Date
Page No.

28th-March → **Classification**

Bayesian Method →

2 type of methods →

① Parametric → To estimate parameters of given distribution

② Non- "

① Parametric Method → Data follows some known distribution.

Duda-Hart → chapter ① & ②.

x_1 → length
x_2 → lightness. } → 2 features.

γ → classes → $\{w_1, w_2\}$.

→ $P_X(x/w_1)$ → Probability Density function for R.V. 'X' of lightness for a particular value 'x' given class 'w_1'. (class-conditional)

Conditional Densities

→ $P(w_1/x)$ → Given the feature 'x', what is the probability that it belongs to 'w_1' class.

$P(w_2/x)$ → "

$P(w_1)$ → Probability of fish being salman(w_1)
$P(w_2)$ → " " " " sea-bass(w_2).

Prior probabilities

Date
Page No.

So, Given a feature 'x' →

Bayesian formula → $P(\omega_j/x) = \frac{P(x/\omega_j) * P(\omega_j)}{P(x)}$

& for 2 categories → $P(x) = \sum_{j=1}^{2} P(x/\omega_j) \cdot P(\omega_j)$

Bayesian formula can be informally expressed in english as →

$$\text{Posterior} \Rightarrow \frac{\text{likelihood} \times \text{Prior}}{\text{evidence}}$$

Decision Criteria → whichever is greater.

$P(\omega_1/x) \gtrless P(\omega_2/x) \Rightarrow$ class ω_i.

$$P(\text{error}/x) = \begin{cases} P(\omega_1/x) & ; \text{ if we decide } \omega_2 \\ P(\omega_2/x) & ; \text{ " " " } \omega_1 \end{cases}$$

for multiclass classification →

if $P(\omega_j/x) \gtrless P(\omega_i/x)$; $\forall i = 1, 2, \ldots K$, $i \neq j$

⇒ Average probability of error →

$$P(\text{error}) = \int_{-\infty}^{+\infty} P(\text{error}, x)\, dx = \int_{-\infty}^{+\infty} P(\text{error}/x) \cdot P(x)\, dx$$

Thus → $P(\text{error}/x) = \min[P(\omega_1/x), P(\omega_2/x)]$

& $\boxed{P(\omega_1/x) + P(\omega_2/x) = 1}$; { for binary classification }

(or) Decide ω_1 if → $\boxed{P(x/\omega_1).P(\omega_1) > P(x/\omega_2).P(\omega_2)}$;

otherwise decide ω_2

Decision based on conditional & prior probability.
(evidence is same).

→ Basically bayesian Decision rule combines conditional & prior probability to achieve minimum probability of error.

⇒ We can extend it for 'c' classes.

04th Apr

* ⇒ The results that we have seen in case of bayesian approach can be biased. Therefore, we need to attach another factor in that

e.g. $\alpha_1 \rightarrow w_1$, $\alpha_2 \rightarrow w_2$, $\alpha_j \rightarrow w_j$ } { (e.g. w_1 = ~~coal~~ diamond, w_2 = coal } ⇒ (α_i) is the action to classify observation (x) to class (w_i).

Loss w.r.t. actions taken →

① $\lambda(\alpha_1/w_1) = 0$ → if accurately classified.

② $\lambda(\alpha_1/w_2) = 1$ → ~~all wrongly classified~~. loss regarding, If stone was coal & it is thrown in the bucket of diamond

③ $\lambda(\alpha_2/w_1) = K$ → vice-versa of ②, very high loss.

④ $\lambda(\alpha_2/w_2) = 0$ → accurate.

Conditional Risk

$$R(\alpha_i/x) = \sum_{j=1}^{c} \lambda(\alpha_i/w_j) P(w_j/x)$$

basically → Risk involved ; loss values ; Probability of assignment.

⇒ and we'll make the decision with min.m risk.

overall Risk ⇒ $R(\alpha_i/x) = \sum_{j=1}^{c} \lambda(\alpha_i/\omega_j) \cdot P(\omega_j/x)$

If → then → $R = \int R(\alpha_i/x) \cdot P(x) \cdot dx$: {where dx is our notation for a d-space volume element - where the integral extends over the entire feature space

overall Risk.

