Name: Yyvaswatharnanu, 2019AIML616, PCAM7C34 Classification EC-R3C3

**Q:01** Use IG to get root node for DT - split.

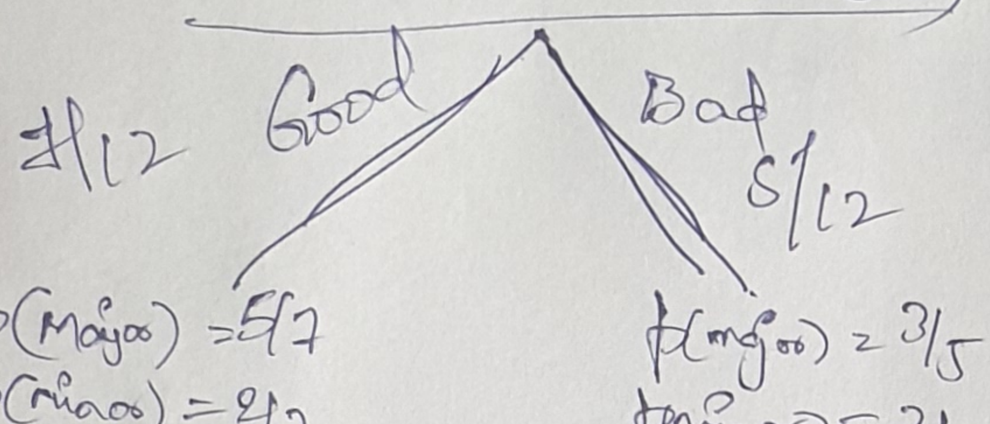
From the dataset, below sample parameters are deduced -

Crash sev. → Total = 12, Major = 8, Minor = 4

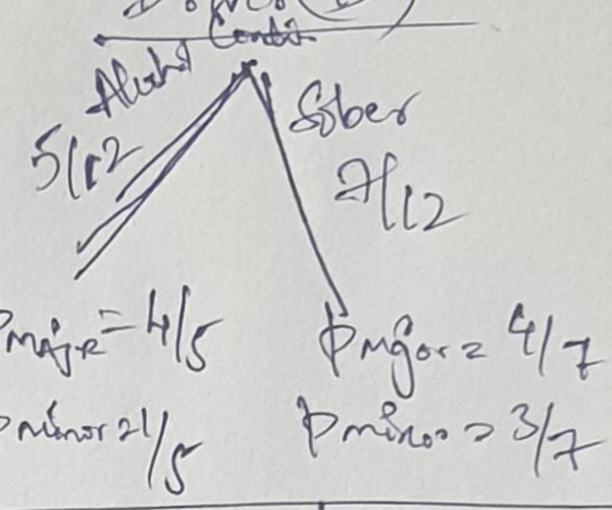
⇒ Entropy (Crash Sev.) = $-\sum_i p_i \log_2 p_i$ , $p_{major} = \frac{8}{12}$, $p_{minor} = 4/12$

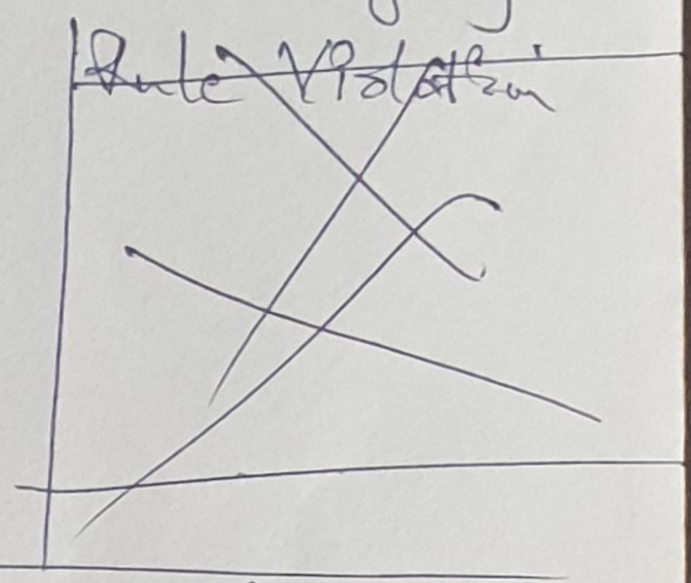
$\therefore E_C[\text{Entropy}_{Crash}] = -\left[\frac{8}{12}\log_2\left(\frac{8}{12}\right) + \frac{4}{12}\log_2\left(\frac{4}{12}\right)\right]$ [Henceforth $\log_2$ is written as $\log$]

