PRML

Handwritten Assignment -2

Name - Mayank Dahiya

Roll no. - B23CS1035

Honesty Pledge :

" I affirm that this assignment is solely my work. I have not used unauthorised assistance, engaged in plagiarism, or voilated ethical standards. Further, all references used and any discussion with anyone have been appropriately cited. Any breach may lead to disciplinary actions as per the course academic honesty policy discussed in Lecture -1".

Mayank Dahiya

Q-1 Given points = (1,2), (3,4), (4,5), (5,5), (5,6), (10,10)

Mean:- $\bar{x} = \frac{1}{n}\sum x_i = \frac{1}{6}(1+3+4+5+5+10) = 4.67$

$\bar{y} = \frac{1}{n}\sum y_i = \frac{1}{6}(2+4+5+5+6+10) = 5.33$

⇒ Mean Vector = (4.67, 5.33)

Covarience Matrix :

$$\begin{bmatrix} Var(X) & COV(X,Y) \\ COV(X,Y) & Var(Y) \end{bmatrix}$$

Varience (X) = $\frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})^2 = \frac{1}{6}\left[(-3.67)^2 + (1.67)^2 + (-0.67)^2 + (0.33)^2 + (0.33)^2 + (5.33)^2\right]$

$= 7.55$

$$\text{Var}(y) = \frac{1}{N}\sum_{i=1}^{N}(y_i-\bar{y})^2 = \frac{1}{6}\left[(-3.33)^2 + (-1.33)^2 + (-0.33)^2 + (-0.33)^2 + (0.67)^2 + (4.67)^2\right]$$

$$= 5.88$$

$$\text{COV}(x,y) = \frac{1}{N}\sum_{i=1}^{N}(x_i-\bar{x})(y_i-\bar{y})$$

$$= \frac{1}{6}\left[(-3.67)(-3.33) + (-1.67)(-1.33) + (-0.67)(-0.33) + (0.33)(-0.33) + (0.33)(0.67) + (5.33)(4.67)\right]$$

$$= 6.61$$

$\Rightarrow$ Covariance Matrix $= \begin{bmatrix} 7.55 & 6.61 \\ 6.61 & 5.88 \end{bmatrix}$ ans

Q-2 Given Points: $(-1,1)$, $(-2,2)$, $(-3,3)$

(a) Zero Mean Data points are calculated as follows:

$$(x', y') = (x_i, y_i) - (\bar{x}, \bar{y})$$

$$\bar{x} = \frac{1}{3}(-1-2-3) = -2$$

$$\bar{y} = \frac{1}{3}(1+2+3) = +2$$

$$(\bar{x}, \bar{y}) = (-2, 2)$$

$(x', y') = (1,-1), (0,0), (-1,1)$ → after performing above subtraction

$$\cancel{x - y = 2x}$$
$$\cancel{-x + y = 2y}$$
$$\cancel{x + y = 0} \Rightarrow \cancel{x = -y}$$

Eigen Vector $= \begin{bmatrix} -y \\ y \end{bmatrix} = -y \begin{bmatrix} +1 \\ -1 \end{bmatrix}$

Eigenvector $= \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

Normalised EigenVector $= \dfrac{\begin{bmatrix} 1 \\ -1 \end{bmatrix}}{\left\| \begin{bmatrix} +1 \\ -1 \end{bmatrix} \right\|} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$

Taking $\lambda = 0$,

$$\frac{2}{3} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow \begin{bmatrix} x - y \\ -x + y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow x = y$$

Normalized EigenVector $= \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

EigenVectors $= \left\{ \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\}$

(b) CoVariance Matrix $= \frac{1}{N} X^T X$

Let $X = \begin{bmatrix} 1 & -1 \\ 0 & 0 \\ -1 & 1 \end{bmatrix}$ $\quad X^T = \begin{bmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \end{bmatrix}$

$$= \frac{1}{3} \begin{bmatrix} 1 & 0 & -1 \\ -1 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 0 & 0 \\ -1 & 1 \end{bmatrix}$$

$$= \frac{1}{3} \begin{bmatrix} 2 & -2 \\ -2 & 2 \end{bmatrix}$$

$$= \frac{2}{3} \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$$

(c) $|A - \lambda I| = 0$

$$\begin{vmatrix} \frac{2}{3} - \lambda & -\frac{2}{3} \\ -\frac{2}{3} & \frac{2}{3} - \lambda \end{vmatrix} = 0$$

$$\Rightarrow \left(\frac{2}{3} - \lambda\right)^2 - \left(\frac{2}{3}\right)^2 = 0$$

$$\Rightarrow \lambda = 0 \text{ or } \frac{4}{3}$$

Take $\lambda = \frac{4}{3}$, $\frac{\cancel{2}}{\cancel{3}}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \frac{\cancel{4}^{2}}{\cancel{3}}\begin{bmatrix} x \\ y \end{bmatrix}$

$$\Rightarrow \begin{bmatrix} x - y \\ -x + y \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix}$$

$$\Rightarrow x - y = 2x, \quad -x + y = 2y$$

$$\Rightarrow x + y = 0 \longrightarrow \boxed{x = -y}$$

Eigen Vector $= \begin{bmatrix} -y \\ y \end{bmatrix} = -y\begin{bmatrix} 1 \\ -1 \end{bmatrix}$