→ Suppose we are give 2 classes →

2- Category classification →

$R(\alpha_1/x) = \lambda_{11} \cdot P(\omega_1/x) + \lambda_{12} \cdot P(\omega_2/x)$

$\lambda_{11} \to \lambda(\alpha_1/\omega_1)$, $\lambda_{12} \to \lambda(\alpha_1/\omega_2)$

Similarly → $R(\alpha_2/x) = \lambda_{21} \cdot P(\omega_1/x) + \lambda_{22} P(\omega_2/x)$

$\lambda_{21} \to \lambda(\alpha_2/\omega_1)$, $\lambda_{22} \to \lambda(\alpha_2/\omega_2)$

So, if → $R(\alpha_1/x) \leq R(\alpha_2/x)$

↓

this will be selected

⇒ $\lambda_{11} \cdot P(\omega_1/x) + \lambda_{12} \cdot P(\omega_2/x) \leq \lambda_{21} \cdot P(\omega_1/x) + \lambda_{22} \cdot P(\omega_2/x)$

↓

this will form the decision-criteria.

04th April →

Decision Criteria → Decide ω_1 if $R(\alpha_1/x) < R(\alpha_2/x)$,

otherwise ω_2. So → if $R(\alpha_1/x) < R(\alpha_2/x) \Rightarrow \omega_1$

To decide ω_1 → ~~ω_2~~ $(\lambda_{21} - \lambda_{11}) \cdot P(\omega_1/x) > (\lambda_{12} - \lambda_{22}) \cdot P(\omega_2/x)$

$$(\lambda_{21} - \lambda_{11}) \cdot P(x/\omega_1) \cdot P(\omega_1) > (\lambda_{12} - \lambda_{22}) \cdot P(x/\omega_2) \cdot P(\omega_2)$$

Positive factor (+ve)

Also, under assumption that → $\lambda_{21} > \lambda_{11}$ → we can say that:

* Likelihood Ratio ←

$$\frac{P(x/\omega_1)}{P(x/\omega_2)} > \frac{(\lambda_{12} - \lambda_{22}) \cdot P(\omega_2)}{(\lambda_{21} - \lambda_{11}) \cdot P(\omega_1)} ;$$

& we can consider $P(x/\omega_j)$ as a fn of ω_j, which is likelihood fn.

Date
Page No.

Minimum error-rate classification →

if action is α_i & true state of nature is ω_j, then provided, the cond.^{n} that we are using symmitrical (or) 0/1 loss fn.

$$\lambda(\alpha_i/\omega_j) = \begin{cases} 0 ; & i=j \\ 1 ; & i \neq j \end{cases}$$

So → Conditional Risk ⇒

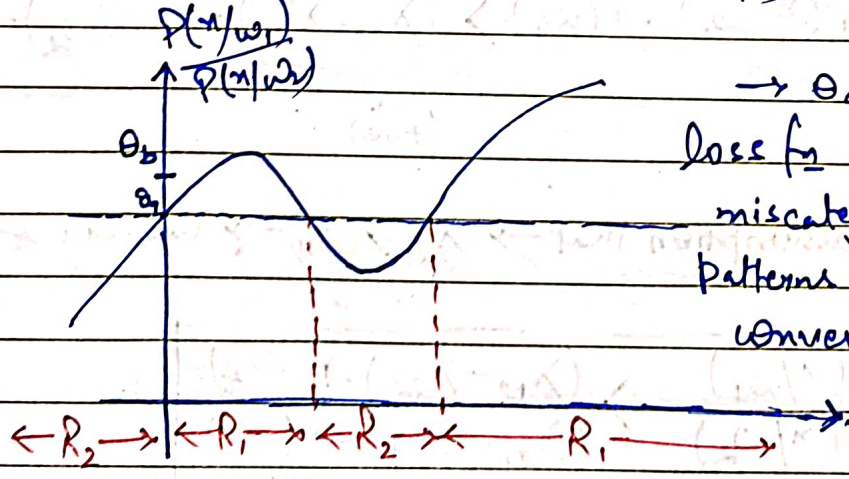
$$R(\alpha_i/x) = \sum_{j=1}^{c} \lambda(\alpha_i/\omega_j) \cdot P(\omega_j/x)$$