**Weather Condition (W)**

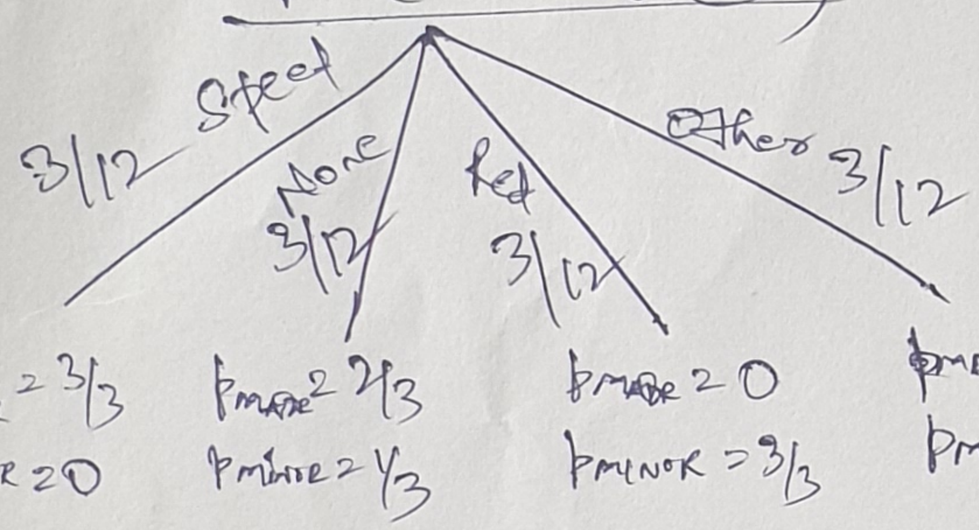



**Driver (D)**

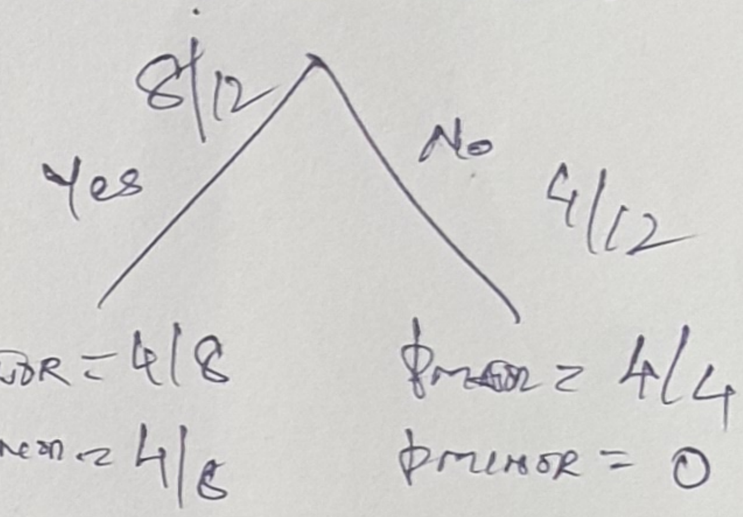



**Rule Violation (R)**




**Seat Belt (S)**




I.G = Info Gain (Attr. X) = Info Gain (Target) − Info Gain (X | Target)

$E_C$ = some +ve constant $= \left(\frac{8}{12}\log\frac{12}{8} + \frac{4}{12}\log\frac{12}{4}\right) = C$.

---

$\text{InfoGain}(W) = C - \left[\frac{7}{12}\left(\frac{5}{7}\log\frac{7}{5} + \frac{2}{7}\log\frac{7}{2}\right) + \frac{5}{12}\left(\frac{3}{5}\log\frac{5}{3} + \frac{2}{5}\log\frac{5}{2}\right)\right]$

$\text{InfoGain}(D) = C - \left[\frac{5}{12}\left(\frac{4}{5}\log\frac{5}{4} + \frac{1}{5}\log 5\right) + \frac{7}{12}\left(\frac{4}{7}\log\frac{7}{4} + \frac{3}{7}\log 7/3\right)\right]$

$\text{InfoGain}(R) = C - \left[\frac{3}{12}\left(1\cancelto{0}{\log 1} + 0\right) + \frac{3}{12}\left(\frac{2}{3}\log\frac{3}{2} + \frac{1}{3}\log 3\right) + \frac{3}{12}(0) + \frac{3}{12}(0)\right]$

$\text{InfoGain}(S) = C - \left[8/12\left(4/8\log 8/4 + 4/8\log 8/4\right) + 4/12\left(\cancelto{0}{\log 1} + 0\right)\right]$

Calculating ⇒ $I_W$ =

Vijayawatharan, 2019AIML618, PGAM7C34 Classification ECB3 C3

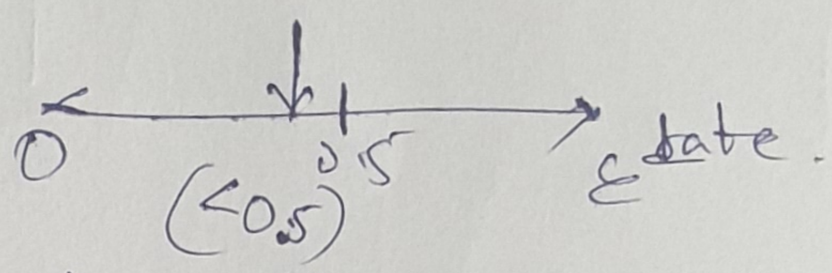
**Q:02** a) Ensemble methods are used for the foll. reason.

i) Avoid overfitting / help with overfitting

ii) Model / Algorithm is noisy & unstable (small changes have large impact)

iii) Classifiers have weak accuracy => high error rate

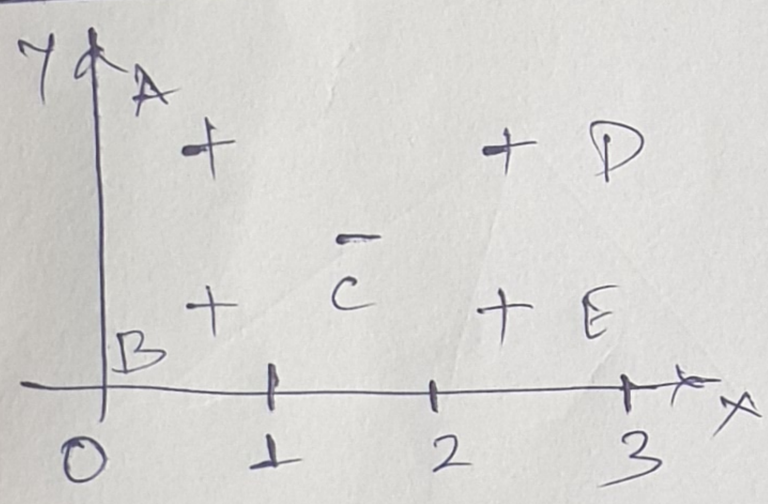
MAC: minimum error rate is better than random chance or coin flip, i.e. $\boxed{0.5}$




iv) Rationale is that weak set of 'experts' is better than a single expert / (Crowd is better than an individual)

