Student Name: TODO

Collaboration Statement:

Total hours spent: TODO

I discussed ideas with these individuals:

- None
- . . .

I consulted the following resources:

- TODO
- TODO
- . . .

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]]`
		74.0

1b: Problem Statement

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

1b: Solution

TODO FILL IN TABLE

μ^1	r^1	$J(x_{1:N}, r^1, \mu^1)$

`[[-3.5   1.125]`
`[-0.75  3.0]`
`[ 2.0   2.0]]`
	`[[1 0 0]`
	` [1 0 0]`
	` [1 0 0]`
	` [1 0 0]`
	` [0 1 0]`
	` [0 1 0]`
	` [0 0 1]]`
		23.8125

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	$J(x_{1:N}, r^2, \mu^1)$

`[[-3.5  1.125]`
`[-0.75  3.0]`
`[ 2.0  2.0]]`
	`[[1 0 0]`
	`[1 0 0]`
	`[1 0 0]`
	`[1 0 0]`
	`[1 0 0]`
	`[0 0 1]`
	`[0 0 1]]`
		18.703

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	$J(x_{1:N}, r^2, \mu^2)$

`[[-3.4  1.5]`
`[0.74  0.199]`
`[1.75  2.5]]`
	`[[1 0 0]`
	`[1 0 0]`
	`[1 0 0]`
	`[1 0 0]`
	`[1 0 0]`
	`[0 0 1]`
	`[0 0 1]]`
		17.703

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

In the example, we see that the no data points are assigned to cluster 2. This is an inefficient use of clusters and might not optimally reduce clustering cost. A possible solution to this problem is randomly re-initializing the location of cluster 2. This could be done by randomly assigning data points from the dataset or using a method like K-means ++ because K-means chooses the initial centers in a way that they are far apart. This increases the chance that all the clusters are used.

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

$$\gamma_{nk} = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$$

Plugging in the given parameters for the GMM, the general formula for γ_{nk} is:

$$= \frac{\pi_k \mathcal{N}(x_n|\mu_k, \epsilon I)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \epsilon I)} \tag{2}$$

Let's investigate the form of the Gaussian distribution in the expression above:

$$\mathcal{N}(x_n|\mu_k, \epsilon I) = \frac{1}{\sqrt{(2\pi)^d |(\epsilon I)|}} \exp\left(-\frac{1}{2}(x - \mu)^T(\epsilon I)^{-1}(x - \mu)\right) \tag{3}$$

Realize that $|(\epsilon I)|$ and $(\epsilon I)^{-1}$ are equivalent to:

- $|(\epsilon I)| = \epsilon^d$ since the determinant of a diagonal matrix is the product of its diagonal entries.
- $(\epsilon I)^{-1} = \frac{1}{\epsilon} I$ because inverse of a diagonal matrix replaces the main diagonal elements of the matrix with their reciprocals.

Inserting these to the Gaussian distribution in expression 3 produces:

$$\mathcal{N}(x_n|\mu_k, \epsilon I) = \frac{1}{(2\pi\epsilon)^{D/2}} \exp\left(-\frac{1}{2}(x_n - \mu_k)^T \frac{1}{\epsilon} I (x_n - \mu_k)\right) \tag{4}$$

An matrix remains the same when multiplied by the identity matrix and therefore:

$$= \frac{1}{(2\pi\epsilon)^{D/2}} \exp\left(-\frac{1}{2\epsilon}(x_n - \mu_k)^T (x_n - \mu_k)\right) \tag{5}$$

Plugging the Gaussian distribution in expression 4 in the expression 2:

$$\gamma_{nk} = \frac{\pi_k \frac{1}{(2\pi\epsilon)^{D/2}} \exp\left(-\frac{1}{2\epsilon}(x_n - \mu_k)^T (x_n - \mu_k)\right)}{\sum_{j=1}^{K} \pi_j \frac{1}{(2\pi\epsilon)^{D/2}} \exp\left(-\frac{1}{2\epsilon}(x_n - \mu_k)^T (x_n - \mu_k)\right))} \tag{6}$$

Simplifying by canceling out $\frac{1}{(2\pi\epsilon)^{D/2}}$ and π_k (since mixture weight are uniform, $\pi_K = \frac{1}{K}$ for all K) in the numerator and denominators, we are left with:

$$\gamma_{nk} = \frac{\exp\left(-\frac{1}{2\epsilon}(x_n - \mu_k)^T (x_n - \mu_k)\right)}{\sum_{j=1}^{K} \exp\left(-\frac{1}{2\epsilon}(x_n - \mu_k)^T (x_n - \mu_k)\right))} \tag{7}$$

2b: Problem Statement

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

2b: Solution

When ϵ approaches 0, the exponential function is much more sensitive to changes, and even small differences in distances lead to large differences in exponential values. In γ_{nk}, each point is assigned to the nearest cluster with a value nearly 1, and γ_{nk} is close to 0 for all other clusters. Therefore, γ_{nk} behaves like a hard assignment in k-means.

3a: Problem Statement

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

$$\text{Cov}_{p^{\text{mix}}(x)}[x] = \sum_{k=1}^{K} \pi_k(\Sigma_k + \mu_k\mu_k^T) - mm^T \tag{8}$$

3a: Solution

Based on hint 3 (ii), we know $Cov(x) = E[xx^T] - E[x]E[x]^T$. Therefore, we have:

$$Cov(pmix(x)) = E_{pmix}[xx^T] - E_{pmix}[x]E_{pmix}[x]^T \tag{9}$$

Next, we derive $E_{pmix}[xx^T]$ and $E_{pmix}[x]E_{pmix}[x]^T$ in expression 9.

For $E_{pmix}[xx^T]$ notice that:

- $E_{pmix}[xx^T] = \sum_{k=1}^{K} \pi_k E_{f_k}[xx^T]$ from hint 3 (i).
- For each k, $E_{f_k}[xx^T] = \sum_{k=1}^{K} +\mu_k\mu_k^T$ based on hint 3 (ii).

Therefore,

$$E_{pmix}[xx^T] = \sum_{k=1}^{K} \pi_k(\sum_{k=1}^{K} +\mu_k\mu_k^T) \tag{10}$$

We know from question description that:

$$m = E_{pmix}[x] = E_{pmix}[x]^T \tag{11}$$

Also, since the transpose of an expectation of a random vector is equivalent to taking the expectation of the transpose of that vector, we have:

$$E[x^T] = (E[x]^T) = m^T \tag{12}$$

Plugging in (10), (11), and (12) in expression (9), we have:

$$Cov(pmix(x)) = \sum_{k=1}^{K} \pi_k(\sum_{k=1}^{K} +\mu_k\mu_k^T) - mm^T \tag{13}$$

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

$$KL\left(\text{CatPMF}(z|\mathbf{r})||\text{CatPMF}(z|\pi)\right) \geq 0 \tag{14}$$

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

4a: Solution

TODO