Eigen Vector $= \begin{bmatrix} 1 \\ -1 \end{bmatrix}$

Normalized Eigen Vector $= \dfrac{\begin{bmatrix} 1 \\ -1 \end{bmatrix}}{\left\| \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\|} = \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{bmatrix}$

Take $\lambda = 0$, $\frac{2}{3}\begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$\begin{bmatrix} x - y \\ -x + y \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\Rightarrow x = y$$

Normalized Eigen Vector $= \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix}$

Eigen Vectors = $\left\{ \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix}, \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \right\}$

(d) Largest eigen Value = $\frac{4}{3}$, hence the principal component will correspond to this eigenValue as it will capture 'maximum variance of data.

Principal Component = $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix}$

$$P(1,-1) = [1 \ -1]\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix} = \frac{2}{\sqrt{2}} = \sqrt{2}$$

$$P(0,0) = [0,0]\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix} = 0$$

$$P(-1,1) = [-1,1]\begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{-1}{\sqrt{2}} \end{pmatrix} = -\sqrt{2}$$

(e) <u>Reconstruction of Points</u> :

$$P_{recons}(+1,-1) = \sqrt{2}\left[\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right] = [1,-1]$$

$$P_{recons}(0,0) = 0\left[\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right] = [0,0]$$

$$P_{recons}(-1,1) = -\sqrt{2}\left[\frac{1}{\sqrt{2}}, \frac{-1}{\sqrt{2}}\right] = [-1,1]$$

Reconstruction error = || Original P − Reconstructed P ||

= ||0|| = 0 (ans)

Reconstruction error = 0 %

Q-3 Given:

$X \rightarrow$ zero mean data

$V \rightarrow$ eigen vectors of the covariance matrix of X.

So, $X \rightarrow n \times p$ dimension

$V \rightarrow p \times p$ dimension

$\{p \leq n\}$

Projected data $(x') = X.V = [x_1, \dots x_n]^T [V_1 V_2 \dots V_p]$

$$= \begin{array}{c} p \text{ variables} \\ \begin{bmatrix} \leftarrow x_1 \rightarrow \\ \leftarrow x_2 \rightarrow \\ \vdots \\ \leftarrow x_n \rightarrow \end{bmatrix}_{n \times p} \end{array} \begin{bmatrix} \uparrow & \uparrow & & \uparrow \\ V_1 & V_2 & \cdots & V_p \\ \downarrow & \downarrow & & \downarrow \end{bmatrix}_{p \times p}$$

$$= \begin{bmatrix} x_1 V_1 & x_1 V_2 & \cdots \\ \vdots & & \\ \vdots & & \end{bmatrix}_{n \times p}$$

So, mean of any column $= S = x_1 V_j + x_2 j + \dots x_n V_j$

$$= \left(\sum_{i=1}^{n} x_i\right) V_j$$

Mean of the given projected data $= \frac{1}{n} \cdot S$

$$= \frac{\left(\sum_{i=1}^{n} x_i\right) V_j}{n} = 0 \quad \text{as } \sum_{i=1}^{n} x_i = 0$$

Hence Proved, the projected data is also zero mean data.

Q-4.

$$J_1 = \frac{1}{n_1 n_2} \sum_{y_i \in Y_1} \sum_{y_j \in Y_2} (y_i - y_j)^2$$

$m_1, m_2 \rightarrow$ means of Y_1, Y_2

$s_1^2, s_2^2 \rightarrow$ sample variance of Y_1, Y_2

(a) Prove : $J_1 = (m_1 - m_2)^2 + \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$

$$\because \quad J_1 = \frac{1}{n_1 n_2} \sum_{y_i \in Y_1} \sum_{y_j \in X_2} (y_i - y_j)^2 = \frac{1}{n_1 n_2} \sum_{y_i \in Y_1} \sum_{y_j \in Y_2} (y_i^2 - 2y_i y_j + y_j^2)$$

$$= \frac{1}{n_1 n_2} \left(\sum_{y_i \in Y_1} y_i^2 \sum_{y_j \in X_2} 1 - 2 \sum_{y_i \in Y_1} y_i \sum_{y_j \in X_2} y_j + \sum_{y_j \in X_2} y_j^2 \sum_{y_i \in Y_1} 1 \right)$$

According to our data initiation step :

We can see that

$$\sum_{y_i \in Y_1} y_i = n_1 m_1 \qquad \sum_{y_j \in Y_2} y_j = n_2 m_2$$

$$\sum_{y_i \in Y_1} y_i^2 = S_1^2 + n_1 m_1^2 \qquad \sum_{y_j \in X_2} y_j^2 = S_2^2 + n_2 m_2^2$$

$$\sum_{y_i \in Y_1} 1 = n_1 \qquad \sum_{y_j \in X_2} 1 = n_2$$

Using above :

$$J_1 = \frac{1}{n_1 n_2} \left((S_1^2 + n_1 m_1^2) n_2 - 2(n_1 m_1)(n_2 m_2) + (S_2^2 + n_2 m_2^2) n_1 \right)$$

$$= \left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} + m_1^2 - 2m_1 m_2 + m_2^2 \right)$$

$$\Rightarrow \quad J_1 = \left[(m_1 - m_2)^2 + \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right]$$

