dimensional data. Be sure to include all code and plots in your assignment submission. **You must only use your own code to solve this problem. Do not download code or use pre-defined libraries or packages.**

(a) Create a program to estimate the means, standard deviations, and weights of a mixture of Gaussians via the EM algorithm. You will probably want to have three functions: one that performs the expectation step, one that performs the maximization step, and one that runs the outer loop.

(b) You will find a data set on Canvas (e.g. data1.txt) that was created with 3 Gaussians. Run your EM algorithm with three randomly initialized means, and standard deviations, but with equiprobable weights. Give the resulting means, standard deviations, and weights (that is, the model parameters). Plot the log likelihood of the data as it changes during training (i.e. plot log-likelihood against the number of iterations of EM so far). Use a stopping threshold of 0.001.

The log-likelihood for a Mixture of Gaussians is as follows:

$$L = \sum_{j=1}^{N} \log\left(P(x_j)\right) = \sum_{j=1}^{N} \log\left(\sum_{i=1}^{I} P(c_i)\mathcal{N}(x_j|\mu_i, \sigma_i)\right) \tag{1}$$

where N is the number of data points, I is the number of Gaussian clusters, and x_j represents the j^{th} data point. Note that $\mathcal{N}(x_j|\mu_i, \sigma_i)$ can be replaced with the equation for a Normal/Gaussian distribution with mean μ_i and standard deviation σ_i.

(c) Repeat part (b), but this time use 2 then 4 Gaussians. Plot the log likelihood when 2 and 4 Gaussians are used, respectively. Display the parameters (e.g. μ_i's, σ_i's, weights) of the model to the screen.