Student Name: Pengcheng Xu

Collaboration Statement:

Total hours spent: 8 hrs

I discussed ideas with these individuals:

- I did it on my own
- ...

I consulted the following resources:

- Course website
- Course slides
- ...

By submitting this assignment, I affirm this is my own original work that abides by the course collaboration policy.

Links: [HW4 instructions] [collab. policy]

Contents

1a: Problem Statement

Find the optimal one-hot assignment vectors r^1 for all $N = 7$ examples, given the initial cluster locations μ^0. Report the value of the cost function $J(x, r^1, \mu^0)$.

1a: Solution

TODO FILL IN TABLE

μ^0	r^1	$J(x_{1:N}, r^1, \mu^0)$

`[[-3.   -2. ]`
` [ 1.5   3. ]`
` [ 2.    2. ]]`
	`[[1 0 0]`
	` [1 0 0]`
	` [1 0 0]`
	` [1 0 0]`
	` [0 1 0]`
	` [0 1 0]`
	` [0 0 1]]`
		74.00000

1b: Problem Statement

Find the optimal cluster locations μ^1 for all K=3 clusters, using the optimal assignments r^1 you found in 2a. Report the value of the cost function $J(x, r^1, \mu^1)$.

1b: Solution

TODO FILL IN TABLE

μ^1	r^1	$J(x_{1:N}, r^1, \mu^1)$

`[[ -3.500   1.125]`
` [ -0.750   3.000]`
` [ 2.000   2.000]]`
	`[[1 0 0]`
	` [1 0 0]`
	` [1 0 0]`
	` [1 0 0]`
	` [0 1 0]`
	` [0 1 0]`
	` [0 0 1]]`
		23.81250

1c: Problem Statement

Find the optimal one-hot assignment vectors r^2 for all N=7 examples, using the cluster locations μ^1 from 1b. Report the value of the cost function $J(x, r^2, \mu^1)$.

1c: Solution

TODO FILL IN TABLE

μ^1	r^2	$J(x_{1:N}, r^2, \mu^1)$
[[-3.500 1.125] [-0.750 3.000] [2.000 2.000]]	[[1 0 0] [1 0 0] [1 0 0] [1 0 0] [1 0 0] [0 0 1] [0 0 1]]	18.70312

1d: Problem Statement

Find the optimal cluster locations μ^2 for all K=3 clusters, using the optimal assignments r^2 from above. Report the value of the cost function $J(x, r^2, \mu^2)$.

1d: Solution

TODO FILL IN TABLE

μ^2	r^2	$J(x_{1:N}, r^2, \mu^2)$
[[-4.400 1.500] [0.000 0.000] [1.750 2.500]]	[[1 0 0] [1 0 0] [1 0 0] [1 0 0] [1 0 0] [0 0 1] [0 0 1]]	17.32500

1e: Problem Statement

What interesting phenomenon do you see happening in this example regarding cluster 2? How could you set cluster 2's location in 1d to better fulfill the goals of K-means (find K clusters that reduce cost the most)?

1e: Solution

The interesting phenomenon is that there's no data belongs to cluster 2.

I would set cluster 2's location on the top of the data in the lower left corner (i.e. [-3.0, -2.0]), this would cause that data belongs to the cluster 2 and reduce the cost.

2a: Problem Statement

Show (with math) that using the parameter settings defined above, the general formula for γ_{nk} will simplify to the following (inspired by PRML Eq. 9.42):

$$\gamma_{nk} = \frac{\exp(-\frac{1}{2\epsilon}(x_n - \mu_k)^T(x_n - \mu_k))}{\sum_{j=1}^{K} \exp(-\frac{1}{2\epsilon}(x_n - \mu_j)^T(x_n - \mu_j))} \tag{1}$$

2a: Solution

Next, we'll show the derivation step-by-step and also provide the comments along-side.

$$\gamma_{nk} = \frac{\pi_k \cdot \mathcal{N}(x_n|u_k, \Sigma_k)}{\Sigma_{j=1}^{K} \pi_j \cdot \mathcal{N}(x_n|u_j, \Sigma_j)}$$

/* Plugging Gaussian Pdf and Using the fact $\pi_{1:K} = \pi_{1:3} = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$, and the symbol D denotes dimension*/

$$= \frac{\frac{1}{3} \cdot \frac{1}{(2\pi)^{\frac{D}{2}} \cdot |\Sigma|^{\frac{1}{2}}} \cdot exp(-\frac{1}{2}(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))}{\frac{1}{3} \cdot \frac{1}{(2\pi)^{\frac{D}{2}} \cdot |\Sigma|^{\frac{1}{2}}} \cdot \Sigma_{j=1}^{K} \cdot exp(-\frac{1}{2}(x_n - \mu_j)^T \Sigma^{-1}(x_n - \mu_j))}$$

