CS-559 Homework Assignment 4

Problem 1

Given: initial centroids are

$A_1 = (2,10)$, $B_1 = (5,8)$, $C_2 = (4,9)$

and the distance function is Euclidean distance

Goal: 1) assign each datapoint to the clusters after the first iteration

2) update centroids after the first iteration

Solution:

given 3 clusters, K_1, K_2 and K_3 and since 3 initial centroids are defined, $A_1 \in K_1$, $B_1 \in K_2$, $C_2 \in K_3$ because the euclidean distance are zero.

consider $A_2 = (2,5)$

$D(A_2, A_1) = (2-2)^2 + (5-10)^2 = 25$

$D(A_2, B_1) = (2-5)^2 + (5-8)^2 = 18$

$D(A_2, C_2) = (2-4)^2 + (5-9)^2 = 20$

$D(A_2, B_1)$ is the minimum distance $\rightarrow J$ is minimized

Therefore, $A_2 \in K_2$

Consider $A_3 = (8,4)$

$D(A_3, A_1) = (8-2)^2 + (4-10)^2 = 72$

$D(A_3, B_1) = (8-5)^2 + (4-8)^2 = 25$

$D(A_3, C_2) = (8-4)^2 + (4-9)^2 = 41$

$D(A_3, B_1)$ is the smallest distance $\rightarrow J$ is minimized

Therefore, $A_3 \in K_2$

Consider $C_1 = (1,2)$

$D(C_1, A_1) = (1-2)^2 + (2-10)^2 = 65$
$D(C_1, B_1) = (1-5)^2 + (2-8)^2 = 52$
$D(C_1, C_2) = (1-4)^2 + (2-9)^2 = 58$

$D(C_1, B_1)$ is the smallest distance $\rightarrow J$ is minimized
Therefore $C_1 \in K_2$

1) The total J will be minimized when $A_2 \in K_2$, $A_3 \in K_2$, $C_1 \in K_2$ Ans

- Consider $J = \sum_N \sum_K r_{nk} \|x_n - \mu_k\|^2$

- The centroids are updated so that J is minimized, keeping r_{nk} fixed

$$\frac{\partial J}{\partial \mu_k} = 0 \Rightarrow \mu_k = \frac{\sum_N r_{nk} x_n}{\sum_N r_{nk}}$$

- r_{nk} is fixed according to $A_2 \in K_2$, $A_3 \in K_2$ and $C_1 \in K_2$
$A_1 \in K_1$
$C_2 \in K_3$

- update μ_k:

$\mu_1 = \frac{(2,10)}{1} = (2,10)$

$\mu_2 = \frac{(5,8) + (2,5) + (8,4) + (1,2)}{4} = \frac{(16,19)}{4} = (4, 4.75)$

$\mu_3 = \frac{(4,9)}{1} = (4,9)$

2) The total J will be minimized when
$\mu_1 = (2,10)$, $\mu_2 = (4, 4.75)$, $\mu_3 = (4,9)$ after 1st iteration.
Ans

Problem 2
Solution:

Define functions.

Gaussian Density Function

```
def jointGaussianPdf(dataset, features_num, k, data_idx, mu, sigma):

    point = dataset[data_idx,:]- mu[k,:]
    a = np.power(2*np.pi,-features_num/2)*np.sqrt(np.linalg.det(sigma[:,:,k]))
    b = np.exp((-1/2)*np.matmul(point,np.matmul(np.linalg.inv(sigma[:,:,k]),np.transpose(point))))

    return a*b
```

Expectation for a kth Gaussian model

```
#expectation
def expectation(dataset,sample_num,k,r,pi,mu,sigma):
    #calculate joint gaussian pdf
    D = dataset.shape[1] #features
    Nk = np.zeros((k,1))
    for s in range(0,sample_num):
        for c in range(0,k):
            r[s,c] = np.multiply(pi[c,:],jointGaussianPdf(dataset,D,c,s,mu,sigma))

    for c in range(0,k):
        r[:,c] = np.divide(r[:,c],np.sum(r,axis=1))
        Nk[c,:] = np.sum(r[:,c],axis=0)
    return r, Nk
```

Maximize mean for a kth Gaussian model

```
def maximizeMean(dataset,sample_num,k,r,Nk,mu):

    weightedSum = np.zeros((sample_num,dataset.shape[1]))
    for c in range(0,k):
        for s in range(0,sample_num):
            weightedSum[s,:] = r[s,c]*dataset[s,:]
        mu[c,:] = np.sum(weightedSum,axis=0)/Nk[c,:]

    return mu
```

