Problem 5: Justification of the K-means Algorithm (10 pts)

Let $x_1, \ldots, x_n \in \mathbb{R}^p$ denote the expression levels of n genes in p samples, with x_{ij} indicating the expression of gene i in sample j. Let $C_1, \ldots, C_K$ denote the K non-overlapping clusters, each containing a subset of $\{1, \ldots, n\}$, with $\cup_{k=1}^K C_k = \{1, \ldots, n\}$. Let $|C_k|$ denote the size of cluster k and $m_k = (m_{k1}, \ldots, m_{kp})'$ be the center of cluster k. The objective function to minimize is

$$f(C_1, \ldots, C_K, m_1, \ldots, m_K) = \sum_{k=1}^K \sum_{i \in C_k} \sum_{j=1}^p (x_{ij} - m_{kj})^2 = \sum_{k=1}^K \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^p (x_{ij} - x_{i'j})^2.$$

1. In the first step of the $(t+1)$-th iteration of the algorithm ($t = 0, 1, \ldots$), given the clusters from the t-th iteration $C_1^{(t)}, \ldots, C_K^{(t)}$. Show that updating the cluster centers as

$$m_{kj}^{(t+1)} = \frac{1}{|C_k^{(t)}|} \sum_{i \in C_k^{(t)}} x_{ij}, \ j = 1, \ldots, p$$

satisfies that

$$f\left(C_1^{(t)}, \ldots, C_K^{(t)}, m_1^{(t+1)}, \ldots, m_K^{(t+1)}\right) \le f\left(C_1^{(t)}, \ldots, C_K^{(t)}, m_1^{(t)}, \ldots, m_K^{(t)}\right).$$

$$f[C_1^k, m_1^k] = \sum_{k=1}^{k} \sum_{i \in C_k} \|x_i - m_k\|_2^2$$

$$\frac{df}{dm_p} = -\sum_{i \in C_p} 2(x_i - m_p) = 0$$

$$\Rightarrow \widehat{m_p} = \frac{1}{|C_p|} \sum_{i \in C_p} x_i$$

$$\frac{d^2f}{dm_p^2} = 2|C_p| > 0$$

$\Rightarrow \widehat{m_p}$ is minimizer for C_1^k

$$\therefore f[C_1^{k(t)}, m_k^{k(t+1)}] \le f[C_1^{k(t)}, m_1^{k(t)}].$$

2. In the second step of the $(t+1)$-th iteration of the algorithm, given the cluster centers from the first step $m_1^{(t+1)}, \ldots, m_K^{(t+1)}$. Show that if we update the cluster membership of gene i as

$$c(i)^{(t+1)} = \underset{k \in \{1,\ldots,K\}}{\arg\min} \sum_{j=1}^{p} \left(x_{ij} - m_{kj}^{(t+1)} \right)^2 ,$$

the resulting updated clusters

$$C_k^{(t+1)} = \{i : c(i)^{(t+1)} = k\},\ k = 1, \ldots, K$$

satisfy that

$$f\left(C_1^{(t+1)}, \ldots, C_K^{(t+1)}, m_1^{(t+1)}, \ldots, m_K^{(t+1)}\right) \le f\left(C_1^{(t)}, \ldots, C_K^{(t)}, m_1^{(t+1)}, \ldots, m_K^{(t+1)}\right).$$

let z_i = the cluster x_i belongs to

$\Rightarrow f[C_1^K, m_1^K] = \sum_{i=1}^{n} \sum_{k=1}^{K} ||x_i - m_k||_2^2 \, I\{z_{ik} = 1\}$

$\Rightarrow f$ is minimize iif $\sum_{k=1}^{K} ||x_i - m_k||_2^2 \, I\{z_i = k\}$ is minimized

Also, iif $k = \underset{k \in \{1,\ldots,K\}}{\arg\min} ||x_i - m_k||_2^2 \overset{\text{iif}}{\Rightarrow} C(i) = \underset{k \in \{1,\ldots,K\}}{\arg\min} ||x_i - m_k||_2^2$

$\therefore f[C_1^{K(t+1)}, m_1^{K(t+1)}] \le f[C_1^{K(t)}, m_1^{K(t+1)}]$

Problem 7: EM Algorithm for the Gaussian Mixture Model (20 pts)

In the following Gaussian Mixture Model

$$X_i|Z_i = 0 \sim N(\mu_0, \sigma_1^2);$$

$$X_i|Z_i = 1 \sim N(\mu_1, \sigma_2^2);$$

$$Z_i \sim Bernoulli(\gamma),\ i = 1, \dots, n,$$

where X_i's are observable random variables, and Z_i's are hidden random variables.

Given observed data points $x_1, \dots, x_n$, derive the EM algorithm for estimating μ_0, μ_1, σ_1^2, σ_2^2 and γ in the following steps .