$$\Rightarrow \sum_{j \neq i} P(\omega_j/x) \; ; \; \{\text{in case of 0/1 loss}\}$$

$$\Rightarrow \{1 - P(\omega_i/x)\}$$

using other way → we can maximise the posterior prob. for decision criteria.

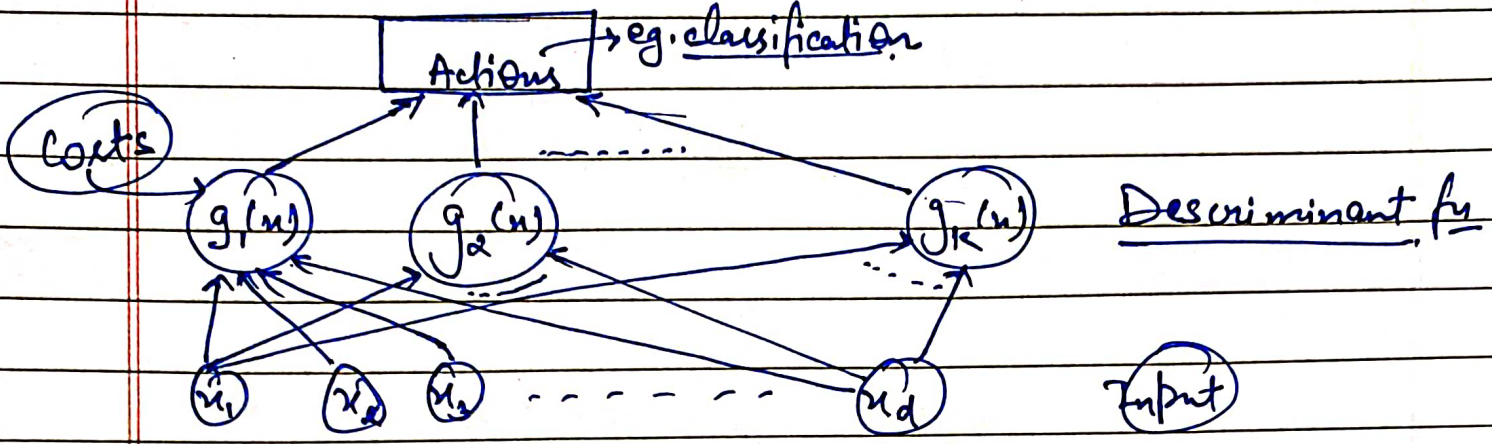
Select (ω_i) if $P(\omega_i/x) > P(\omega_j/x)$; for all $j \neq i$



→ θ_a is threshold, if loss fn penalises miscategorizing ω_2 as ω_1 patterns more than the converse (ie. $\lambda_{12} > \lambda_{21}$), we get larger threshold θ_b and hence Range of x values for which we classify a pattern as ω_i gets smaller.

Date
Page No.

Classifiers & Discriminant functions →



classifier is said to assign a feature vector 'x' to class 'ω_i'. if →

$$g_i(x) > g_j(x) \text{ ; for all } j \neq i$$

→ Minimum error rate classification →

$$R(\alpha_i / x) = 1 - P(\omega_i / x)$$

→ all the model displayed in slide, comes under umbrella of generating models.

⇒ $g_1(x)$, $g_2(x)$, ..., $g_K(x)$

$$\boxed{\arg \max_k \left(g_i(x) \right)}$$

most probably o/p class

So → $g_j(x) = -R(\alpha_j/x)$ ← framework.

$\Rightarrow -\left(1 - P(w_j/x)\right)$: { check on what cond.ⁿ, this is true }.

So → if $g_i(x) > g_j(x)$; if $i \neq j$
& $i = j = 1, 2, \ldots k$.