---

**b) & c)** Will explain both together with an example.

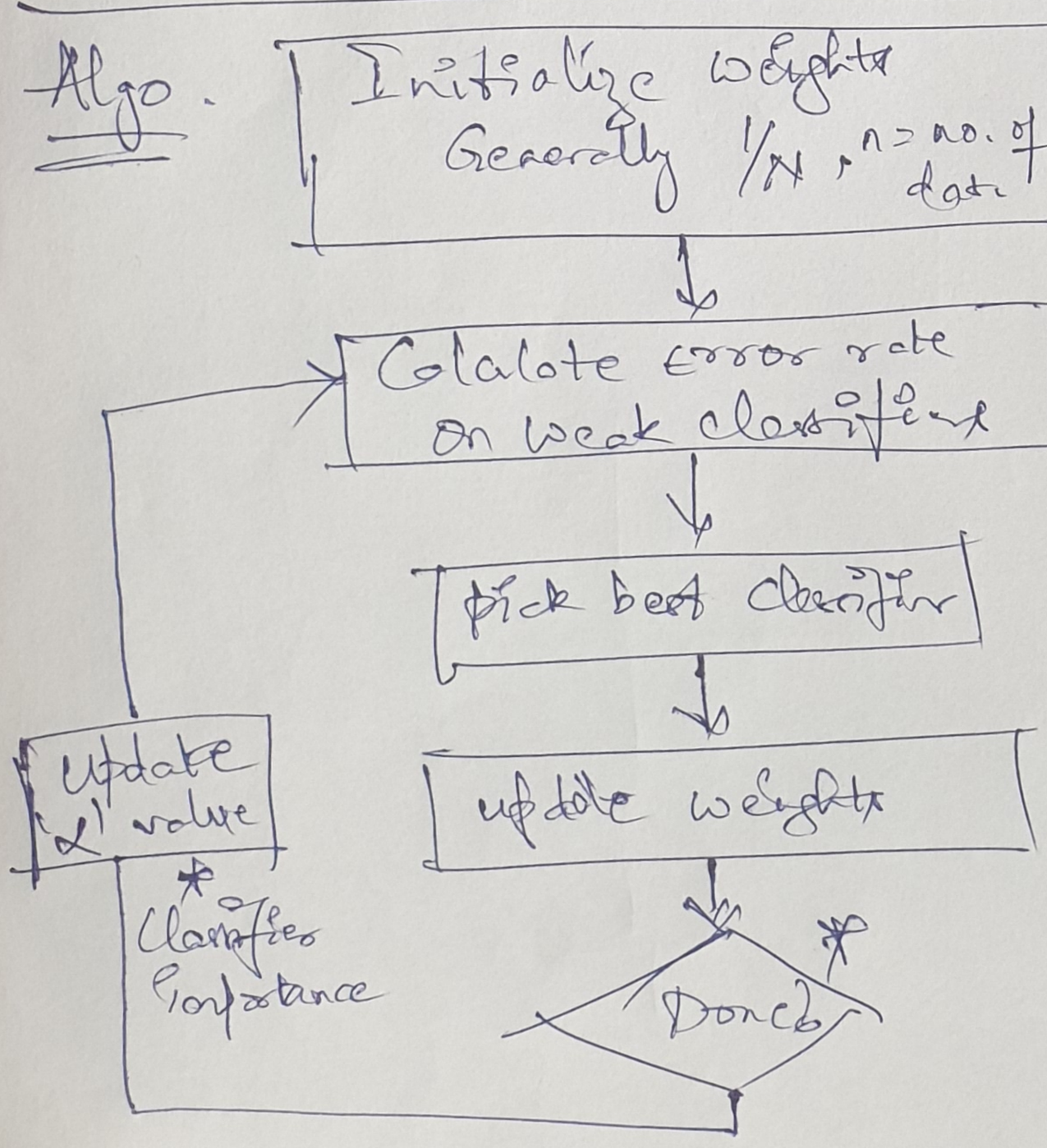



Consider the dataset on the left. Points are labelled A - E & classified as '+'ve & '−'ve, X axis is labelled 1, 2, 3.

=> Lets do adaboost on this toy problem

---

**Algo.**




* **Condition for done**

i) Best Error rate = 1/2

ii) Hypothesis is good enough

iii) Enough iterations done

**Weight update**

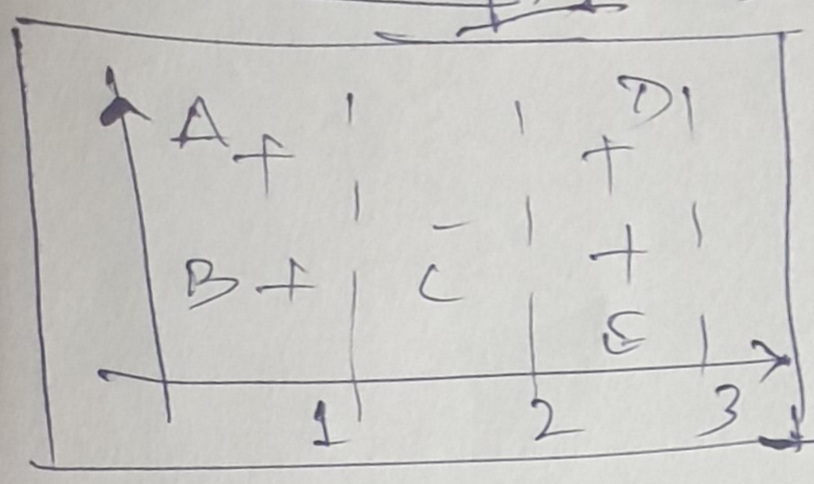
Done to 'boost' feature importance so that it is more emphasized to be classified correct in the next round / iteration

**α - Classifier importance**

Needed to pick best classifier output for final output.

Vyravattamonu, 2019AIML618, PGAM-IE 34 Classification EC-R3 CB

Q: 02 Contd. [b,c] Lets calculate errors & weights for each round.




Pick 3 classifiers, $x>1$, $x>2$, $x>3$ where in each, all samples to the left are considered +ve & right are -ve

Error rates

| classifier | Iter 1 | Iter 2 | Iter 3 | Wrong |
|---|---|---|---|---|
| $x>1$ | 2/5 | 1/4 | 1/2 | DE |
| $x>2$ | 3/5 | 3/4 | 5/6 | CDE |
| $x>3$ | 1/5 | 1/2 | 1/3 | C |

Weights

| points | Round 1 | Round 2 | Round 3 |
|---|---|---|---|
| A | 1/5 | 1/8 | 1/6 |
| B | 1/5 | 1/8 | 1/6 |
| C | 1/5 | 1/2 | 1/3 |
| D | 1/5 | 1/8 | 1/6 |
| E | 1/5 | 1/8 | 1/4 |

Note: Error rate = no. of wrong classifications. Round 1 weights are updated first and best classifier is picked with Error rate, $x>1$ which classifies D & E, $x>2 \to$ CDE, $x>3 \to$ C for $\boxed{\text{round 1}}$

the table iteratively updates for each round.

$$\alpha = \frac{1}{2}\ln\frac{1-\varepsilon}{\varepsilon}, \quad w^{t+1} = \frac{w^t}{Z} e^{-\alpha y_i(x) h_t(x)}$$

$h_t(x) \to$ prediction, $y_i(x) \to$ actual.