(b) For total scatter matrix

$$J_2 = \frac{1}{n_1} \sum_{y_i \in Y_1} (y_i - m_1)^2 + \frac{1}{n_2} \sum_{y_j \in X_2} (y_j - m_2)^2$$

$$S_1^2 = \sum_{y_i \in Y_1} (y_i - m_1)^2 , \quad S_2^2 = \sum_{y_j \in X_2} (y_j - m_2)^2$$

$\Rightarrow \quad J_2 = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}$

Using $y = w^T x$ and $J_2 = 1$

Lagrangian $\rightarrow L(w, x) = J_1 + \lambda J_2$

Computing gradient of x wrt w.

$$\nabla_w L = \nabla_w J_1 + \lambda \nabla_w J_2$$

$$\nabla_w L = -(y - w^T x)x + \lambda(w - w_0)$$

Solving for w when $\nabla_w L = 0$

$$\Rightarrow -(y - w^T x)x + \lambda(w - w_0) = 0$$

$$\Rightarrow (X^T X - \lambda I)\, w = yX - \lambda w_0$$

$$\Rightarrow w = (X^T X - \lambda I)^{-1}(yX - \lambda w_0)$$

Expressing λ in terms of m_1, m_2, n_1, n_2 and S_1, S_2

$$\hookrightarrow \lambda = \frac{(m_1 - m_2)t}{n_1 S_2 + 1}$$

Here, m_1 and m_2 are constants and S_1 and S_2 are defined as

$$S_1 = \sum_{x \in D} \frac{-1}{n_2 S_2 (m_1 - m_2)} \qquad S_2 = \sum_{x \in D} \frac{1}{n_1 (m_1 - m)(m_2 - m)^T}$$

Substitute λ into expression for w to get, find expression for

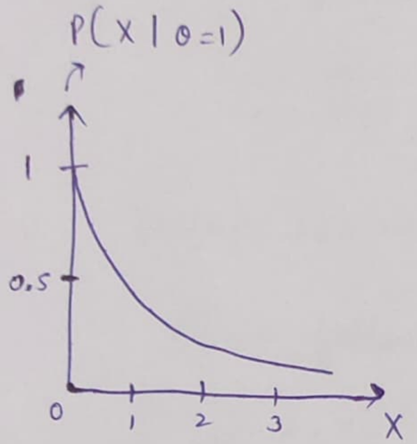
$$\boxed{w = \left(X^T X - \frac{(m_1 - m_2)t}{n_1 S_1 + 1} + I\right)^{-1}\left(yX - \frac{(m_1 - m_2)t}{n_1 S_1 + 1} + w_0\right)}$$

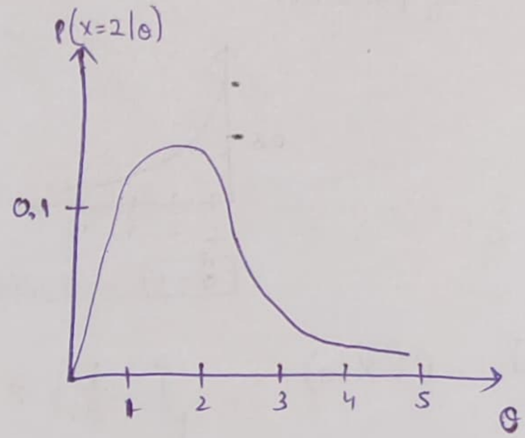
Q-5

1. $P(x|\theta) = \begin{cases} \theta e^{-\theta x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$

(a) (i) $P(x|\theta)$ vs x for $\theta = 1$

(ii) $P(x|\theta)$ vs θ $(2 \leq \theta \leq 5)$ for $X = 2$





(b) log likelihood $\Rightarrow$ $l(\theta) = \sum_{k=1}^{n} \ln p(x_k|\theta)$

$$= \sum_{k=1}^{N} [\ln\theta - \theta x_k] = n\ln\theta - \theta \sum_{k=1}^{n} x_k$$

$\nabla_\theta l(\theta) = 0$ for $\hat{\theta}$

$$\nabla_\theta l(\theta) = \frac{\partial}{\partial\theta}\left(n\ln\theta - \theta\sum_{k=1}^{n} x_k\right)$$

$$= \frac{n}{\theta} - \sum_{k=1}^{n} x_k = 0$$

$$\Rightarrow \boxed{\hat{\theta} = \frac{n}{\sum_{k=1}^{n} x_k}}$$

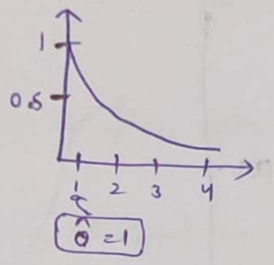
{ maximum likelihood function }

(c) Mean $= \frac{1}{n} \sum_{n=1}^{n} x_k$

Using Integration. $\int_0^\infty x\, p(x)\, dx$

$= \int_0^\infty x e^{-x} dx = 1$

Using these we can see that $\hat{\theta} = 1$, so plotting it on the graph of part (a)



2 $P(x|\theta) = \begin{cases} \frac{1}{\theta}, & 0 \leq x \leq \theta \\ 0 & \text{otherwise} \end{cases}$ distribution f^n

(a) Let us assume $I(.)$ indicator function which outputs 1.0 if condition is satisfied else $0 = 0$.