Then we'll select class 'i'.

$$g_i(x) = P(w_i/x) = \frac{P(x/w_i) \cdot P(w_i)}{P(x)}$$

$$g_j(x) = \frac{P(x/w_j) \cdot P(w_j)}{P(x)}$$ → ignore this.

→ Take log →

$g_i'(x)$ ← $\log g_i(x) = \log P(x/w_i) + \log(P(w_i))$ ——— (1)

$g_j'(x)$ ← $\log g_j(x) = \log P(x/w_j) + \log P(w_j)$

⇒ Now, how to calculate → $P(x/w_i)$ & $P(x/w_j)$

Decision Boundries are formed.

$$\Rightarrow \boxed{g(x) = g_1(x) - g_2(x)}$$

So, if $g(x) > 0 \Rightarrow w_1$

otherwise $\Rightarrow w_2$

Decision Boundry → $g(x) = g_1(x) - g_2(x)$

So → $g(x) = P(\omega_1/x) - P(\omega_2/x)$

$$g(x) = \ln\left\{\frac{P(x/\omega_1)}{P(x/\omega_2)}\right\} + \ln\left\{\frac{P(\omega_1)}{P(\omega_2)}\right\}$$

→ Descriminant fn for Normal Density →

So → if $x \in R^d$

&

$P(x/\omega_i) \sim N(\mu_i, \Sigma_i)$

So →

$$P(x/\omega_i) = \frac{1}{(2\pi)^{d/2} \cdot (\Sigma_i)^{1/2}} \cdot \exp\left\{\frac{-1}{2}(x-\mu_i)^t \cdot \Sigma_i^{-1}(x-\mu_i)\right\}$$

$$g_i(x) = \frac{-1}{2}(x-\mu_i)^t \cdot \Sigma_i^{-1} \cdot (x-\mu_i) - \frac{d}{2}\ln(2\pi) - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)$$

Date
Page No.

if we have 'c' classes →

such that:-

$\omega_1 \rightarrow \mu_1, \Sigma_1$

$\omega_2 \rightarrow \mu_2, \Sigma_2$

$\vdots$

$\omega_c \rightarrow \mu_c, \Sigma_c$

There are multiple cases, that can occur →

Case① → $\boxed{\Sigma_i = \sigma^2 . I}$ → when features are statistically independent. each feature has same variance σ^2 & co-variance of 2 features are 0.

in 2-D: $x \in \mathbb{R}^2$. So →

$$\Sigma_i = \begin{bmatrix} \sigma_i^2 & 0 \\ 0 & \sigma_i^2 \end{bmatrix}$$

Note → ①. if variance in X & Y directions are same & co-variance is 0, then contours will be co-centric circles.

②. if co-variance is 0 but variance in both directions are different, then contours will be ellipse.

③. if co-variance is also (+ve), the contours will be tilted ellipse.

See book for all 3-cases →

Duda-Hart ~~Page~~

→ chapter 2 - (2.6).

Q → Convert the discriminant fn into a linear Discriminant fn.

Date
Page No.

So, Discriminant fn for case(i) is →

$$g_i(x) = -\frac{1}{2}\frac{(x-\mu_i)^t}{\sigma^2 I} - \frac{1}{2}\ln|\sigma^2 I| + \ln P(\omega_i)$$

⇓

$$g(x) = g_1(x) - g_2(x)$$

Date
Page No.

(or) Similarly :→

$$g_2(x) = \frac{-1}{2}\frac{(x-\mu_2)}{\sigma^2} - \frac{1}{2}\cdot\ln|\sigma^2| + \ln P(\omega_2)$$