Deduce that, $\boxed{e^{-\alpha y_i h_i} = e^{-\frac{1}{2}\ln\frac{1-\varepsilon}{\varepsilon}\, y_i h_i}}$

$y_i h_i \to$ +ve for correct prediction & -ve for wrong

$$\Rightarrow w^{t+1} = \frac{w^t}{Z}\cdot\begin{cases}\left(-\frac{1}{2}\sqrt{\frac{1-\varepsilon}{\varepsilon}}\right) & \text{correct} \\ \left(-\frac{1}{2}\sqrt{\frac{\varepsilon}{1-\varepsilon}}\right) & \text{wrong.}\end{cases} \quad \text{classification}$$

Also. $\sum_{\text{correct}} w_t = \varepsilon \Rightarrow \sqrt{\frac{1-\varepsilon}{\varepsilon}}\sum_{\text{correct}} w_t + \sqrt{\frac{\varepsilon}{1-\varepsilon}}\sum_{\text{wrong}} w_t = Z$

$$\Rightarrow \sqrt{\frac{1-\varepsilon}{\varepsilon}}\cdot\varepsilon + \sqrt{\frac{\varepsilon}{1-\varepsilon}}\cdot(1-\varepsilon) = Z \Rightarrow Z = 2\sqrt{\varepsilon(1-\varepsilon)}$$

$$\Rightarrow w^{t+1} = \frac{w^t}{2}\cdot\frac{1}{1-\varepsilon} \text{ if correct } \& \ \frac{w^t}{2\varepsilon} \text{ for wrong classification}$$

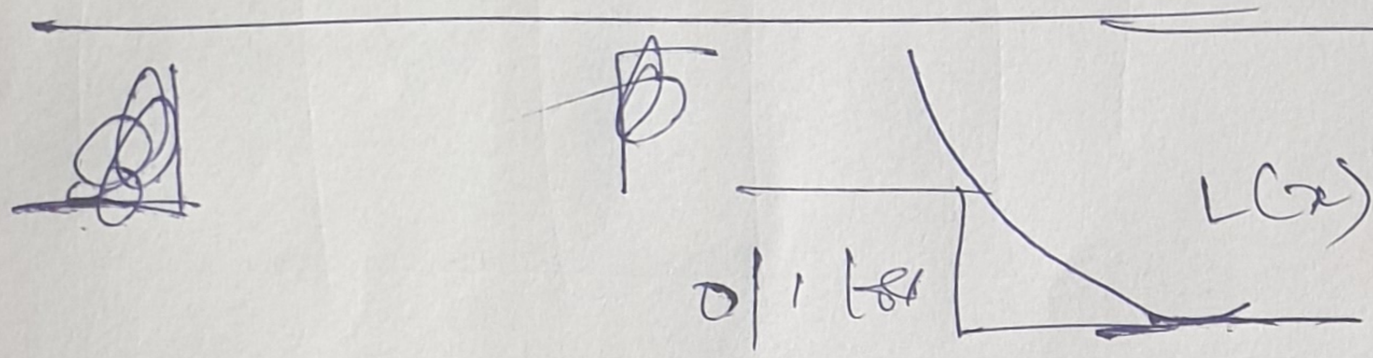
Vyjayanthi Ramani, 2019AIML618, PCAM ZC311 Classification EC R3 C3

Q.02 contd. b) c). So, using these formulae & updating the table, we get the values as seen.

The final output is $H(x) = \boxed{\alpha_1 \operatorname{sign}(h_1) + \alpha_2 \operatorname{sign}(h_2) + \alpha_3 \operatorname{sign}(h_3) \ldots}$

The '$\alpha$' is calculated by minimizing $C(x) = \sum_{\forall i} e^{-y^i \theta^T \phi(x^i)}$ then $\phi(x)$ is a feature mapping.

This is a general loss for Reproducing Kernel Hilbert Space (RKHS).




$L(x)$ = exp. loss.

Use co-ordinate descent to solve & get '$\alpha$'

$$\boxed{\alpha = \frac{1}{2} \ln \frac{1-\epsilon}{\epsilon}}$$

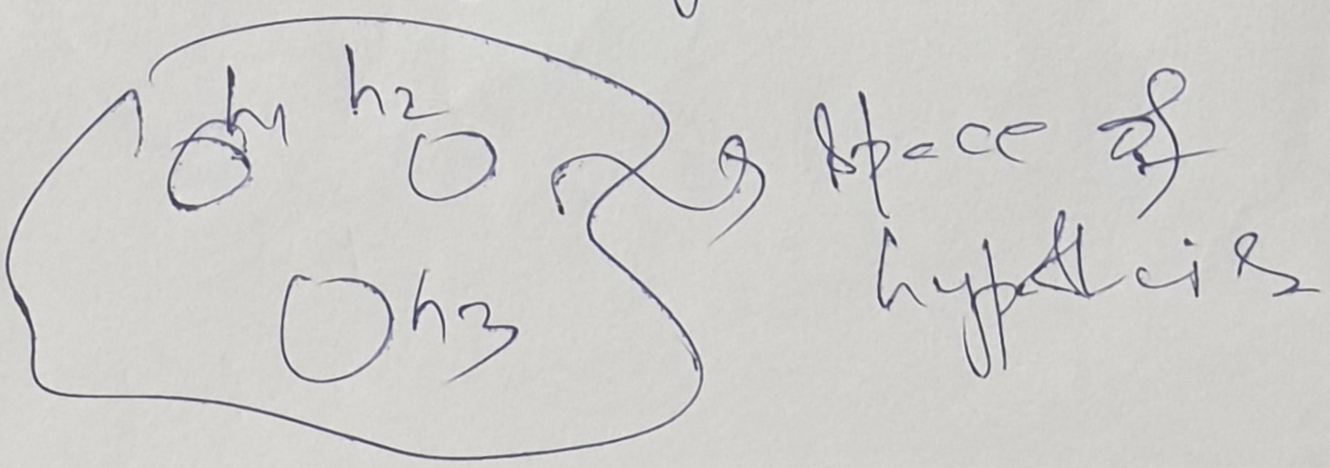
Round 1 → best class $= x > 3$ (0.2 error) $\Rightarrow \alpha = \frac{1}{2} \ln \frac{1 - 1/5}{1/5} = \boxed{\frac{1}{2} \ln 4}$

Round 2 → best class $= x > 1$ (0.25 error) $\Rightarrow \alpha = \frac{1}{2} \ln \left(\frac{1 - 1/4}{1/4}\right) = \boxed{\frac{1}{2} \ln 3}$

Round 3 → best class $= x > 3$ (0.33 error) $\Rightarrow \alpha = \frac{1}{2} \ln \left(\frac{1 - 1/3}{1/3}\right) = \boxed{\frac{1}{2} \ln 2}$

$$\Rightarrow H = \operatorname{sign}\left[\frac{1}{2} \ln 4 \left(\text{output of } x > 3\right) + \frac{1}{2} \ln 3 \left(\text{output of}\right) + \frac{1}{2} \ln 2 \left(\text{output of}\right)\right]$$

Disjoint sets of hypothesis will produce 100% accuracy.



