2 b)

$$P_\theta(x) = \sum_{k=1}^{K} P_\theta(z=k) P_\theta(x|z=k)$$

$$= \phi_1 N(x; \mu_1, \Sigma_1) + \dots + \phi_K N(x; \mu_K, \Sigma_K)$$

c)

$$\mu_K = \frac{\sum_{i=1}^{n} P(z=k|x^i) x^i}{n_K}$$

$$\Sigma_K = \frac{\sum_{i=1}^{n} P(z=k|x^i)(x^i - \mu_K^*)(x^i - \mu_K^*)}{n_K}$$

where $n_K = \sum_{i=1}^{n} P(z=k|x^i)$

$$\phi_K^* = \frac{n_K}{n}$$

The optimal ϕ_K^K is just the proportion of data points with class k.

WRITTEN EXERCISES

1) After updating cluster assignment, for each x^i, there are two possible conditions:

i) x^i stays in its original cluster. Then, obviously, we have:

$$\|x^i - c_{t-1}^{(f_t(x^i))}\|_2 = \|x^i - c_{t-1}^{(f_{t-1}(x^i))}\|_2$$

ii) x^i moves to a new cluster since it is closer to the centroid in the new cluster, i.e.

$$\|x^i - c_{t-1}^{(f_t(x^i))}\|_2 < \|x^i - c_{t-1}^{(f_{t-1}(x^i))}\|_2$$

Therefore, we can get:

$$\|x^i - c_{t-1}^{(f_t(x^i))}\|_2 \leq \|x^i - c_{t-1}^{(f_{t-1}(x^i))}\|_2 .$$

Hence, using the definition of the K-means optimization objective function, we can deduce:

$$\cancel{J(c_{t-1}, f_t) = \sum_{i=1}^{n} \|x^i - c_{t-1}^{(f_t(x^i))}\|_2}$$

$$J(c_{t-1}, f_t) = \sum_{i=1}^{n} \|x^i - c_{t-1}^{(f_t(x^i))}\|_2 \leq \sum_{i=1}^{n} \|x^i - c_{t-1}^{(f_{t-1}(x^i))}\|_2$$

$$= J(c_{t-1}, f_{t-1}).$$

b) Select a random cluster $K=k$ and identify a point where the sum of its distances to all x^i in the same cluster is represented as μ^k, which means we want to find a $\mu^k \in \mathbb{R}^d$ where

$$\underset{}{\text{argmin}} \sum_{i: f_t(x^i)=k} \| x^i - \mu^k \|_2 .$$

Mathematically, the goal requires us to get

$$\frac{\partial \sum_{i: f_t(x^i)=k} \| x^i - \mu^k \|_2}{d \mu^k} = 0$$

By definition of the L2 norm, we know

$$\sum_{i: f_t(x^i)=k} \| x^i - \mu^k \|_2 = \sqrt{(x_1^i - \mu_1^k)^2 + (x_2^i - \mu_{1,2}^k)^2 + \dots + (x_d^i - \mu_d^k)^2}$$

Since for each $j = 1, 2, 3, \dots, d$, we have $(x_j^i - \mu_j^k)^2 \geq 0$,

we can simplify the calculation to try to find the value of μ^k such that

$$\frac{\partial \sum_{i: f_t(x^i)=k} (x_1^i - \mu_1^k)^2 + (x_2^i - \mu_{1,2}^k)^2 + \dots + (x_d^i - \mu_d^k)^2}{d \mu^k}$$

$$= 0$$

Thus, we can get:

$$\frac{d\ \sum_{i:f_t(x^i)=k}[(x_1^i-\mu_1^k)^2+(x_2^i-\mu_2^k)^2+\dots+(x_d^i-\mu_d^k)^2]}{\mu^k}$$

$$= \sum_{i:f_t(x^i)=k}[(-2x_1^i+2\mu_1^k)+(-2x_2^i+2\mu_2^k)+\dots+(-2x_d^i+2\mu_d^k)]$$

$$= -2\sum_{i:f_t(x^i)=k}[(x_1^i-\mu_1^k)+(x_2^i-\mu_2^k)+\dots+(x_d^i-\mu_d^k)]=0$$

Arithmetically, we can have:

$$(\sum_{i:f_t(x^i)=k} x_1^i - \sum_{i:f_t(x^i)=k}\mu_1^k)+\sum_{i:f_t(x^i)=k} x_2^i - \sum \mu_2^k + \dots + \sum_{i:f_t(x^i)=k} x_d^i - \sum_{i:f_t(x^i)=k}\mu_d^k = 0$$

which means

$$\sum_{i:f_t(x^i)=k} x_1^i - S^k\mu_1^k + \sum_{i:f_t(x^i)=k} x_2^i - S^k\mu_2^k + \dots + \sum_{i:f_t(x^i)=k} x_d^i - S^k\mu_d^k = 0$$

Obviously, for each $j=1,2,3,\dots,d$, we can keep

$$\sum_{i:f_t(x^i)=k} x_j^i - S^k\mu_j^k = 0$$

i.e. $\mu_j^k = \frac{\sum_{i:f_t(x^i)=k} x_j^i}{S^k}$, which means

$$\mu^k = \frac{\sum_{i:f_t(x^i)=k} x^i}{S^k} = c_t^{(k)}$$

then the partial derivative will stay 0.

This concludes that given a cluster $K=k$, for any $a \in \mathbb{R}^d$, we have

$$\sum_{i:f_t(x^i)=k} \|x^i - c_t^k\|_2 \leqslant \sum_{i:f_t(x^i)=k} \|x^i - a\|_2 ,$$

including the condition that $a = c_{t-1}^k$.

Thus, expanding to all clusters, it is safe for us to conclude:

$$J(c_t, f_t) = \sum_{i=1}^{n} \|x^i - c_t^{(f_t(x^i))}\|_2$$

$$\leq \sum_{i=1}^{n} \|x^i - c_{t-1}^{(f_t(x^i))}\|_2$$

$$= J(c_{t-1}, f_t).$$

② LOCAL OPTIMA IN K-MEANS

Consider the following dataset with six points {0.1, 0.6, 1.5, 3, 3.7, 8}

FIRST INITIALIZATION:

Let's assume:

- 1st initial centroid: ~~2~~ 0.6
- 2nd initial centroid: 8

When K-means is run with these initial centroids it might converge to Cluster 1: (0.1 0.6 1.5 3 3.7) or Cluster 2: 8

SECOND INITIALIZATION:

Now, let's assume:

- 1st initial centroid: ~~1~~ 1
- 2nd initial centroid: ~~3~~ 4

When K-means is run with these new initial centroids, it might converge to Cluster 1: (0.1 0.6 1.5) or Cluster 2: {3 3.7 8}