$P(D|\theta) = \prod_{k=1}^{n} p(x_k|\theta)$

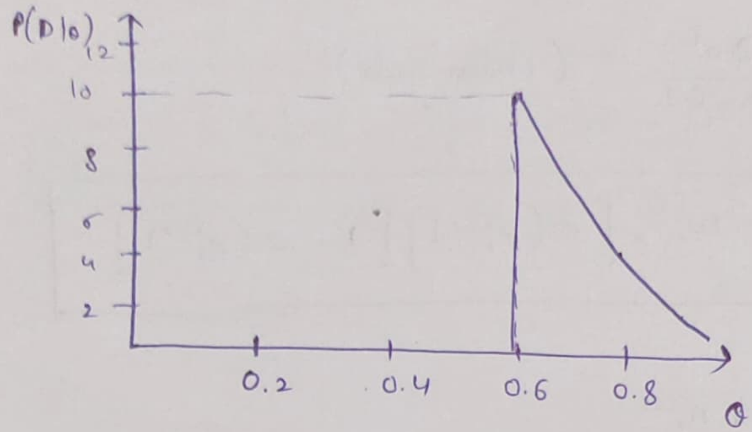
$= \prod_{k=1}^{n} \frac{1}{\theta} I(0 \leq x_k \leq \theta)$

$= \frac{1}{\theta^n} I(\theta \geq \max_k x_k) \, I(0 \leq \min_k x_k)$

$P(D|\theta)$ decreases as value of θ increases and also $I(\theta \geq \max_k x_k)$ will become zero if θ goes below max value of x_k.

The likelihood function is maximized at $\hat{\theta} = \max_{k}(X_k)$

(b)



Q-6 Discrete random variable with support $\{-2,-1,0,1,2\}$

$X \rightarrow$ follows uniform distribution

$$P(X=x) = \begin{cases} \frac{1}{n} & x \in \{-2,-1,0,1,2\} \\ 0 & \text{else} \end{cases}$$

$$= \begin{cases} \frac{1}{5} & x \in \{-2,-1,0,1,2\} \\ 0 & \text{else} \end{cases}$$

$$E(X) = \sum P(X=x) \cdot x$$

$$= (-2)\frac{1}{5} + (-1)\frac{1}{5} + 0\left(\frac{1}{5}\right) + 1\left(\frac{1}{5}\right) + \left(\frac{2}{5}\right) = 0$$

$$\Rightarrow \underline{E[X] = 0}$$

$$Var(X) = E[X^2] - (E[X])^2 = \left[\left(-\frac{2}{5}\right)^2 + \left(-\frac{1}{5}\right)^2 + (0)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2\right] - (0)^2$$

$$= \frac{4}{5} + \frac{1}{5} + \frac{1}{5} + \frac{4}{5}$$

$$= 2$$

$$\underline{Var(X) = 2}$$

Q-7 $f = w_1^{[3]} a_1^{[2]} + w_2^{[3]} a_2^{[2]}$; $\dfrac{\partial[\sigma(z)]}{\partial z} = [\sigma(z)][1-\sigma(z)]$

(i) $$\frac{\partial f}{\partial z_1^{[2]}} = \frac{\partial f}{\partial a_1^{[2]}} \cdot \frac{\partial a_1^{[2]}}{\partial z_1^{[2]}}$$ (chain rule)

$$\Rightarrow \boxed{\frac{\partial f}{\partial z_1^{[2]}} = w_1^{[3]} \cdot [\sigma(z_1^{[2]})][1-\sigma(z_1^{[2]})]}$$

(ii) $$\frac{\partial f}{\partial z_2^{[2]}} = \frac{\partial f}{\partial a_2^{[2]}} \times \frac{\partial a_2^{[2]}}{\partial z_2^{[2]}}$$

$$\Rightarrow \boxed{\frac{\partial f}{\partial z_2^{[2]}} = w_2^{[3]} \cdot [\sigma(z_2^{[2]})][1-\sigma(z_2^{[2]})]}$$