Vyravarothamam, 2019AIML618, PGAM-IC 311 Classification ELR3 C3

**Q.02** a)

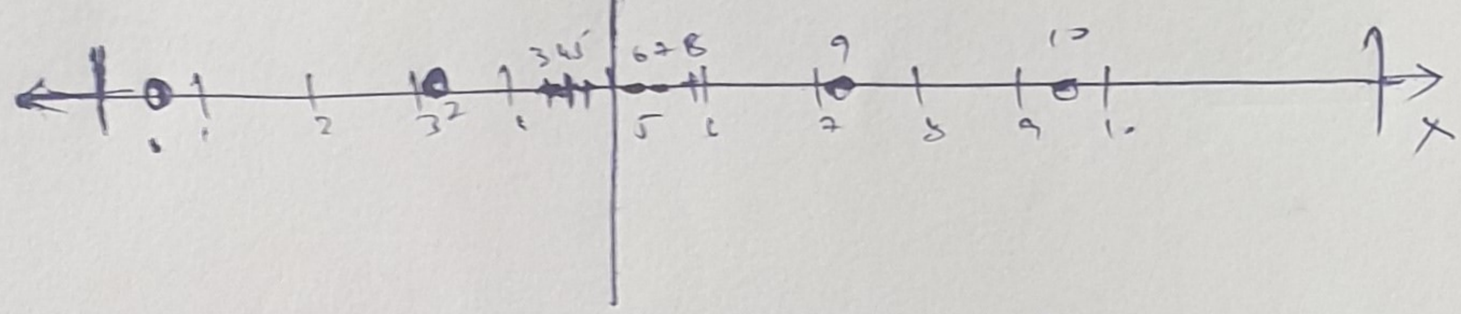
| probabilistic gener. | prob. discr. |
|---|---|
| (Generative) Naive Bayes | (Deterministic) Logistic Reg. / SVM |
| closed soln. | iterative soln. |
| (modelling) strong assumptions | weaker modelling assumptions |
| MAP / Baye's estimation | Likelihood / ML estimation |
| Linear / Quadratic boundary | Linear / non-linear (Kernels) boundary |
| (Linear → LDA, Quadratic → QDA) | Decision boundary is property of hypothesis. |
| Probability Distribution is property of hypothesis | |

b)

| x | y |
|---|---|
| 0.6 | − |
| 3.2 | − |
| 4.5 | + |
| 4.6 | + |
| 4.9 | + |
| 5.2 | − |
| 5.6 | − |
| 5.8 | + |
| 7.1 | − |
| 9.5 | − |

Legend [0 = −ve, + = +ve]

test point. x = 5




Lets use absolute $L_1$ norm for distance as data is (1-D)

$L_1 \rightarrow \left(\sum (x_i)^p\right)^{1/p}$ $p = 1$

$\Rightarrow |x_i - x_i| = L_1$-norm

| x | x−5 | Rank |
|---|---|---|
| 0.6 | 4.4 | 9 |
| 3.2 | 1.8 | 7 |
| 4.5 | 0.5 | 4 |
| 4.6 | 0.4 | 3 |
| 4.9 | 0.1 | 1 |
| 5.2 | 0.2 | 2 |
| 5.6 | 0.6 | 5 |
| 5.8 | 0.8 | 6 |
| 7.1 | 2.1 | 8 |
| 9.5 | 4.5 | 10 |

$K = 3 \Rightarrow [4.9, 5.2, 4.6]$

$= \underset{+/-}{\text{mode}} [+, +, -] = +$

$K = 5 \Rightarrow [4.9, 5.2, 4.6, 4.5, 5.6]$

$= \underset{+/-}{\text{mode}} [+, -, +, +, -] = +$

For both $K = 3$ & $5$ $+$ is output by our model.

Vyraswathamani, R019AIML618, Classification PCAM7C311
EC-R3C3

Q:03 Contd.

c) From the table

| | 1 | 0 |
|---|---|---|
| 1 | 50 | 115 |
| 0 | 72 | 5000 |

Actual

predicted

TP = 50 FP = 72

TN = 5000 FN = 115

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN}, \quad \text{precision} = \frac{TP}{TP+FP}, \quad \text{recall} = \frac{TP}{TP+FN} = TPR.$$

$$\text{Acc.} = \frac{50+5000}{50+5000+72+115} = \frac{5050}{5237} \simeq 96.43\%$$

$$\text{precision} = \frac{50}{50+72} \simeq 0.41, \quad \text{Recall} = \frac{50}{50+115} \simeq 0.303$$

F1 is defined as $\frac{2 \times \text{prec.} \times \text{recall}}{\text{precision} + \text{recall}} = \frac{2 \times 0.3 \times 0.41}{0.3 + 0.41}$

$$\Rightarrow \text{F1-score} = \frac{0.248}{0.71} \simeq 0.3485 \simeq 0.35 \text{ or } 35\%$$

---

When the situation is actually impacted → roughly

i) 41% of the time (precision) we are able to predict the correct situation.

ii) Out of the total situations in which situation is impacted, we are able to predict 30% (roughly) of the time (recall) correctly.

iii) Though accuracy is good, precision, recall, & F1-score (harmonic mean of prec & recall) are pretty bad.