07/04/22 Assignment → $\Sigma_1 = \Sigma_2 = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$

$$g_1(x) = \log P(\omega_1/x) = \frac{P(x/\omega_1) * P(\omega_1)}{P(x)}$$

So → $\log g_1(x) = \log P(x/\omega_1) + \log P(\omega_1)$

Similarly → $\log g_2(x) = \log P(x/\omega_2) + \log P(\omega_2)$

$$g(x) = g_1(x) - g_2(x)$$

→ if $g(x) \geq 0 \Rightarrow$ belongs to $g_1(x)$

else $\Rightarrow$ belongs to $g_2(x)$

$$g_1(x) = \log\left(\frac{1}{(2\pi)^{d/2}\cdot\Sigma^{1/2}}\right) \exp\left(-(x-\mu_1)^T\cdot\Sigma^{-1}(x-\mu_1)\right) + \log P(\omega_1)$$

$\Rightarrow$ (expand)

Similarly for $g_2(x) \Rightarrow$

$$\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$$

$$\Sigma^{-1} = 1/\sigma^2$$

Date
Page No.

$$g_1(x) = \frac{-1}{2\sigma^2} \cdot (x^T x - 2\mu_1^T x + \mu_1^T \mu_1) + \log P(\omega_1)$$

$$g_2(x) = \frac{-1}{2\sigma^2} (x^T x - 2\mu_2^T x + \mu_2^T \mu_2) + \log P(\omega_2)$$

So → $g(x) = g_1(x) - g_2(x) = 0$

$$\boxed{g_1(x) = \left(\frac{1}{\sigma^2}\mu_1\right)^T x - \frac{1}{2\sigma^2}\mu_1^T\mu_1 + \log P(\omega_1)}$$

So →

$$g(x) = g_1(x) - g_2(x) = 0$$

then, we'll get sol.n →

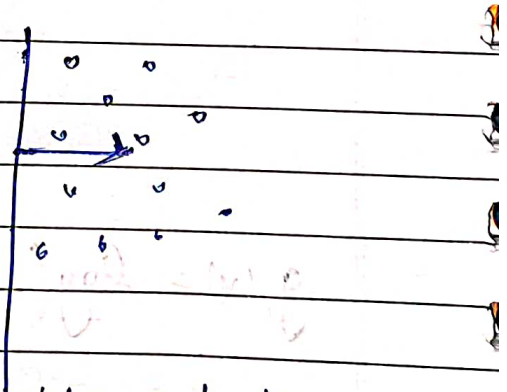
$$g(x) \Rightarrow \boxed{w^T(x - x_0) = 0} \rightarrow \text{and } w = (\mu_i - \mu_j)$$

$$\& \; x_0 = \frac{1}{2}(\mu_i + \mu_j) - \sigma^2 / \|\mu_i - \mu_j\|^2 \cdot \ln\frac{P(\omega_i)}{P(\omega_j)} \cdot (\mu_i - \mu_j)$$

example from assign. (03) →

$$g_1(x) = -\frac{\|x - \mu_1\|_2}{2\sigma^2} + \log P(\omega_1)$$

$$g_2(x) = -\frac{\|x - \mu_2\|_2}{2\sigma^2} + \log P(\omega_2)$$



Mahalanobis Distance

$$x_0 = \frac{1}{2}(\mu_i + \mu_j) - \frac{\sigma^2}{\|\mu_i - \mu_j\|^2} \ln\frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

$\Sigma_i = \Sigma$ → variance is same in both the dir.ⁿ but co-variance exists.

Case (2) → $g_i(x) = w_i^T x + w_{i0}$

$w_i = \Sigma^{-1} \mu_i$

$w_{i0} = -\frac{1}{2} \mu_i^T \Sigma^{-1} \mu_i + \log P(w_i)$

$g(x) = g_1(x) - g_2(x) = 0$

$\Rightarrow w^T (x - x_0)$; { Decision criteria }.

