Statistics 5525: Homework 3

For each homework assignment, turn in at the beginning of class on the indicated due date. Late assignments will only be accepted with special permission. Write each problem up *very* neatly (LaTeX is preferred). Show all of your work.

Problem 1

Consider the p-dimensional Gaussian Mixture Model:

$$x_i \sim \sum_k^K \pi_k p(x|\mu_k, \Sigma_k, C_k), \qquad \text{for } i = 1, \ldots, N, \tag{1}$$

where $p(x|\mu_k, \Sigma_k) = \frac{1}{\pi^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)'\Sigma_k^{-1}(x-\mu_k)}$.

In the special case where $\Sigma_k = \sigma I \; \forall k = 1 \ldots, K$, discuss the connections between the K-means algorithm and EM for fitting model (1). Additionally, show that as $\sigma \to 0$ the two methods coincide.

Problem 2

Download "ClusterSet1.txt" from the course webpage. Apply the k-means clustering procedure to this data set. You may code this up from scratch or use the built in functions in either R or Matlab. Discuss how you selected 'K', and why you believe it is correct.

Problem 3

Recall that the EM algorithm for fitting model (1) iterates over the following updates: For $t = 1, \ldots, T$

1. $\pi_{i,k}^{(t)} = p(x_i \in C_k | \mu_k^{(t-1)}, \Sigma_k^{(t-1)})$
2. $\mu_k^{(t)} = \sum_{i=1}^N \pi_{i,k}^{(t)} x_i / \sum_{i=1}^N \pi_{i,k}^{(t)}$
3. $\Sigma_k^{(t)} = \sum_{i=1}^N \pi_{i,k}^{(t)} (x_i - \mu_k^{(t)})(x_i - \mu_k^{(t)})' / \sum_{i=1}^N \pi_{i,k}^{(t)}$

part a

Given $\pi_{i,k}$s, show that the M.L.E for μ_ks are given by $\sum_{i=1}^N \pi_{i,k} x_i / \sum_{i=1}^N \pi_{i,k}$.

part b

Given μ_ks and $\pi_{i,k}$s, show that the M.L.E for Σ_ks are given by $\sum_{i=1}^N \pi_{i,k}(x_i - \mu_k)(x_i - \mu_k)' / \sum_{i=1}^N \pi_{i,k}$ (Given μ_ks).

part c

Implement the EM algorithm and using 'ClusterSet1.txt', compare results to those found in Problem 2.

Problem 4

Using a hierarchical clustering method with 'ClusterSet1.txt', compare results to those found in Problem 2 and 3. Show dendrograms, and discuss the distance function you settled on for your link function.

Problem 5 (old problem 6)

Download "ClusterSet2.txt" from the course webpage. Using any method you find appropriate, determine the number of clusters and and assignment labels for each data point.