Vyrosavathamam, PCAM 2C3U, Classification EC-R3 C3

**Q:04** a) $h_\theta(x) = 0.2$. In LogReg. $h_\theta(x)$ is the probability of classification

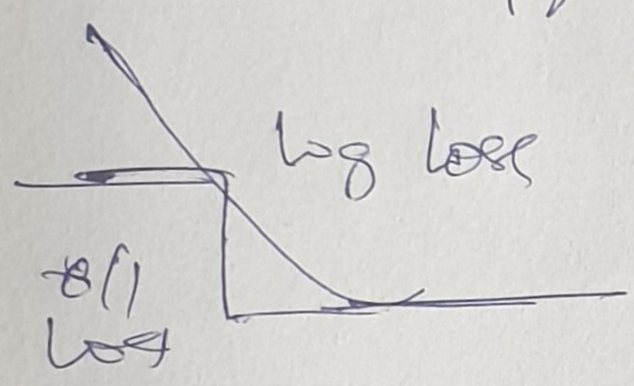
It is defined as.

$p(y^{(i)}=1 \mid x^{(i)}; \theta) = h_\theta(x)$ & $p(y^{(i)}=0 \mid x^{(i)}; \theta) = 1 - h_\theta(x)$

& we get $p(y^{(i)} \mid x^{(i)} \mid \theta) = (h_\theta(x^{(i)}))^{y^{(i)}} [1 - h_\theta(x^{(i)})]^{(1-y^{(i)})}$

Maximize log-likelihood $= \log \prod_{i=1}^{m} (p\, y^{(i)} \mid x^{(i)}; \theta)$

to get the cost function (cross entropy or log logistic loss)




$\therefore$ prob. of $x \in$ class 1 (belongs) $= 0.2$

$\Rightarrow$ prob. of $x \in$ class 0 $= 1 - 0.2 = \boxed{0.8}$

---

b) **N.B** $\Rightarrow$ $p(y=k \mid x) = \left[\prod_{i=1}^{n} p(\vec{x} \mid y=k) \cdot p(y=k)\right]$ (Max)

$k$ = classes, $n$ = no. of features

Calculate

$p(y=\text{High} \mid \vec{x})$, $\vec{x} = [\text{Age} \quad \text{Gender} \quad \text{Occupation}]$
$[\text{Above 40} \quad \text{M} \quad \text{Self-Emp}]$

$p(y=\text{Low} \mid \vec{x})$

& $p(y=\text{Middle} \mid \vec{x})$

$\text{argmax} \left[\prod_{i=1}^{B} p(x \mid y=k)\, p(y=k)\right]$

i) $p(y=\text{High} \mid \vec{x})$; $p(\text{Above-40} \mid \text{High}) \cdot p(\text{M} \mid \text{High}) \cdot p(\text{Self-Emp} \mid \text{High}) \cdot p(\text{High})$

$p(\text{High}) = 3/5$

$p(\text{above-40} \mid \text{high}) = 1/3$

$p(\text{M} \mid \text{high}) = 1/3$

$p(\text{self-emp} \mid \text{high}) = 2/3$

$\Rightarrow p(y=\text{High} \mid \vec{x})$ posterior $= \frac{1}{3} \cdot \frac{1}{3} \cdot \frac{2}{3} \cdot \frac{3}{5}$

$\boxed{p(y=\text{High} \mid \vec{x}) = 2/45}$

Vyavartha PGAM7C34 Cluster EC-R3 CS 2019-AIML 618

Q.04 Contd. b) ii) $P(y = \text{Middle} \mid \vec{x}) = P(\text{middle}) \cdot P(\text{above 40} \mid \text{middle}) \cdot P(M \mid \text{middle}) \cdot P(\text{Bat eng} \mid \text{middle})$

$P(\text{middle}) = 1/5$

$P(\text{above}40 \mid \text{middle}) = 0$, $P(M \mid \text{middle}) = 1$, $P(\text{Bat eng} \mid \text{middle}) = 0$

$$\Rightarrow \boxed{P(y = \text{middle} \mid \vec{x}) = \frac{1}{5} \cdot 1 \cdot 0 \cdot 0 = 0}$$

iii) $P(y = \text{Low} \mid \vec{x}) = P(\text{low}) \cdot P(\text{above}40 \mid \text{low}) \cdot P(\text{Bat eng} \mid \text{low}) \cdot P(M \mid \text{low})$

$P(\text{low}) = 1/5$, $P(M \mid \text{low}) = 1$

$P(\text{above } 40 \mid \text{low}) = 0$, $P(\text{Bat eng} \mid \text{low}) = 0$

$$\Rightarrow \boxed{P(y = \text{low} \mid \vec{x}) = 1/5 \cdot 1 \cdot 0 \cdot 0 = 0}$$

$$\underset{P}{\operatorname{argmax}} \left[ P(y = \text{high} \mid \vec{x}), P(y = \text{low} \mid \vec{x}), P(\text{middle} \mid \vec{x}) \right]$$

$= \text{argmax}(2/45, 0, 0) = 2/45$.

$\therefore$, The tuple will be classified as "High" to some bracket.

Note Laplace smoothing solution: $P(\text{middle} \mid m) = P(\text{low} \mid a)$

$= 1/5 \cdot 1 \cdot 1/3 \cdot 1/3 = 1/45$ which is still lesser than $2/45$ of high bracket.

$\boxed{P(0) \text{ was replaced with } 1/3 \text{ for smoothing.}}$

Nyrowethamanu, PCAM 7E-3U, Classification EC-R3-C3

**Q.05** a) Pros & Cons of SVM.

| Pros. | Cons. |
|---|---|
| → Very fast, always $O(n^2)$ as it only needs dot products (no. of samples) | → $\phi(x)$ → feature mapping can be expensive for high dimensions |
| → Can work with low dimensional data also, with Kernel trick | → Affected by noise |
| → Works when dimensions are too high with average dataset size | → Decision boundary affected by outliers. |
| → both linear & non-linear boundaries (Kernels in RKHS) | → Can struggle with very large data set size in computation. |

b) The optimization of SVM (with slack) is defined in general for RKHS as follows -

$$\text{max. dual} \left[ \text{min.} \ \frac{\|w\|^2}{2} + C\sum_{i=1}^{m} \varepsilon_i \right] \text{ subject to } (m = \text{no. of samples})$$

