Assigned: 31 October 2019

Project #7 – Expectation Maximization

EE 511: Fall 2019

Due: Thursday, 14 November 2019 at 16:00. Late penalty: 15% per day before 16 November at 16:00.

1. Implement a random number generator for a random vector $X = [X_1, X_2, X_3]^T$ having multivariate Gaussian distribution with

$$\mu = \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 3 & -1 & 1 \\ -1 & 5 & 3 \\ 1 & 3 & 4 \end{bmatrix}.$$

2. Implement a random number generator for a random variable with the following mixture distribution: $f(x) = 0.4N(-1,1) + 0.6N(1,1)$. Generate a histogram and overlay the theoretical p.d.f. of the random variable.

3. Implement a 2-dimensional random number generator for a Gaussian mixture model (GMM) pdf with 2 sub-populations. Use the expectation maximization (EM) algorithm to estimate the pdf parameters of the 2-D GMM from samples. Compare the quality and speed of your GMM-EM estimates using 300 samples from different GMM distributions (e.g. spherical vs ellipsoidal covariance, close vs well-separated subpopulations, etc.).

4. A geyser is a hot spring characterized by an intermittent discharge of water and steam. Old Faithful is a famous cone geyser in Yellowstone National Park, Wyoming. It has a predictable geothermal discharge and since 2000 it has erupted every 44 to 125 minutes. Refer to the addendum data file that contains waiting times and the durations for 272 eruptions.
   a. Generate a 2-D scatter plot of the data. Run a k-means clustering routine on the data for $k = 2$. Show the two clusters on a scatterplot.
   b. Use a GMM-EM algorithm to fit the dataset to a GMM pdf. Draw a contour plot of your final GMM pdf. Overlay the contour plot with a scatterplot of the data set. How can you use the GMM pdf estimates to cluster the data?