Maximize covariance for a kth Gaussian model

```
def maximizeCovariance(dataset,sample_num,k,r,Nk,mu,sigma):

    weightedCov = np.zeros((dataset.shape[1],dataset.shape[1],k))
    for c in range(0,k):
        points = dataset - mu[c,:]
        weightedCov[:,:,c] = np.matmul(np.multiply(r[:,c],np.transpose(points)),points)
        sigma[:,:,c] = weightedCov[:,:,c]/Nk[c,:]
    return sigma
```

Maximize mixture for a kth Gaussian mode

```
def maximizeMixture(k,sample_num,r,Nk,pi):
    for c in range(0,k):
        pi[c,:] = Nk[c,:]/sample_num
    return pi
```

EM includes all functions above to update means, covariances, and model mixtures for a predefined number of iterations.

```
def EM(dataset,sample_num,k,r,pi,mu,sigma,iter_num):
    D = dataset.shape[1] #features
    for i in range(0,iter_num):

        #1 evaluate the log-likelihood for sigma pi mu
        log_likelihood = np.zeros((sample_num,1))
        for d in range(0,sample_num):
            likelihood = 0
            for c in range(0,k):
                #evaluate conditional prob.(gaussian)
                likelihood += np.multiply(pi[c,:],jointGaussianPdf(dataset,D,c,d,mu,sigma))
            log_likelihood[d,:] = np.log(likelihood)
        #print log-likelihood for iteration i
        print('log-likelihood for iteration ',i,' is: ',np.sum(log_likelihood,axis=0))

        #2 E step: evaluate expectation given current pi, mu, sigma
        r, Nk = expectation(dataset,sample_num,k,r,pi,mu,sigma)
        #3 M step: evaluate new pi, mu, sigma
        mu = maximizeMean(dataset,sample_num,k,r,Nk,mu)
        sigma = maximizeCovariance(dataset,sample_num,k,r,Nk,mu,sigma)
        pi = maximizeMixture(k,sample_num,r,Nk,pi)

    return pi, mu, sigma
```

Define parameters and initialize means, covariances, and mixtures.

```
    #define variables
    k = 3 #clusters
    f = train_data.iloc[0].count() #features
    n = train_data[0].count() #samples
    X = train_data.to_numpy() #training data: dim = n x f
    r = np.zeros((n,k)) #membership weights: dim = n x k
    pi = np.zeros((k,1)) #mixture weights: dim = k x 1
    mu = np.zeros((k,f)) #means: dim = k x f
    sigma = np.zeros((f,f,k)) #covars: dim = f x f x k
    nIter = 10 #run EM for 10 iterations
    #initialize sigma, pi, mu
    for c in range(0,k):
        pi[c,:] = 1/k
        mu[c,:] = np.mean(X,axis=0)
        sigma[:,:,c] = np.cov(X,rowvar=False)
```

Values for initial means, covariances, and mixtures are shown below.

```
print('initial pi: \n', pi,'\n')
print('initial mu: \n', mu,'\n')
print('initial sigma: \n')
for c in range(0,k):
    print(sigma[:,:,c])
```

```
initial pi:
 [[0.33333333]
 [0.33333333]
 [0.33333333]]

initial mu:
 [[-0.82042865 -0.58219397]
 [-0.82042865 -0.58219397]
 [-0.82042865 -0.58219397]]

initial sigma:

[[ 7.9328436  -0.10559592]
 [-0.10559592  6.2470177 ]]
[[ 7.9328436  -0.10559592]
 [-0.10559592  6.2470177 ]]
[[ 7.9328436  -0.10559592]
 [-0.10559592  6.2470177 ]]
```

The values of log-likelihood are shown for 10 iterations. Note that the updated values converges after 4th iteration.

```
pi,mu,sigma = EM(X,n,k,r,pi,mu,sigma,nIter)
```

```
log-likelihood for iteration  0  is:  [-885.43171119]
log-likelihood for iteration  1  is:  [-1521.53552285]
log-likelihood for iteration  2  is:  [-1061.65237995]
log-likelihood for iteration  3  is:  [-890.3059187]
log-likelihood for iteration  4  is:  [-887.41565901]
log-likelihood for iteration  5  is:  [-887.43168481]
log-likelihood for iteration  6  is:  [-887.43218431]
log-likelihood for iteration  7  is:  [-887.43221109]
log-likelihood for iteration  8  is:  [-887.43221153]
log-likelihood for iteration  9  is:  [-887.43221153]
```

