CSECE 8735 Fall 2015
Unsupervised Learning

Assignment 1
due Thursday 9/17/2015

Problem 1

(1) Implement the EM algorithm for Gaussian mixture density (GMD) parameter estimation by using MATLAB (note that the textbook has some useful MATLAB codes in the exercise section, see page 79-83).

(2) Use your code to estimate the GMD parameters for the dataset GMD based on the assumption that the GMD has four component Gaussian pdfs, with each having a full covariance matrix. Terminate your EM estimation at the 100th iteration, i.e., use $\theta^{(100)}$ as the estimate for the model. The following two initialization methods are to be used in your experiments:

a) Specified $\theta^{(0)}$ and termination:

$$\pi_1^{(0)} = \pi_2^{(0)} = \pi_3^{(0)} = \pi_4^{(0)} = 1/4,$$
$$\mu_1^{(0)} = [10 \quad 2]^T, \mu_2^{(0)} = [5 \quad 6]^T, \mu_3^{(0)} = [0 \quad 1]^T, \mu_4^{(0)} = [4 \quad 3]^T,$$

$$\Sigma_1^{(0)} = \Sigma_2^{(0)} = \Sigma_3^{(0)} = \Sigma_4^{(0)} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix};$$

b) Your choice:
Use your own judgement to initialize the EM procedure with a set of parameter values different from a).

(3) Include the following items in your report:

a) Your code (with clear comments);

b) The estimated model parameters (mean vectors, covariance matrices, and mixture weights) produced in the final iteration for the initializations a) and b), respectively (include your initialization parameters in the report).

c) For the specified initialization a), compute the log likelihood scores of the observed data $\log p(X \mid \theta^{(t)})$ for $t = 1, 2, \cdots, 100$, and show in a plot the function values vs. t for the specified range of iterations .

d) For the specified initialization a), use the maximum posterior probability rule to assign each data sample x_n to one of the four clusters $k = 1,2,3,4$:

$$k^* = \arg\max_{1 \le k \le 4} P\left(z_n = k \mid x_n; \theta^{(t)}\right),$$

with $t = 10, 50, 100$. Color-code the data samples and make a 2D plot, i.e., plot the data samples assigned to the 1st, 2nd, 3rd and the 4th clusters by using the colors of blue, green, red, and magenta, respectively. (note: you need to do three plots for the three specified t values).

Problem 2

(1) For the coin-tossing example discussed in class, implement the EM algorithm in MATLAB.
(2) Run the EM procedure with the initialization of $\theta^{(0)} = \left(\theta_A^{(0)},\ \theta_B^{(0)}\right) = (0.6,\ 0.4)$ for 10 iterations.
(3) Include in your report the following items:
  a) Your MATLAB code (with clear comments)
  b) A table containing the parameter estimates $\theta^{(t)} = \left(\theta_A^{(t)},\ \theta_B^{(t)}\right)$ for $t = 1, 2, \cdots, 10$.
  c) The posterior probabilities $P\left(z^n = A \mid x^n; \theta^{(t)}\right)$ and $P\left(z^n = B \mid x^n; \theta^{(t)}\right)$, at $t = 1$ and $t = 10$.
  d) The log probabilities $P\left(X; \theta^{(t)}\right)$ for $t = 1, 2, \cdots, 10$.

Problem 3

Assume that in a 2-dimensional Gaussian mixture density model all the component Gaussian pdfs share the same covariance matrix

$$\Sigma_1 = \Sigma_2 = \cdots = \Sigma_K = \Sigma$$

Derive the EM estimation equations for the GMD parameters π_k, $\mu_k = \left[\mu_{k,1}\ \ \mu_{k,2}\right]^T$, $k = 1, 2, \cdots, K$, as well as Σ.

This assignment is complete.