(iii) $$\frac{\partial f}{\partial w_{11}^{[1]}} = \left(\frac{\partial f}{\partial z_1^{[2]}} \cdot \frac{\partial z_1^{[2]}}{\partial z_1^{[1]}} \cdot \frac{\partial z_1^{[1]}}{\partial w_{11}^{[1]}}\right) + \left(\frac{\partial f}{\partial z_2^{[2]}} \cdot \frac{\partial z_2^{[2]}}{\partial z_1^{[1]}} \cdot \frac{\partial z_1^{[1]}}{\partial w_{11}^{[1]}}\right)$$

$$\frac{\partial f}{\partial w_{11}^{[1]}} = \left(\frac{\partial z_1^{[1]}}{\partial w_{11}^{[1]}}\right) \cdot \left[\frac{\partial f}{\partial z_1^{[2]}} \cdot \frac{\partial z_1^{[2]}}{\partial z_1^{[1]}} + \frac{\partial f}{\partial z_2^{[2]}} \cdot \frac{\partial z_2^{[2]}}{\partial z_1^{[1]}}\right]$$

<u>Answer</u> $\rightarrow$

$$\boxed{\frac{\partial f}{\partial w_{11}^{[1]}} = [x_1 . [\sigma(z_1^{[1]})] . [1-\sigma(z_1^{[1]})]] . \left[(w_{11}^{[2]}) . \sigma(z_1^{[2]})(1-\sigma(z_1^{[2]})) . w_1^{[3]} + (w_{21}^{[2]}) . (\sigma(z_2^{[2]}))(1-\sigma(z_2^{[2]})) . w_2^{[3]} \right]}$$

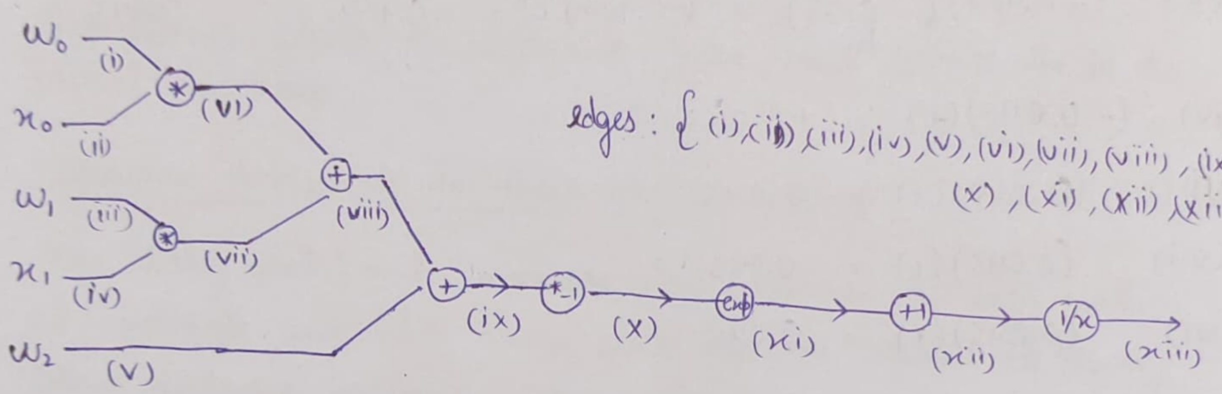
(iv) $$\frac{\partial f}{\partial w_{22}^{[1]}} = \left(\frac{\partial z_2^{[1]}}{\partial w_{22}^{[1]}}\right)\left[\frac{\partial f}{\partial z_1^{[2]}} \cdot \frac{\partial z_1^{[2]}}{\partial z_2^{[1]}} + \frac{\partial f}{\partial z_2^{[2]}} \cdot \frac{\partial z_2^{[2]}}{\partial z_2^{[1]}}\right]$$

Answer: $$\boxed{\frac{\partial f}{\partial w_{22}^{[1]}} = \left[x_2 [\sigma(z_2^{[1]})][1-\sigma(z_2^{[1]})]\right] \left[(w_1^{[3]}) . (w_{12}^{[2]})(\sigma(z_1^{[2]}))(1-\sigma(z_1^{[2]})) + (w_2^{[3]}) . (w_{22}^{[2]})(\sigma(z_2^{[2]}))(1-\sigma(z_2^{[2]})) \right]}$$

(V) $$\frac{\partial f}{\partial w_{12}^{[1]}} = \left(\frac{\partial z_1^{[1]}}{\partial w_{12}^{[1]}}\right) \cdot \left(\frac{\partial f}{\partial z_1^{[2]}} \cdot \frac{\partial z_1^{[2]}}{\partial z_1^{[1]}} + \frac{\partial f}{\partial z_2^{[2]}} + \frac{\partial z_2^{[2]}}{\partial z_1^{[1]}} \right)$$

Answer: $$\boxed{\frac{\partial f}{\partial w_{12}^{[1]}} = \left[x_2[\sigma(z_1^{[1]})][1-\sigma(z_1^{[1]})]\right] \cdot \left[(w_1^{[3]})(w_{11}^{[2]})(\sigma(z_1^{[2]}))(1-\sigma(z_1^{[2]})) + (w_2^{[3]})(w_{21}^{[2]})(\sigma(z_2^{[2]}))(1-\sigma(z_2^{[2]})) \right]}$$

Q-8.



edges : { (i), (ii), (iii), (iv), (V), (VI), (vii), (viii), (ix), (X), (xi), (xii), (xiii) }