The final parameters are shown below.

```
print('final pi is as follows: \n', pi,'\n')
print('final mu is as follows: \n', mu,'\n')
print('final sigma is as follows: \n')
for c in range(0,k):
    print(sigma[:,:,c])
```

```
final pi is as follows:
 [[1.00000000e+00]
 [7.95890073e-17]
 [6.10258298e-17]]

final mu is as follows:
 [[-0.82042865 -0.58219397]
 [-0.24552594  1.64909829]
 [-0.24558495  1.64925094]]

final sigma is as follows:

[[ 7.92491076 -0.10549033]
 [-0.10549033  6.24077068]]
[[0.19124412 0.09158298]
 [0.09158298 0.12875877]]
[[0.19064662 0.09136423]
 [0.09136423 0.1283201 ]]
```

Problem 3 : Bayesian Classification

Given: 20 samples with 2 classes (C_0, C_1)

Goal: classify a new data $X = \{Gender = M, CarType = Family, Shirt\ Size = Large\}$

Solution

1) Using Naive Bayes → assume that attributes are independent

- Class prior probability:

$$P(C_0) = \frac{N_0}{N} = \frac{10}{20} = \frac{1}{2}, \quad P(C_1) = \frac{N_1}{N} = \frac{10}{20} = \frac{1}{2}$$

- conditional probability for C_0

$$P(X \mid C_0) = P(Gender=M \mid C_0) \times P(CarType=Family \mid C_0) \times P(Shirt\ Size=Large \mid C_0)$$

based on the dataset

$$P(Gender=M \mid C_0) = \frac{6}{10}$$

$$P(CarType=Family \mid C_0) = \frac{1}{10}$$

$$P(Shirt\ Size=Large \mid C_0) = \frac{2}{10}$$

$$P(X \mid C_0) = \frac{6}{10} \times \frac{1}{10} \times \frac{2}{10}, \quad P(C_0) = \frac{1}{2}$$

- based on these factorizations

$$P(C_0 \mid X) = P(X \mid C_0) P(C_0) = \frac{12}{1000} \times \frac{1}{2} = \frac{6}{1000} = 0.006$$

- conditional probability for C_1

$$P(X \mid C_1) = P(Gender=M \mid C_1) \times P(CarType=Family \mid C_1) \times P(Shirt\ Size=Large \mid C_1)$$

$$P(Gender=M \mid C_1) = \frac{4}{10}$$

$$P(Car\ Type=Family \mid C_1) = \frac{3}{10}$$

$$P(Shirt\ Size=Large \mid C_1) = \frac{2}{10}$$

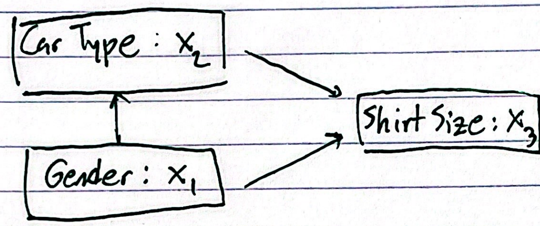
$$P(X \mid C_1) = \frac{4}{10} \times \frac{3}{10} \times \frac{2}{10}, \quad P(C_1) = \frac{1}{2}$$

- based on these factorizations: $P(C_1 \mid X) = P(X \mid C_1) \cdot P(C_1) = \frac{24}{1000} \times \frac{1}{2} = 0.012$

- Compare conditional probability for C_0 & C_1

$$P(C_1|x) > P(C_0|x) = 0.012 > 0.006 \rightarrow \text{Class} = C_1 \quad \underline{\text{Ans}}$$

2) using Bayesian Network



- Class prior probability:

$$P(C_0) = \frac{1}{2}$$
$$P(C_1) = \frac{1}{2}$$

- Conditional Probability given class C_0

$$P(x_1, x_2, x_3 | C_0) \propto P(x_1|C_0)P(x_2|C_0, x_1)P(x_3|C, x_1, x_2)\cancel{P(C_0)}$$
$$P(C_0|x_1, x_2, x_3) = P(x_1|C_0)P(x_2|C_0, x_1)P(x_3|C, x_1, x_2)P(C_0)$$

