CSC4008 hw-EM

Name：费祥 Student ID：120090414

T1.

A. According to the assumption, variances are shared spherical, $\Sigma_k = \sigma I$, and we have $\pi_k = \frac{1}{K}$

$$\underset{z,\mu}{\operatorname{argmax}} \log P(D, z|\pi,\mu,\sigma) = \underset{z,\mu}{\operatorname{argmax}} \sum_i \log P_\pi(z_i) P_{\mu,\sigma}(x_i|z_i)$$

$$= \underset{z,\mu}{\operatorname{argmax}} \sum_i \left(\log\frac{1}{K} + \log N(x_i; \mu_{z_i}, \sigma)\right) = \underset{z,\mu}{\operatorname{argmin}} \sum_i \frac{\|\mu_{z_i} - x_i\|^2}{2\sigma^2}$$

$$= \underset{z,\mu}{\operatorname{argmin}} \sum_i \|\mu_{z_i} - x_i\|^2 = \underset{z,\mu}{\operatorname{argmin}} \sum_i \sum_k r_{ik} \|\mu_k - x_i\|^2$$

∴ the above is the same as k-means

∴ Q.E.D.

B. according to the objective function, we don't need to compute the covariance matrices.

T2.

(i) (11.87) is the term $\sum_{z_i} q_i(z_i) \log \frac{P(x_i, z_i|\theta)}{q_i(z_i)}$ from the left hand of (11.85), and the lower bound in (11.85) is a sum over i of this term.

Then, since joint probability $P(x_i, z_i|\theta) = P(z_i|x_i, \theta)P(x_i|\theta)$, we can reach (11.88)

(ii) $Q(\theta^t, \theta^t) = \sum_i L(\theta^t, q_i^t)$

since $q_i^t(z_i) = P(z_i|x_i, \theta^t)$ ∴ KL divergence is 0

∴ $L(\theta^t, q_i^t) = \log P(x_i|\theta^t) \Rightarrow Q(\theta^t, q_i^t) = \sum_i \log P(x_i|\theta^t)$

and we have $l(\theta^t) = \sum_i \log\left[\sum_{z_i} P(x_i, z_i|\theta^t)\right] = \sum_i \log P(x_i|\theta^t)$

∴ $Q(\theta^t, \theta^t) = \sum_i \log P(x_i|\theta^t) = l(\theta^t)$

(iii) since $\log p(X_i|\theta) \geq L(\theta, q_i)$ always holds

$\therefore L(\theta^{t+1}) \geq Q(\theta^{t+1}, \theta^t)$

and since $\theta^{t+1} = \arg\max_\theta Q(\theta, \theta^t)$

$\therefore Q(\theta^{t+1}, \theta^t) \geq Q(\theta^t, \theta^t)$

and here we have $Q(\theta^t, \theta^t) = l(\theta^t)$

$\therefore$ even though the hidden variable Z do not appear, we still know that $q_i^t(z_i) = P(z_i|x_i, \theta^t)$ since $Q(\theta^t, \theta^t) = L(\theta^t)$

$\therefore$ they still have the right values in iteration t,

T3.

To explain why M-step works, we first need to know what M-step does.

In fact, our final purpose is to increase the log-likelihood to a maximum. but for the original optimized problem, there is no closed form solution since:

$$\log P(D;\theta) = \sum_{n=1}^{N} \log P(X^{(n)};\theta) = \sum_{n=1}^{N} \log\left(\sum_{Z^{(n)}} P(Z^{(n)}, X^{(n)};\theta)\right)$$

so there is a summation inside the log, it is hard to optimize.

therefore, we use the auxiliary distribution of hidden variables $q_n(Z^{(n)})$, and we have:

$$\ln P(D;\theta) = L(q,\theta) + KL(q(Z)\|P(Z|D;\theta)$$

$L(q,\theta)$ is the lower bound, and in fact it is $Q(\theta;\theta_{old})$.

and it is much easier to optimize, so we can get the θ_{new} in M-step.

Then, we just need to illustrate that this optimization can increase the original log-likelihood.

In E-step, we get the q_i^t that make $L(q,\theta)$ and $l(\theta;X,y)$ tightest, and since $q_i^t(z) = P(z_i|x_i,\theta^t)$, here, θ^t is θ_{old}, θ^{t+1} is θ_{new}.

So the KL divergence is $0 \Rightarrow L(\theta^t) = Q(\theta^t, \theta^t)$

and since $\log p(X_i|\theta) \geq L(\theta, q_i)$ always holds

$\therefore L(\theta^{t+1}) \geq Q(\theta^{t+1}, \theta^t)$

and since $\theta^{t+1} = \arg\max_\theta Q(\theta, \theta^t)$

$\therefore Q(\theta^{t+1}, \theta^t) \geq Q(\theta^t, \theta^t) = L(\theta^t)$

$\therefore L(\theta^{t+1}) \geq L(\theta^t) \Rightarrow L(\theta_{new}) \geq L(\theta_{old})$

so the likelihood increase.

Therefore, M-step works.