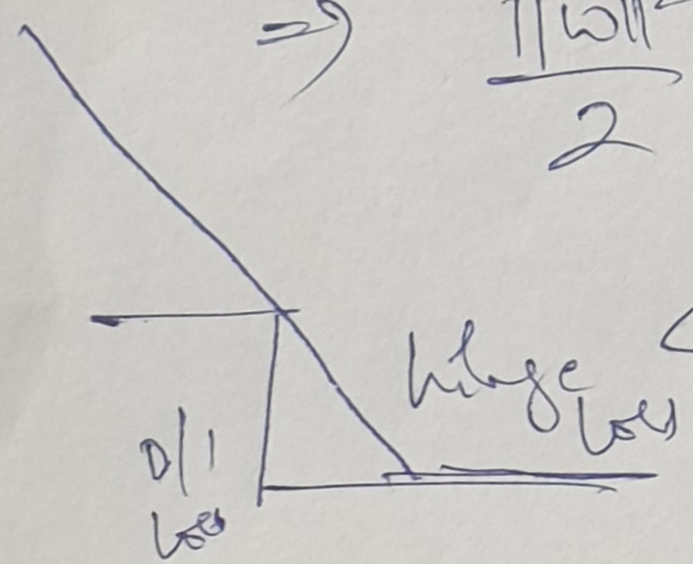
$$\text{s.t.} \quad y^{(i)}\left[(w^T x^{(i)} + b)\right] \geq 1 - \varepsilon_i$$

$$\varepsilon_i \geq 0$$

($C$ = regularization hyper-param)

Viewed in Regularized Empirical Risk Minimization framework (RERM)

$$\Rightarrow \frac{\|w\|^2}{2} + C\sum_{i=1}^{m} \max\left(0, 1 - y^{(i)}\left[w^T x^{(i)} + b\right]\right)$$




(or) minimize $\frac{\|w\|^2}{2} + C\sum_{i=1}^{m} \alpha_i \left(1 - \varepsilon_i - y^{(i)}\left[w^T x^{(i)} + b\right]\right) - \sum_{i=1}^{m} \beta_i \varepsilon_i$, s.t. $(\beta_i \varepsilon_i \geq 0)$ (KKT)

Vyroshetharan, PGA19ZC34, Classification ECR3 C3

Q.05 contd. Solving yields the dual

$$\max_{\alpha} \left[ \sum \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j (x_i \cdot x_j) \right] \text{ s.t. } \sum \alpha_i y_i = 0, \ \alpha_i \geq 0$$

Define $H_{ij} = y_i y_j (x_i \cdot x_j)$, where $y_i, y_j = (m \times 1)$ vector

$$\Rightarrow \max_{\alpha} \ \alpha^T i - \frac{1}{2} \alpha^T H \alpha. \quad \text{s.t. } \alpha^T y = 0$$

Q.P where $i = \vec{1}_{(m \times 1)}$ (vector of 1's)

solving $\Rightarrow \boxed{0 \leq \alpha \leq C}$

Where $H$ = Kernel / Gram matrix in RKHS.

$H = \phi_K(x^i)^T \phi_K(x^j)$ where $\phi(x)$ = the feature mapping of $x$ to RKHS

---

For Kernel to be valid, lets suppose $K$ is a valid Kernel $\Rightarrow$ $G(x, z) = \phi(x_i)^T \phi(z_i)$

$= \phi(z_i)^T \phi(x)$

Since we need dot products in RKHS, $G$ = Gram matrix (supposedly)

$\Rightarrow$ $G$ should be symmetric

Take any random vector $\vec{u}$

$$\Rightarrow u^T G u = \sum_i u_i (Gu)_i = \sum_i \sum_j u_i G_{ij} u_j$$

$$= \sum_i \sum_j \sum_k \left( u_i u_j \phi_k(x^i) \phi_k(x^j) \right)$$

$$= \sum_k \sum_i \sum_j u_i u_j \phi_k(x^{(i)}) \phi_k(x^j) = \boxed{\sum_k \left( \sum_i u_i \phi_k(x^i) \right)^2}$$

always $\geq 0$

Vyravathan, PGAM ZC311 Classification EC-R3 C3

→ Since $\vec{v}$ is arbitrary ⇒ G is always

Q:05 Contd

SYMMETRIC POSITIVE SEMI DEFINITE

Thus, any valid Kernel has a equivalent Gram or Kernel matrix which satisfies (& is a necessary & sufficient condition) of Mercer's Theorem.

---

Examples $K = (x^T z + c)^d$, $K = e^{\frac{\|x - z\|^2}{2\sigma^2}}$

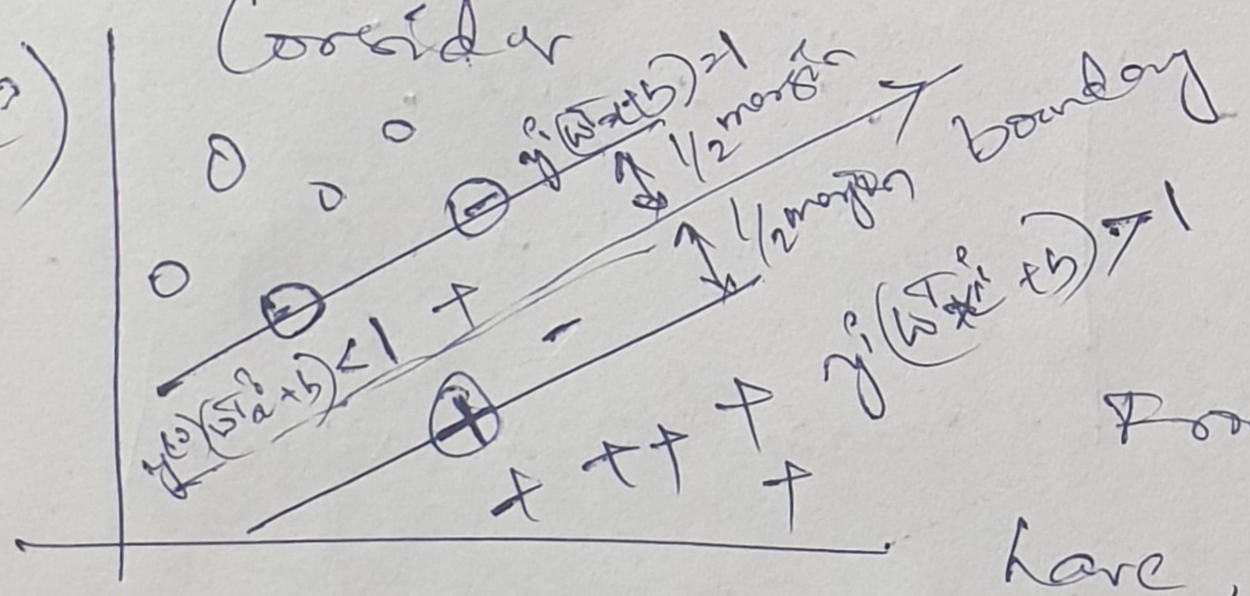
$K = \tanh(\beta x^T z + c)$

---

These Kernels convert lower dimentions to higher dimensions (Reproducing Kernel Hilbert Spaces) through a feature map $\phi$ & satisfy Mercer's theorem

---

c) Consider




Circled points are called support vectors.