Computational Graph

(a) Forward pass on the computation graph

$w = [w_0, w_1, w_2] = [1, 1, 1]$

$x = [x_0, x_1] = [1, 1]$

edges	Value
(vi)	$w_0 x_0 = 1$
(vii)	$w_1 x_1 = 1$
(viii)	$w_0 x_0 + w_1 x_1 = 2$
(ix)	$w_0 x_0 + w_1 x_1 + w_2 = 3$
(X)	$-(w_0 x_0 + w_1 x_1 + w_2) = -3$
(xi)	$e^{-(w_0 x_0 + w_1 x_1 + w_2)} = 0.0497$
(xii)	$1 + e^{-(w_0 x_0 + w_1 x_1 + w_2)} = 1.0497$
(xiii)	$\frac{1}{1+e^{-(w_0 x_0 + w_1 x_1 + w_2)}} = 0.952$

(i) $w_0 = 1$

(ii) $x_0 = 1$

(iii) $w_1 = 1$

(iv) $x_1 = 1$

(v) $w_2 = 1$

(b) local gradient times upstream gradient

edges $\qquad$ Value = (upstream gradient) x (local gradient)

(xiii) : 1

(xii) : $1\left[\frac{d}{dx}\left(\frac{1}{x}\right)\right] = \frac{-1}{x^2} = \frac{-1}{(1.0497)^2} = -0.907$

(xi) : $(-0.907)(1) = -0.907$

(x) $(-0.907)\left(\frac{d}{dx}e^x\right) = (-0.907)\,e^x = -0.907\,e^{-3} = -0.045$

(ix) $(-0.045)(-1) = +0.045$

(viii) $(0.045)(1) = 0.045$

(vii) $(0.045)(1) = 0.045$

(vi) $(0.045)(1) = 0.045$

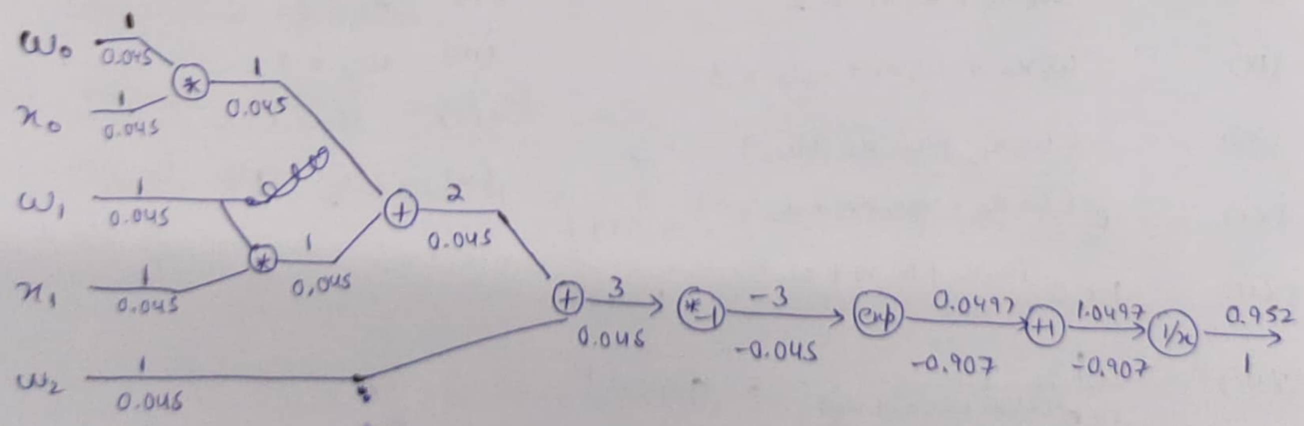
(v) $0.045(1) = 0.045$

(iv) $0.045(w_1) = 0.045$

(iii) $0.045(x_1) = 0.045$

(ii) $0.045(w_0) = 0.045$

(i) $0.045(x_0) = 0.045$



Q - 9. Steps of k means clustering algorithm.

① **Define the number of clusters (K)**

There's no single best value of K and hence it often involves some exploration and experimentation to find the optimal number of clusters that best represent the structure of data.

② **Initializing Random Centroids**

Randomly choose K datapoints from your dataset to be the initial centroids.

③ **Assign datapoints to their closest centroids**

For every datapoint, measure the euclidian distance with all centroids and label them with the one nearest to them. Other distance metrics such as Manhattan distance can also be used.

④ **Recompute the centroids corresponding to each cluster**

Once all data points have been labelled according to their nearest centroid, recompute new centroids for each cluster averaging the co-ordinates of all such datapoints belonging to the respective clusters.

⑤ **Repeat steps ③ and ④ until convergence**

Repeat steps 3 and 4 until the centroids no longer change or the convergence condition is achieved.

→ **Limitations of k means clustering algorithm**

① **Predefined number of clusters (K)**

The optimal value of K is many times not clearly visible from data but the K means clustering algorithm demands you to specify a value of K initially which is very challenging.

(2) Sensitive to initial centroids and outliers

Different initializations might lead to slightly different cluster assignments and also, the algorithm is highly effected by outliers as it tries to provide a definite label to every datapoint.

(3) Not Ideal for High Dimensional Data

K Means becomes less effective as the no. of dimensions in our data increases.

