ECEN 758 Data Mining and Analysis

Assignment 3: due 11:59pm, Monday October 22, 2017

Procedure: Please Read

Please follow these guidelines to ensure your solutions reach me, and help me attribute your marks correctly

- *Format*: solutions must be typeset (using e.g. Microsoft Word or LaTex) and rendered in pdf.
- *Transmittal*: email your pdf solutions to me at duffieldng AT tamu DOT edu using the required subject line for the assignment: "DMA Assignment n" where is the number of the assignment (1,2,3, etc).
- *File name*: use file name DMA-n-UIN.pdf where n is the number of the assignment (1,2,3, etc), and UIN is your UIN.
- *Identification*: please include your name and UIN near the top of the first page of your solutions.
- *Numerical Computations*: you may use packages or write code etc. to do the numerical computations. You must include function calls or your code in your solutions.
- *Algebraic Computations:* You must include your derivation to receive full credit.

1 K-means: [25 marks]

	X_1	X_2
$\boldsymbol{x}_1$	0	2
$\boldsymbol{x}_2$	0	0
$\boldsymbol{x}_3$	1.5	0
$\boldsymbol{x}_4$	5	0
$\boldsymbol{x}_5$	5	2

Data for Question 1

For the two-dimensional points in the table below, assume $k = 2$ clusters initially assign as $C_1 = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_4\}$ and $C_2 = \{\boldsymbol{x}_3, \boldsymbol{x}_5\}$. Apply the K-means algorithm until convergence, i.e., until the clusters do not change, using usual Euclidean distance $\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2 = (\sum_{a=1,2} |x_{i,a} - x_{j,a}|^2)^{1/2}$.

2 Gaussian Mixture Models : 37 Marks

	X_1	X_2
$\boldsymbol{x}_1$	0.5	4.5
$\boldsymbol{x}_2$	2.2	1.5
$\boldsymbol{x}_3$	3.9	3.5
$\boldsymbol{x}_4$	2.1	1.9
$\boldsymbol{x}_5$	0.5	3.2
$\boldsymbol{x}_6$	0.8	4.3
$\boldsymbol{x}_7$	2.7	1.1
$\boldsymbol{x}_8$	2.5	3.5
$\boldsymbol{x}_9$	2.8	3.9
$\boldsymbol{x}_{10}$	0.1	4.1

Data for Question 2

$\boldsymbol{x}_1, \ldots \boldsymbol{x}_{10}$ are ten data point with two attributes: see the table below. This question will use three Gaussian clusters with initial means $\mu_1 = (0.5, 4.5)^T$, $\mu_2 = (2.2, 1.6)^T$ and $\mu_3 = (3, 3.5)^T$, initial covariance matrices $\Sigma_1 = \Sigma_2 = \Sigma_3 = \{\{1, 0\}, \{0, 1\}\}$ and initial mixture probabilities $\mathbb{P}(C_1) = \mathbb{P}(C_2) = \mathbb{P}(C_3) = 1/3$.

In the following parts (A), (C) and (D), quote the relevant general formulae, then apply it to the data.

(A) Compute the first EM iterates of the cluster means.

(B) Show the data on a scatter plot, together with the initial and iterated means. Comment on your answer.

(C) Compute the first EM iterates of the mixture probabilities.

(D) Compute the first iterates of the covariance matrices for the three clusters.