---

From b) solution, we have, $y^{(i)}\left[w^T x^{(i)} + b\right] \geq 1 - \xi_i$

$\xi_i \geq 0$, $\beta_i \xi_i \geq 0$

$\alpha_i \geq 0$ & $0 \leq \alpha_i \leq C$

This means support vectors are the vectors on the margin deciding the decision boundary

---

i) $y^{(i)}\left[w^T x^{(i)} + b\right] = 1 \Rightarrow \alpha_i \geq 0, \xi_i = 0, \beta_i \geq 0, \alpha \leq C$

These are support vectors circled in the diagram

Nyvabodhanam PCAMZC311 Classification EC-R3C3

ii) $y^i(w^T x^i + b) < 1$, $\Rightarrow \xi_i \neq 0$ & $\beta_i = 0$ (for black) & $\alpha = C$

These are misclassified points

iii) $y^i(w^T x^i + b) \geq 1 \Rightarrow \alpha_i = 0$, $\xi_i = 0$ & $\beta_i \geq 0$

$\Rightarrow$ These are perfectly classified points

d) $x_1 = [1, 4]$, $x_2 = [-1, 4]$, $x_3 = [0, -1]$

$y_1 = 1$, $y_2 = 1$, $y_3 = -1$, $\alpha_1 = 0.65 = \alpha_2$, $\alpha_3 = 4.5$

$\boxed{\text{Question is wrong}}$ as $\sum_{\forall i} \alpha_i y_i = 0$ is not satisfied.

In dual Lagrange, soln. $L = \sum_{i=1}^{m} \alpha_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \langle x_i \cdot x_j \rangle$

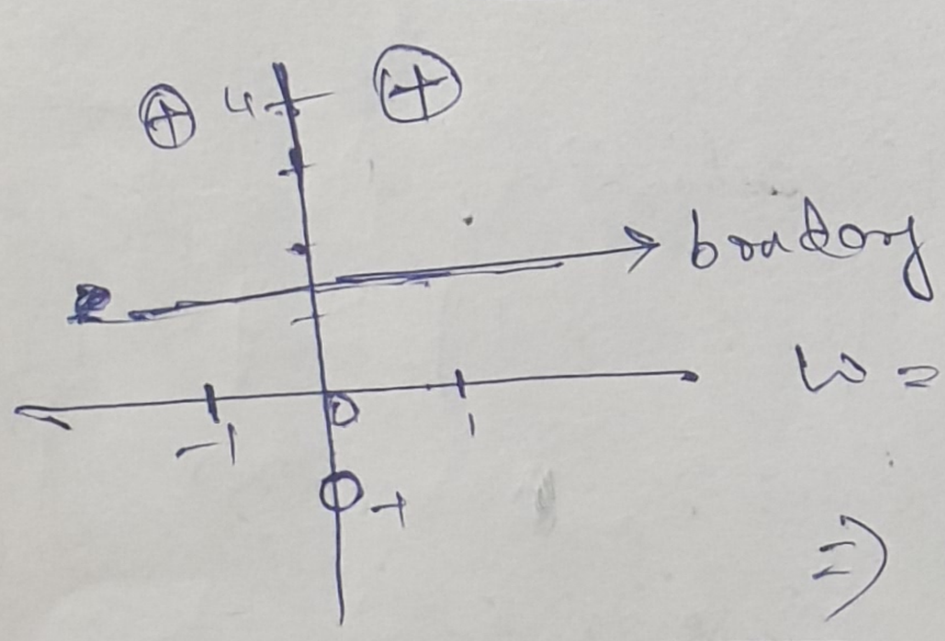
$w = \sum_{i=1}^{m} \alpha_i y_i (x_i \cdot x_j) = 0$. & $\sum_i \alpha_i y_i = 0$.

$x_1 \cdot x_1 = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} = 1 + 16 = 17$, $x_1 \cdot x_2 = x_2 \cdot x_1 = \begin{bmatrix} 1 \\ 4 \end{bmatrix} \begin{bmatrix} -1 \\ 4 \end{bmatrix} = -1 + 16 = 15$

$x_3 \cdot x_1 = x_1 \cdot x_3 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \begin{bmatrix} 1 \\ 4 \end{bmatrix} = 0 - 4 = -4$, $x_2 \cdot x_2 = \begin{bmatrix} -1 \\ 4 \end{bmatrix} \begin{bmatrix} -1 \\ 4 \end{bmatrix} = 1 + 16 = 17$

$x_3 \cdot x_3 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = 0 + 1 = 1$, $x_2 \cdot x_3 = x_3 \cdot x_2 = \begin{bmatrix} -1 \\ 4 \end{bmatrix} \begin{bmatrix} 0 \\ -1 \end{bmatrix} = 0 - 4 = -4$

$\Rightarrow L = (0.65 + 0.65 + 4.5) - \frac{1}{2} \left[ \begin{array}{l} 0.65 \cdot 0.65 \cdot 17 + 0.65 \cdot 0.65 \cdot 15 \\ + 0.65 \times 4.5 \times 4 + 0.65 \times 4.5 \times 4 \\ + 0.65 \times 0.65 \times 17 + 4 \times 4.5 \times 4.5 \end{array} \right]$




$w = \frac{4 - 1}{2} = 3/2 = 1.5$, $y = 1.5$

$\Rightarrow w = [0, 1.5]$, $b = -1.5$.

$wx + b = 0 \Rightarrow [0 \;\; 1.5] \begin{bmatrix} x \\ y \end{bmatrix} - 1.5 = 0$.