STAT 598Z: Homework 6

Due: 16th April 2013

1. This homework will contribute 10 points towards your final score.
2. Attempt as many problems as possible.
3. Only neatly handwritten solutions will be accepted. Alternatively you may use LaTeX to typeset your solutions.
4. Hand in your HW (including print outs of your source code) at the beginning of the class on 16th April 2013. Additionally source code (if any) should be emailed to `stat598z@gmail.com` **before** the assignments are submitted in the class. No late submissions will be accepted!
5. Program files should be named after the problem (e.g. solution to problem 1 should be problem1.py etc).
6. Remember to seed your random number generators!

Problem 1 (10 pt) Draw 20,000 samples from a mixture of 2 dimensional Gaussian distributions with the following parameters

Mean	Variance	Proportion
$\begin{pmatrix} 0 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$	0.7
$\begin{pmatrix} 1 \\ 0 \end{pmatrix}$	$\begin{pmatrix} 0.1 & 0 \\ 0 & 1 \end{pmatrix}$	0.1
$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}$	0.2

1. Write a Python program of for K-Means clustering. Set $K = 2$. Plot $J(r, \mu)$ as a function of the number of iterations. Visually represent the results of the produced clustering. Report what happens when you change the initial seed.
2. Report the minimum value of $J(r, \mu)$ you obtain for $K = 2, \ldots, 6$. Which one has the lowest $J(r, \mu)$? Visually represent the clustering result for each K. Which K do you think is the best fit amongst $K = 2, \ldots, 6$? Please explain your reasons.
3. Let $K = 3$, initialize the algorithm with 10 different random seeds. Check if all the results are similar given different initialization. Briefly explain.

4. Vary the separation between the classes by varying the mean values and perform K-Means clustering with $K = 3$. Comment on your results.
5. Vary the spread of the clusters by varying the variance values and perform K-Means clustering with $K = 3$. Comment on your results.
6. Write a Python program for the Gaussian Mixture Model (GMM). Visually represent the clustering assignments produced by the GMM algorithm (Hint: Think about how to visually represent the GMM clusters).