- given that $x_1 = M$, $x_2 = \text{Family}$, $x_3 = \text{Large}$ for a test data, the relationship is found

$$P(x_1 = M, C_0) = \frac{6}{10}$$

$$P(x_2 = \text{Family} | C_0, x_1 = M) = \frac{1}{6}$$

$$P(x_3 = \text{Large} | C_0, x_1 = M, x_2 = \text{Family}) = \frac{0}{1} = 0$$

$$P(C_0) = \frac{1}{2}$$

$$\therefore P(C_0 | x_1, x_2, x_3) = \left(\frac{6}{10} \times \frac{1}{6} \times \frac{0}{1}\right) \times \frac{1}{2} = 0$$

- Conditional Probability given class C_1

$$P(x_1, x_2, x_3 | C_1) = P(x_1|C_1)P(x_2|C_1, x_1)P(x_3|C_1, x_1, x_2)$$
$$P(C_1|x_1, x_2, x_3) = P(x_1|C_1)P(x_2|C_1, x_1)P(x_3|C_1, x_1, x_2)P(C_1)$$

- given that $x_1 = M, x_2 = \text{Family}, x_3 = \text{Large}$, the relationship is found

$$P(x_1 = M | C_1) = \frac{4}{10}$$

$$P(x_2 = \text{Family} | C_1, x_1 = M) = \frac{3}{4}$$

$$P(x_3 = \text{Large} | C_1, x_1 = M, x_2 = \text{Family}) = \frac{1}{3}$$

$$P(C_1) = \frac{1}{2}$$

$$\therefore P(C_1|x_1, x_2, x_3) = \left(\frac{4}{10} \times \frac{3}{4} \times \frac{1}{3}\right) \times \frac{1}{2} = \frac{1}{20}$$

- Compare conditional probability for C_0 & C_1

$$P(C_1|x_1, x_2, x_3) > P(C_0|x_1, x_2, x_3) = \frac{1}{20} > 0 \rightarrow \text{class} = C_1 \quad \underline{\text{Ans}}$$

Problem 4

Given : Zero biases in each layer

Sigmoid function for 2-hidden layers : $a^{[i]} = \sigma(z^{[i]})$

Softmax function for output layer : $y_i = s(z^{[3]})$

Goal: calculate $\frac{\partial L}{\partial w^1}$, $\frac{\partial L}{\partial w^2}$, $\frac{\partial L}{\partial w^3}$ using backpropagation

Solution:

$$X \longrightarrow \underbrace{w^1 x}_{z^1} \; ; \; \underbrace{\sigma(z^1)}_{a^1} \longrightarrow \underbrace{w^2 a^1}_{z^2} \; ; \; \underbrace{\sigma(z^2)}_{a^2} \longrightarrow \underbrace{w^3 a^2}_{z^3} \; ; \; \underbrace{s(z^3)}_{\hat{y}} \longrightarrow L(\hat{y}, y)$$

- Calculate $\frac{\partial L}{\partial w^3}$

$$\frac{\partial L}{\partial w^3} = \frac{\partial L}{\partial z^3} \cdot \frac{\partial z^3}{\partial w^3}$$

$$= \left(\frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^3}\right)\frac{\partial z^3}{\partial w^3} = \frac{\partial L}{\partial \hat{y}}\left(\frac{\partial \hat{y}_1}{\partial z^3} + \frac{\partial \hat{y}_2}{\partial z^3}\right)\frac{\partial z^3}{\partial w^3}$$

$$= \frac{\partial L}{\partial \hat{y}}\left(\frac{\partial \hat{y}_1}{\partial z_1^3} + \frac{\partial \hat{y}_1}{\partial z_2^3} + \frac{\partial \hat{y}_2}{\partial z_2^3} + \frac{\partial \hat{y}_2}{\partial z_1^3}\right)\frac{\partial z^3}{\partial w^3}$$

$$= \left[\frac{\partial L}{\partial \hat{y}_1}\left(\frac{\partial \hat{y}_1}{\partial z_1^3} + \frac{\partial \hat{y}_1}{\partial z_2^3}\right) + \frac{\partial L}{\partial \hat{y}_2}\left(\frac{\partial \hat{y}_2}{\partial z_2^3} + \frac{\partial \hat{y}_2}{\partial z_1^3}\right)\right]\frac{\partial z^3}{\partial w^3}$$

