Student Name: Guangyu Lin

Collaboration Statement:

Total hours spent: 2 hours

I discussed ideas with these individuals:

- TODO
- TODO
- . . .

I consulted the following resources:

- Office Hour
- bishop's textbook
- . . .

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

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

`[[ -3.5 1.125]`
`[ -0.75 3]`
`[ 2 2]]`
	`[[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]`
` [ 2   2]]`
	`[[1 0 0]`
	` [1 0 0]`
	` [1 0 0]`
	` [1 0 0]`
	` [1 0 0]`
	` [0 0 1]`
	` [0 0 1]]`
		18.703125

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   0]`
` [ 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]]`
		17.325

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

When we update the second μ, there are no points that assigned to cluster 2. I will choose a data point that has the greatest cost in this case (-3, -2) and choose that point as the new cluster 2. In this way, the cost is much lower than before since the biggest cost becomes 0 now.

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):

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

2a: Solution

As we set $\pi_{1:K}$ over uniform distribution with $K = 3$ clusters, we can regard π_k and π_j as constant $\frac{1}{3}$. Therefore, based on the γ_{nk} 's definition, we can move the π_j outside from the $\sum$. Then, the denominator of the definition becomes $\pi_j \sum_{j=1}^{K} \mathcal{N}(x_n|\mu_j, \sum_j)$. Now we can cancel out π_k and π_j since they are the same thing. Then we have $\gamma_{nk} = \frac{\mathcal{N}(x_n|\mu_k,\Sigma_k)}{\sum_{j=1}^{K}\mathcal{N}(x_n|\mu_j,\Sigma_j)}$. Then by substibute with MVN-PDF definition we can have $\frac{\frac{1}{(2\pi\epsilon)^{\frac{1}{2}}} exp(-\frac{1}{2\epsilon}(x_n-\mu_k)^T\Sigma^{-1}(x_n-\mu_k))}{\sum_{j=1}^{K}\frac{1}{(2\pi\epsilon)^{\frac{1}{2}}} exp(-\frac{1}{2\epsilon}(x_n-\mu_j)^T\Sigma^{-1}(x_n-\mu_j))}$ Now $\frac{1}{(2\pi\epsilon)^{\frac{1}{2}}}$ and the $\Sigma^{-1} = \epsilon^{-1}I$ based on the assumption ,these two terms can be canceled out from numerator and denominator so that we have $\frac{exp(-\frac{1}{2\epsilon}(x_n-\mu_k)^T(x_n-\mu_k))}{\sum_{j=1}^{K} exp(-\frac{1}{2\epsilon}(x_n-\mu_j)^T(x_n-\mu_j))}$.

2b: Problem Statement

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

2b: Solution

```python
import numpy as np

# Create an array with the float64 data type
gamma = np.array([[0, 0, 0],
                  [0, 0, 0],
                  [0, 0, 0],
                  [0, 0, 0],
                  [0, 0, 0],
                  [0, 0, 0],
                  [0, 0, 0]], dtype=np.float64)  # Specify the data type as np.float64
epsilon = 0.05 # set epsilon as 0.05, when epsilon get closer to 0, the result will
    become nan
def gamma_nk(x, mu, k, epsilon):
    a = np.exp(-(1/(2*epsilon))*(x-mu[k]).T@(x-mu[k]))
    b = 0
    for i in range(3):
        b += np.exp(-(1/(2*epsilon))*(x-mu[i]).T@(x-mu[i]))
    return a/b
for i in range(3):
    for j in range(7):
        result = gamma_nk(x_ND[j],mu_KD,i,epsilon)
        gamma[j][i] = result

print(gamma)
```

```
[[1.00000000e+000 2.16853573e-077 9.19644842e-133]
 [1.00000000e+000 2.71579428e-048 9.43728467e-143]
 [1.00000000e+000 2.86495542e-030 4.83011747e-116]
 [1.00000000e+000 8.76396511e-037 2.19286994e-120]
 [1.00000000e+000 3.02230902e-012 2.47211306e-089]
 [2.04851575e-113 2.34969834e-021 1.00000000e+000]
 [4.27723516e-127 1.48151224e-036 1.00000000e+000]]
```

Based on the output, we can see when ϵ approaching 0, the probability vector will approach to 1 for one certain cluster and the probability of other two cluster will close to 0 which is similar to one hot indicator that used in K-means that hard assigned each data points to one cluster.

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

Based on hint 3(ii) and the definition of m, we can write 3a as $\sum_{k=1}^{K} \pi_k(\Sigma_k + \mu_k\mu_k^T) = Cov_{p^{\text{mix}}(x)}[x] + mm^T$. Then the task becomes derive the $\mathbb{E}_{p^{mix(x)}}[xx^T]$ to $\sum_{k=1}^{K} \pi_k(\Sigma_k + \mu_k\mu_k^T)$
Now based on the definition of expectation and the pdf form of p^{mix} we can write $\mathbb{E}_{p^{mix(x)}}[xx^T]$ as $\int xx^T \sum_{k=1}^{K} \pi_k f_k(x|\mu_k, \Sigma_k)dx$.
Now we can move the summation and $\pi_k$outside of the expectation due to linearity and derived the previous formula as $\sum_{k=1}^{K} \pi_k(\int xx^T f_k(x|\mu_k, \Sigma_k)dx)$
Now we can reuse the hint 3(ii) towards the term $(\int xx^T f_k(x|\mu_k, \Sigma_k)dx)$ is just $\mathbb{E}_{f_k}[xx^T]$ we can write it as $Cov_{fk}(x) + \mathbb{E}_{fk}(x)\mathbb{E}_{fk}(x)^T$
According to the problem statement, we know $Cov_{fk}(x)$ is Σ_k and $\mathbb{E}_{fk}(x)$ is μ_k so we can just substitute into the formula and get $(\Sigma_k + \mu_k\mu_k^T)$
And now we derived that $\mathbb{E}_{p^{mix(x)}}[xx^T] = \sum_{k=1}^{K} \pi_k(\Sigma_k + \mu_k\mu_k^T)$

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

$$KL\left(\text{CatPMF}(z|\mathbf{r})||\text{CatPMF}(z|\pi)\right) \geq 0 \tag{3}$$

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

4a: Solution

Based on the definition of Kl divergence, we know $KL(q(z)||p(z)$ is $\mathbb{E}_{q(z)}[-log\frac{p(z)}{q(z)}]$
Then based on the Jensen inequality, we can regard the f as $-log$ and we only need to prove $f(\mathbb{E}[A]) = 0$
Now we have
$-log\mathbb{E}[\frac{p(z)}{q(z)}]$
$= -log\sum_{k=1}^{K} r_k\frac{\pi_k}{r_k}$ by the definition of $p(z)$ and $q(z)$
$= -log\sum_{k=1}^{K} \pi_k$
Since $r, \pi \in \Delta_+^K$ $\sum_{k=1}^{K} \pi_k$ is 1
now we have $-log1$ which is just 0.
Therefore the $KL(q(z)||p(z) = \mathbb{E}_{q(z)}[-log\frac{p(z)}{q(z)}]$ is ≥ 0