(4) Spherical Cluster Assumption

K Means works best for data forming roughly spherical clusters and is not very ideal for data with varying densities or any other shape.

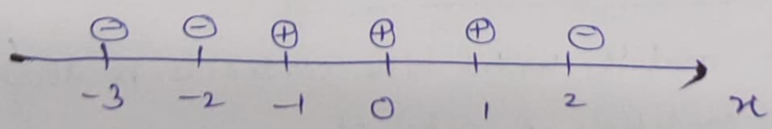
(5) Distance Based Clustering

K Means Clustering is solely based on distance as a similarity metric between two points neglecting any other attributes.

(6) No guarantee of optimal solution

K Means finds the local minima which not be the best solution globally. Therefore, you cannot be sure if the solution you have reached is optimal.

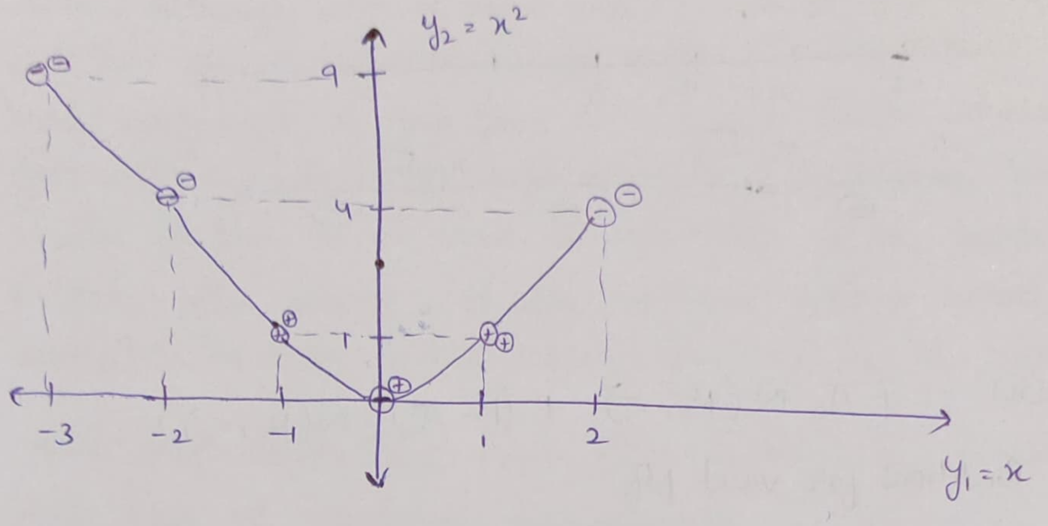
Q-10



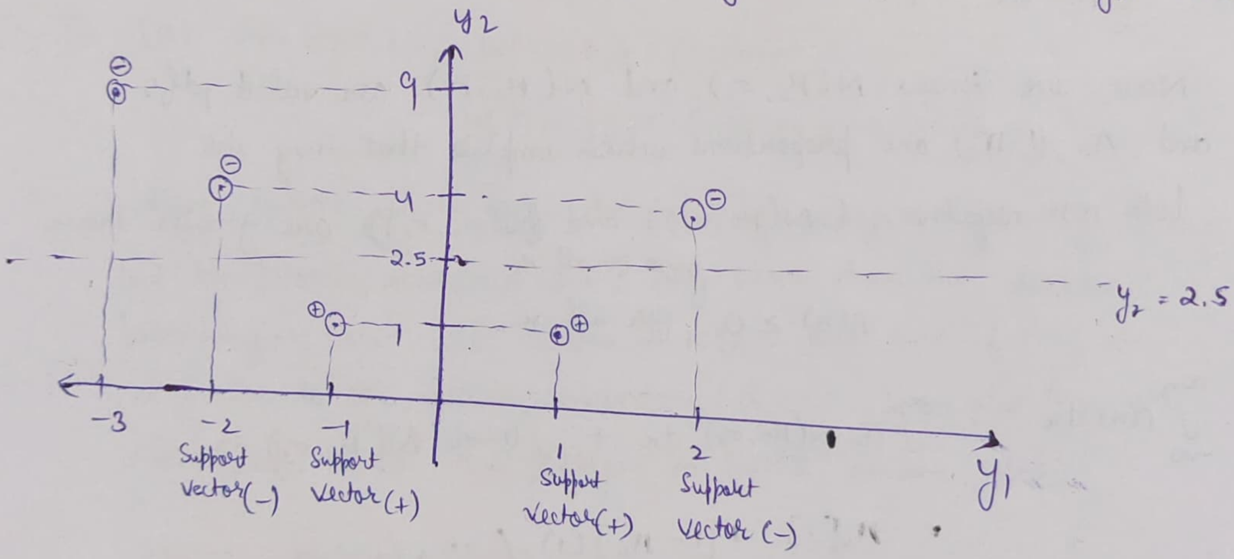
(a) Feature map to make the data linearly seperable is as follows: →