- The loss function is given by $L = (y_1 - \hat{y}_1)^2 + (y_2 - \hat{y}_2)^2$

$$\frac{\partial L}{\partial \hat{y}_1} = -2(y_1 - \hat{y}_1) \quad , \quad \frac{\partial L}{\partial \hat{y}_2} = -2(y_2 - \hat{y}_2)$$

- The output layer is $y_i = \frac{e^{z_i^3}}{\sum_j e^{z_j^3}}$

$$\hat{y}_1 = \frac{e^{z_1^3}}{\sum e^{z_i^3}} \quad , \quad \hat{y}_2 = \frac{e^{z_2^3}}{\sum e^{z_i^3}}$$

$$\frac{\partial \hat{y}_1}{\partial z_1^3} = \hat{y}_1(1 - \hat{y}_1) \quad , \quad \frac{\partial \hat{y}_2}{\partial z_2^3} = \hat{y}_2(1 - \hat{y}_2) \quad , \quad \frac{\partial \hat{y}_1}{\partial z_2^3} = -\hat{y}_2\hat{y}_1 \quad , \quad \frac{\partial \hat{y}_2}{\partial z_1^3} = -\hat{y}_2\hat{y}_1$$

$$\frac{\partial z^3}{\partial w^3} = a^2$$

$$\therefore \frac{\partial L}{\partial w^3} = -2(y_1-\hat{y}_1)\left[\hat{y}_1(1-\hat{y}_1)-\hat{y}_1\hat{y}_2\right]-2(y_2-\hat{y}_2)\left[\hat{y}_2(1-\hat{y}_2)-\hat{y}_1\hat{y}_2\right] \quad \underline{\underline{\text{Ans}}}$$

$$\frac{\partial L}{\partial w^2} = \frac{\partial L}{\partial z^2} \cdot \frac{\partial z^2}{\partial w^2}$$

$$= \left(\frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^2}\right)\frac{\partial z^2}{\partial w^2} = \left(\frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial z^2} + \frac{\partial L}{\partial \hat{y}_2} \cdot \frac{\partial \hat{y}_2}{\partial z^2}\right)\frac{\partial z^2}{\partial w^2}$$

$$= \left[\frac{\partial L}{\partial \hat{y}_1}\left(\frac{\partial \hat{y}_1}{\partial z_1^2} + \frac{\partial \hat{y}_1}{\partial z_2^2}\right) + \frac{\partial L}{\partial \hat{y}_2}\left(\frac{\partial \hat{y}_2}{\partial z_1^2} + \frac{\partial \hat{y}_2}{\partial z_2^2}\right)\right]\frac{\partial z^2}{\partial w^2}$$

$$\frac{\partial L}{\partial \hat{y}_1} = -2(y_1-\hat{y}_1), \quad \frac{\partial L}{\partial \hat{y}_2} = -2(y_2-\hat{y}_2)$$

$$\frac{\partial \hat{y}_1}{\partial z_1^2} = \frac{\partial \hat{y}_1}{\partial z_1^3} \times \left(\frac{\partial z_1^3}{\partial z_1^2}\right) = \hat{y}_1(1-\hat{y}_1)a^2(1-a^2)w^3$$

$$\frac{\partial \hat{y}_1}{\partial z_2^2} = \frac{\partial \hat{y}_1}{\partial z_2^3} \times \left(\frac{\partial z_2^3}{\partial z_2^2}\right) = -\hat{y}_1\hat{y}_2\, a^2(1-a^2)w^3$$

$$\frac{\partial \hat{y}_2}{\partial z_1^2} = \frac{\partial \hat{y}_2}{\partial z_1^3} \times \left(\frac{\partial z_1^3}{\partial z_1^2}\right) = -\hat{y}_1\hat{y}_2 a^2(1-a^2)w^3$$

$$\frac{\partial \hat{y}_2}{\partial z_2^2} = \frac{\partial \hat{y}_2}{\partial z_2^3} \times \left(\frac{\partial z_2^3}{\partial z_2^2}\right) = \hat{y}_2(1-\hat{y}_2)a^2(1-a^2)w^3$$

$$\frac{\partial z^3}{\partial w^2} = a^1$$

$$\therefore \frac{\partial L}{\partial w^2} = a^1\Big[-2(y_1-\hat{y}_1)\left(\hat{y}_1(1-\hat{y}_1)a^2(1-a^2)w^3 + a^2(1-a^2)w^3(-\hat{y}_1\hat{y}_2)\right) - 2(y_2-\hat{y}_2)\left(-\hat{y}_1\hat{y}_2 a^2(1-a^2)w^3 + \hat{y}_2(1-\hat{y}_2)a^2(1-\right.$$