/* Cancel out the common items on the left part */

$$= \frac{exp(-\frac{1}{2}(x_n - \mu_k)^T \Sigma^{-1}(x_n - \mu_k))}{\Sigma_{j=1}^{K} exp(-\frac{1}{2}(x_n - \mu_j)^T \Sigma^{-1}(x_n - \mu_j))}$$

/* Using the fact that "Covariance $\Sigma_{1:K}$ set to I_D for all clusters, for some $\epsilon > 0$" */

$$= \frac{exp(-\frac{1}{2\epsilon}(x_n - \mu_k)^T (x_n - \mu_k))}{\Sigma_{j=1}^{K} exp(-\frac{1}{2\epsilon}(x_n - \mu_j)^T (x_n - \mu_j))}$$

Thus, we've proved equation (1).

2b: Problem Statement

What will happen to the vector γ_n as $\epsilon \to 0$? How is this related to K-means?

2b: Solution

As $\epsilon \to 0$, the vector γ_n would become a hot-spot vector (i.e. only one element is 1, the rest elements are 0's). Thus, γ_n would reduce to K-means assignment distribution (i.e. r_n) in this case.

The intuition behind this is that, as $\epsilon \to 0$, each data point x_n would become closer and closer to the cluster point μ_n that generate x_n (cuz $x_n \sim \mathcal{N}(\mu_n|\sigma_n)$, as $\epsilon \to 0$, σ_n also $\to 0$, which means x_n becomes closer and closer to μ_n).

As a result, as $\epsilon \to 0$, only one item (i.e. the item where μ_n generate x_n) $exp(-\frac{1}{2\epsilon}(x_n - \mu_k)^T (x_n - \mu_k))$ is 1 (cuz when x_n is close enough to μ_n, $x_n - \mu_k$ would become 0), all others are 0. That's why γ_n would become a hot-spot vector.

3a: Problem Statement

Given: $m = \mathbb{E}_{p^{\text{mix}(x)}}[x]$. Prove that the covariance of vector x is:

$$\text{Cov}_{p^{\text{mix}}(x)}[x] = \sum_{k=1}^{K} \pi_k(\Sigma_k + \mu_k\mu_k^T) - mm^T \tag{2}$$

3a: Solution

Next, we'll derive equation (2) step by step, and add the comments alongside.

/* Based on Hint (3), Covariance corollary */

$$\text{Cov}_{p^{\text{mix}}}[x] = E_{p^{\text{mix}}}[xx^T] - E_{p^{\text{mix}}}[x]E_{p^{\text{mix}}}[x]^T$$

/* Using given $m = \mathbb{E}_{p^{\text{mix}}}[x]$ */

$$= E_{p^{\text{mix}}}[xx^T] - mm^T$$

/* Replacing $p^{min}(x)$'s by $f_k(x)$, and using Expectation's linearity */

$$\begin{aligned} &= E[p^{\text{mix}}(xx^T)] - mm^T \\ &= E[\Sigma_{k=1}^K \pi_k f_k(xx^T)] - mm^T \\ &= \Sigma_{k=1}^K \pi_k E[f_k(xx^T)] - mm^T \end{aligned} \tag{3}$$

Now, Comparing equation (2) and (3), the only thing we need to do is to show that $E[f_k(xx^T)] = \Sigma_k + \mu_k\mu_k^T$. This could be shown by the following:

/* Using Hint(3) */

$$E_{f_k}[xx^T] = \text{Cov}_{f_k}[x] + E_{f_k}[x]E_{f_k}[x]^T$$

/* Using the given info of expectation and covariance about $f_k(x)$ */

$$= \Sigma_k + \mu_k \mu_k^T$$

Thus, we've proved equation (2).

4a (OPTIONAL): Problem Statement

Consider any two Categorical distributions $q(z)$ and $p(z)$ that assign positive probabilities over the same size-K sample space. Show that their KL divergence is non-negative. That is, show that

$$KL\left(\text{CatPMF}(z|\mathbf{r})||\text{CatPMF}(z|\pi)\right) \geq 0 \tag{4}$$

when $\mathbf{r} \in \Delta_+^K$ and $\pi \in \Delta_+^K$.

4a: Solution

TODO