1. Write down the complete log-likelihood $\ell(\mu_0, \mu_1, \sigma_1^2, \sigma_2^2, \gamma)$ in terms of $x_1, \dots, x_n$ and $Z_1, \dots, Z_n$.

$$\sum_{i=1}^{n} \log \mathbb{P}(x_i z_i|\theta) = \sum_{i=1}^{n} \log \mathbb{P}(x_i z_i|\theta) + \sum_{i=1}^{n} \log \mathbb{P}(z_i|\theta)$$

$$= \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \mathbb{P}(x_i|z_i = k, \theta) + \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik} \log \lambda_k$$

$\Rightarrow k=2,\ \lambda_1 \neq r,\ \lambda_2 = r,\ z_{i1} = 1 - z_{i2}$,

$$\mathbb{P}(x_i|z_i=k,\theta) = \frac{1}{\sqrt{2\pi}\delta_{k+1}} \exp\left[1 - \frac{1}{2\delta_{k+1}^2}(x_i - \mu_k)^2\right]$$

$\therefore\ \ell(\mu_0, \mu_1, \delta_1^2, \delta_2^2, r)$

$$= \sum_{i=1}^{n}\sum_{k=1}^{2} \left\{ z_{ik}\left[-\frac{1}{2}\log 2\pi - \log \delta_k - \frac{1}{2\delta_k^2}(x_i - \mu_{k-1})^2\right] + z_{ik}\log r_k \right\}$$

$$= n\left[\log(1-r) - \log\sqrt{2\pi}\delta_2\right] + \left(\sum_i z_i\right)\left[\log\frac{\delta_2}{\delta_1} + \log\frac{r}{1-r}\right]$$

$$- \frac{\sum_i (x_i - \mu_i)^2}{2\delta_2^2} + \sum_i \frac{z_i}{2}\left[\frac{(x_i - \mu_1)^2}{\delta_2^2} - \frac{(x_i - \mu_0)^2}{\delta_1^2}\right]$$

2. In the E-step of the $(t+1)$-th iteration ($t = 0, 1, 2, \ldots$), derive the conditional expectation of Z_i given x_i and the current parameter estimates $\left(\widehat{\mu}_0^{(t)}, \widehat{\mu}_1^{(t)}, (\widehat{\sigma}_1^{(t)})^2, (\widehat{\sigma}_2^{(t)})^2, \widehat{\gamma}^{(t)}\right)$:

$$\tau_i^{(t+1)} = E\left[Z_i | x_i, \widehat{\mu}_0^{(t)}, \widehat{\mu}_1^{(t)}, (\widehat{\sigma}_1^{(t)})^2, (\widehat{\sigma}_2^{(t)})^2, \widehat{\gamma}^{(t)}\right] .$$

3. In the M-step of the $(t+1)$-th iteration, derive the updated parameter estimates based on $x_1, \ldots, x_n$ and $\tau_1^{(t+1)}, \ldots, \tau_n^{(t+1)}$.

$$\left(\widehat{\mu}_0^{(t+1)}, \widehat{\mu}_1^{(t+1)}, (\widehat{\sigma}_1^{(t+1)})^2, (\widehat{\sigma}_2^{(t+1)})^2, \widehat{\gamma}^{(t+1)}\right) .$$

2) $\mathbb{E}(z_{ik} | x_i, \theta^t) = \mathbb{P}(z_{ik} = 1 | x_i, \theta^t)$

$= \dfrac{\mathbb{P}(z_{ik}=1, x_i | \theta^t)}{\sum_{k=1}^{k} \mathbb{P}(z_{ik}=1, x_i | \theta^t)}$

$\Rightarrow \mathbb{P}(z_{ik}=1, x_i | \theta^t) = \mathbb{P}(x_i | z_{ik}=1, \theta^t)\,\mathbb{P}(z_{ik}=1 | \theta^t) = \varphi\left(\dfrac{x_i - \mu_{k-1}}{\delta_k}\right) \cdot \gamma_k$

$\Rightarrow \varphi(\cdot)$ is pdf for $N(0,1)$ and $\gamma_1 = \gamma$, $\gamma_2 = 1-\gamma$.

$\Rightarrow \mathbb{E}(z_i | x_i, \theta^t)$ is $\dfrac{\gamma^{(t)} \cdot \varphi\left(\dfrac{x_i - \mu_1^t}{\delta_2^{2(t)}}\right)}{\left[\gamma^{(t)} \varphi\left(\dfrac{x_i - \mu_1}{\delta_2^2}\right) + (1-\gamma^{(t)}) \varphi\left(\dfrac{x_i - \mu_0^{(t)}}{\delta_2^{2(t)}}\right)\right]}$

3) z_i is Bernoulli distribution

$\Rightarrow \hat{\gamma}^{(t+1)} = \dfrac{1}{n} \sum_i \mathbb{E}(z_i / x_i, \theta^{(t)})$, $\theta = \underbrace{(\gamma, \mu_0, \mu_1, \delta_1^2, \delta_2^2)}_{\text{parameters.}}$

estimation:

$\hat{\mu}_1^{(t+1)} = \sum_i [x_i \mathbb{E}(z_i / x_i, \theta^{(t)})] \,/\, [\sum_i \mathbb{E}(z_i / x_i, \theta^{(t)})]$

$\hat{\mu}_0^{(t+1)} = \sum_i [x_i - x_i \mathbb{E}(z_i / x_i, \theta^{(t)})] \,/\, [n - \sum_i \mathbb{E}(z_i / x_i \theta^{(t)})]$

$\hat{\delta}_1^2 = [\sum_i x_i^2 (1 - \mathbb{E}(z_i / x_i, \theta^{(t)})] \,/\, \sum_i (1 - \mathbb{E}(z_i / x_i, \theta^{(t)})) - (\hat{\mu}_0^{(t+1)})^2$

$\hat{\delta}_2^2 = [\sum_i x_i^2 \mathbb{E}(z_i / x_i, \theta^{(t)})] \,/\, \sum_i (1 - \mathbb{E}(z_i / x_i, \theta^{(t)})) - (\hat{\mu}_1^{(t+1)})^2$