$$= a^1\left[-2(y_1-\hat{y}_1)\left(\hat{y}_1(1-\hat{y}_1)-\hat{y}_1\hat{y}_2\right)a^2(1-a^2)w^3 - 2(y_2-\hat{y}_2)\left(\hat{y}_2(1-\hat{y}_2)-\hat{y}_1\hat{y}_2\right)a^2(1-a^2)w^3\right]$$

$$= a^1 a^2(1-a^2)w^3\left[\frac{\partial L}{\partial w^3}\right] \quad \underline{\underline{\text{Ans}}}$$

$$\frac{\partial L}{\partial w^1} = \frac{\partial L}{\partial z^1} \times \frac{\partial z^1}{\partial w^1}$$

$$= \left(\frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial z^1}\right)\frac{\partial z^1}{\partial w^1} = \left(\frac{\partial L}{\partial \hat{y}_1} \cdot \frac{\partial \hat{y}_1}{\partial z^1} + \frac{\partial L}{\partial \hat{y}_2} \times \frac{\partial \hat{y}_2}{\partial z^1}\right)\frac{\partial z^1}{\partial w^1}$$

$$= \left[\frac{\partial L}{\partial \hat{y}_1}\left(\frac{\partial \hat{y}_1}{\partial z_1^1} + \frac{\partial \hat{y}_1}{\partial z_2^1}\right) + \frac{\partial L}{\partial \hat{y}_2}\left(\frac{\partial \hat{y}_2}{\partial z_1^1} + \frac{\partial \hat{y}_2}{\partial z_2^1}\right)\right]\frac{\partial z^1}{\partial w^1}$$

$$\frac{\partial L}{\partial \hat{y}_1} = -2(y_1 - \hat{y}_1) \quad , \quad \frac{\partial L}{\partial \hat{y}_2} = -2(y_2 - \hat{y}_2)$$

$$\frac{\partial \hat{y}_1}{\partial z_1^1} = \left(\frac{\partial \hat{y}_1}{\partial z_1^3} \times \frac{\partial z_1^3}{\partial z_1^2}\right)\frac{\partial z_1^2}{\partial z_1^1} = \hat{y}_1(1-\hat{y}_1)a^2(1-a^2)w^3a^1(1-a^1)w^2$$

$$\frac{\partial \hat{y}_1}{\partial z_2^1} = \left(\frac{\partial \hat{y}_1}{\partial z_2^3} \times \frac{\partial z_2^3}{\partial z_2^2}\right)\frac{\partial z_2^2}{\partial z_2^1} = -\hat{y}_1\hat{y}_2 a^2(1-a^2)w^3a^1(1-a^1)w^2$$

$$\frac{\partial \hat{y}_2}{\partial z_1^1} = \left(\frac{\partial \hat{y}_2}{\partial z_1^3} \times \frac{\partial z_1^3}{\partial z_1^2}\right)\frac{\partial z_1^2}{\partial z_1^1} = -\hat{y}_1\hat{y}_2 a^2(1-a^2)w^3a^1(1-a^1)w^2$$

$$\frac{\partial \hat{y}_2}{\partial z_2^1} = \left(\frac{\partial \hat{y}_2}{\partial z_2^3} \times \frac{\partial z_2^3}{\partial z_2^2}\right)\frac{\partial z_2^2}{\partial z_2^1} = \hat{y}_2(1-\hat{y}_2)a^2(1-a^2)w^3a^1(1-a^1)w^2$$

$$\frac{\partial z^1}{\partial w^1} = x$$

$$\therefore \frac{\partial L}{\partial w^1} = a^1a^2(1-a^1)(1-a^2)w^2w^3x\left[-2(y_1-\hat{y}_1)(\hat{y}_1(1-\hat{y}_1) - \hat{y}_1\hat{y}_2) - 2(y_2-\hat{y}_2)(\hat{y}_2(1-\hat{y}_2) - \hat{y}_1\hat{y}_2)\right]$$

$$= (1-a^1)w^2x\left(\frac{\partial L}{\partial w^2}\right)\left[\frac{\partial L}{\partial w^3}\right] \quad \underline{\underline{\text{Ans}}}$$