STAT 440

Homework #6
Due March 19th by midnight

Goal: EM Algorithm

For this homework, submit your R code for this assignment electronically via Canvas. Note the deadline on Canvas. In addition to the R code, you must also turn in your typed-in answers. **Please submit your R code and the typed up answers as separate documents**. Points will be deducted if R code is very poorly formatted.

1. For this problem, you will be deriving and implementing the EM algorithm for a Gaussian Mixture Model. Each part will guide you through a step in implementing the EM algorithm. A Gaussian Mixture Model is where the data Y comes from one of several normal distributions with probabilities π_k, i.e.

$$Y_i \sim N(\mu_k, \sigma_k^2) \tag{1}$$

for groups k in $1, \ldots, K$. We can write the probability that $Y_i = y$ as

$$P(Y_i = y) = \sum_{k=1}^{K} \pi_k P(Y_i = y | Z_i = k) = \sum_{k=1}^{K} \pi_k N(Y_i; \mu_k, \sigma_k^2) \tag{2}$$

where Z_i denotes what group that Y_i comes from. We are interested in estimating $\theta = (\mu_k, \sigma_k^2, \pi_k)$.

This problem is well-known and there are extensive online notes detailing solutions. You may use these online resources, but you must understand what happens at each step. That is why homework will be graded primarily on the amount of work you show and your explanations. Show ALL steps in your derivation. This homework is less about the coding and more about the theory, since the hard part of EM algorithm is the theoretical component.

   (a) Write down the joint likelihood of $Y_1, \ldots, Y_n$. This is the observed likelihood, since we are missing our latent variables Z_i. Explain why this joint likelihood is difficult to maximize directly with respect to θ.

   (b) Now assume we know Z_i. Write down the complete-data likelihood and the corresponding complete-data log-likelihood.

(c) Now take the conditional expectation to find $Q(\theta|\theta^{(t)})$. There will be an expectation that depends on $\mu_k^{(t)}$, $\sigma_k^{2(t)}$ and $\pi_k^{(t)}$, which you can denote $\gamma_j(x_i)$ for $i \in 1, \ldots, n$. The update step for that term is given by

$$\gamma_j(x_i) = \frac{\pi_j N(x_i|\mu_j, \sigma_j^2)}{\sum_{k=1}^{K} \pi_k N(x_i|\mu_k, \sigma_k^2)} \tag{3}$$

(d) Maximize this conditional expectation with respect to μ_k, σ_k^2, and π_k. Your update steps should be the following:

$$\mu_j = \frac{\sum_{i=1}^{n} \gamma_j(x_i) x_i}{\sum_{i=1}^{n} \gamma_j(x_i)} \tag{4}$$

$$\sigma_j^2 = \frac{\sum_{i=1}^{n} \gamma_j(x_i)(x_i - \mu_j)^2}{\sum_{i=1}^{N} \gamma_j(x_i)} \tag{5}$$

$$\pi_j = \frac{\sum_{i=1}^{n} \gamma_j(x_i)}{n} \tag{6}$$

(e) You now have update steps for all parameters in the model. Clearly write each update step. Make sure you include the update step derived in step (c).

(f) Fill in the code provided in **HW_6_Guide.R** for $K = 2$. Note that for $K = 2$, you can keep track of only π and $(1 - \pi)$, and similarly, just one vector of $\gamma(x_n)$ since $\gamma_1(x_i) = 1 - \gamma_1(x_i)$.

(g) Run your code on the data set for $K = 2$. **hw6_prob1_data.txt** found on Canvas. Clearly provide estimates for θ (and only θ, not the γ values).