$$y_1 = x$$
$$y_2 = x^2$$

Plotting in 2D space $\{y_1, y_2\}$



(b) $w_2 y_2 + w_1 y_1 + w_0$ by hard margin linear SVM is $y_2 = 2.5$



Positive support vectors $\rightarrow$ $\{(-1,1), (1,1)\}$

Negative Support vectors $\rightarrow$ $\{(-2,4), (2,4)\}$

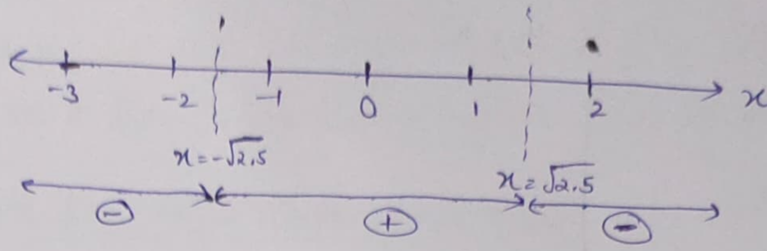
(c) Equivalent Decision Boundary in the original input space is

$x = \pm\sqrt{2.5} \Rightarrow x^2 = 2.5$

$x \in [-\sqrt{2.5}, \sqrt{2.5}]$: prediction is +ve

$x \notin [-\sqrt{2.5}, \sqrt{2.5}]$: prediction is -ve

Input space is 1D and hence the decision boundary is an interval.



Ques 11(b))

(a) $P(x) = \left(\pi_0 N(\mu_0, \sigma_0)\right) + (1-\pi_0)\left(N(\mu_1, \sigma_1)\right)$

Conditions for valid pdf

① $P(x) \geq 0$ for all x

② $\int_{-\infty}^{\infty} P(x)\, dx = 1$

Now, we know $N(\mu_0, \sigma_0)$ and $N(\mu_1, \sigma_1)$ are valid pdfs and $\pi_0, (1-\pi_0)$ are proportions which implies that they are both non-negative. $(N(\mu_0, \sigma_0))$ and $(N(\mu_1, \sigma_1))$ are greater than zero for all x.

$P(x) \geq 0$ for all x.

$$\int_{-\infty}^{\infty} P(x)\, dx = \int_{-\infty}^{\infty} \pi_0 N(\mu_0, \sigma_0)\, dx + \int_{-\infty}^{\infty} (1-\pi_0) N(\mu_1, \sigma_1)\, dx$$

$$= \pi_0[1] + [1-\pi_0](1) \quad (\because N(\mu_0, \sigma_0) \text{ and } N(\mu_1, \sigma_1) \text{ are valid pdfs.})$$

$$\int_{-\infty}^{\infty} P(x)\, dx = \pi_0 + (1-\pi_0) = 1$$

$P(x)$ is valid probability distribution function.

(b) The clustering results from K Means and GMM might not be the same. Although both of them will be able to find the clusters well but the cluster centres might differ because K Means uses hard assignment of each point to a specific cluster whereas GMM goes with a probabilistic soft assignment that causes GMM's cluster centres to be closer to each other as the points in K Means other cluster will also contribute with a small weight associated to them while calculating GMM's cluster center.

(c) Circle shifts to the left while stars shifts to the right in the next stage of expectation maximization

Q-12 (a) For SVM 1, $w^T x + b \geq 1$ (then $x \rightarrow +$)

$w^T x + b \leq -1$ (than $x \rightarrow -$)

There is some finite non zero value of margin given by $2/\|w\|$ but in SVM-2, margin is zero, any point above the decision boundary is (+) and below it is (−) and even if they are very close to the decision boundary ($d \rightarrow 0$), then also they are classified. Hence, the margin in SVM 2 can be zero.

Hence, Margin of SVM-1 > Margin of SVM 2 (0) [given data is linearly seperable]

(b) Now, if we take a SVM 3, with $w = \frac{w}{2}$ (SVM-1 trained weight)

$b = \frac{b}{2}$ (SVM-1 trained weight

We know that all points already satisfy,

$w^T x + b \geq 1$ (for +)

So, if we kept $w = w/2$ and $b = b/2$

Decision function equation becomes $\frac{1}{2}(w^T x + b) \geq \frac{1}{2}$ $\Rightarrow$ $f(x) \geq \frac{1}{2} > 0$ (for +)

$f(x) < 0$ (for −)

We assign points class ⊕ if $w^Tx+b>0$ and ⊖ otherwise, as changing the weight from $w \to w/2$ and $b \to b/2$ didnot alter the decision of labels of points as the sign ~~of~~ remains same,

Now, $w^Tx+b \geq \frac{1}{2}$ and $w^Tx+b \leq \frac{1}{2}$ are the new margins ie. magnitude of decision ~~boundary~~ function get scaled but sign is perserved. So, SVM 3 correctly identifies all training points because decision boundary remains same.

(c) As $w \to w/2$ in SVM-3 and Margin $= 2/\|w\|$

So, if margin of SVM-1 is $2/\|w\|$

Margin of SVM 3 $= \frac{2}{\|w\|/2} = \frac{4}{\|w\|} = 2\left(2/\|w\|\right)$

$\Rightarrow$ Margin of SVM 3 is twice that of SVM - 1