$w = \Sigma^{-1} (\mu_1 - \mu_2)$

& $x_0 = \frac{1}{2}(\mu_1 + \mu_2) - \frac{\ln [P(w_1)/P(w_2)] (\mu_1 - \mu_2)}{(\mu_1 - \mu_2)^t \cdot \Sigma^{-1} \cdot (\mu_1 - \mu_2)}$

missing, it's min. Complete

Case (3) → arbitary →

$g_i(x) = x^t W_i x + w_i^t x + w_{i0}$

$W_i = -\frac{1}{2} \cdot \Sigma_i^{-1}$

$w_i = \Sigma_i^{-1} \cdot \mu_i$

and → $w_{i0} = -\frac{1}{2} \mu_i^t \Sigma_i^{-1} \mu_i - \frac{1}{2} \ln |\Sigma_i| + \ln P(w_i)$

Similarly → $g_2(x) =$ ——

& → $g(x) = g_1(x) - g_2(x) = 0$

Que. → What is generative model?

Que. → " " Descriminant fn. / model?

12-April

→ Regression Methods for binary classification →

we'll try to contain the regression values b/w (0 & 1). So, that the o/p. values denote the probability values.

$X \xrightarrow{f} Y$

~~$P(Y=1/X)$~~ $P(Y=1/x)$

s.f → $\beta_1 X + \beta_0$, $w^T X + b$ → it is not going to work

~~$\log\left(\frac{P(Y=1/X)}{1-P(Y=1/X)}\right)$~~ $X \xrightarrow{f} \log\frac{P(Y=1/X)}{1-P(Y=1/X)}$

⇒ our estimate is that, this is linear

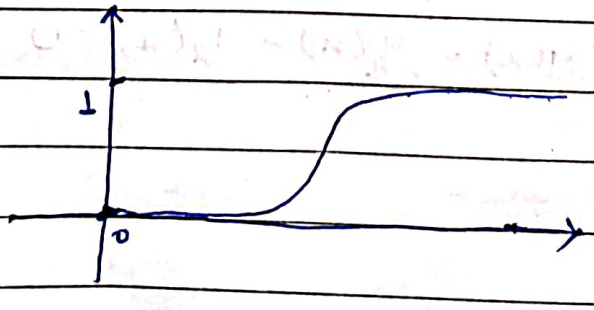
$\log\left(\frac{P(x)}{1-P(x)}\right) = \beta^T X$, $X = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$, $\beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix}$

⇒ it is a linear fn. of (x)

$\beta^T X + \beta_0 = \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \beta_0$

$\frac{P(x)}{1-P(x)} = e^{\beta^T X}$

$P(X) = \left(\frac{e^{\beta^T X}}{1+e^{\beta^T X}}\right) = \frac{1}{1+e^{-\beta^T X}}$ → estimated probability.



Date
Page No.

$$\sum_{i=1}^{n} \left(P(X_i) - \left(\frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}} \right) \right)^2$$

$P(X_i)$ → actual, $Y_i \sim f(X_i)$

$\left(\frac{e^{\beta^T X_i}}{1+e^{\beta^T X_i}} \right)$ → estimated

If (Y/X) follows Normal distribution, then least square loss function is optimal.

$$(X_i, Y_i) \in R^n \times \{0,1\}$$

$$\{(X_1, Y_1), (X_2, Y_2), \ldots (X_n, Y_n)\}$$

$$\mathcal{L}(\beta) = \prod_{i=y_i=1} P(X_i) \prod_{i=y_i=0} 1 - P(X_i)$$

$\rightarrow P(Y_i=1/X_i)$

[$\prod \rightarrow$ product, $\sum \rightarrow$ summation]

for a given value of X, Y can have 2 values, either '0' or '1', so it follows bernoulli distribution. {Bcoz most of the data pts. are around '0' & '1'.}

$$\mathcal{L}(\beta) = \prod_{i=1}^{n} P(X_i)^{y_i} (1 - P(X_i))^{1-y_i}$$

$$\text{Max } \log \mathcal{L}(\beta) = \sum_{i=1}^{n} y_i \log P(X_i) + (1 - y_i) \log (1 - P(X_i))$$

$$= \sum_{i=1}^{n} y_i \log \left(\frac{P(X_i)}{1 - P(X_i)} \right) + \log (1 - P(X_i))$$

$$= \sum_{i=1}^{n} y_i (\beta^T X_i) - \log (1 + e^{\beta^T X_i})$$

$$\min_{\beta} \; - \sum_{i=1}^{n} y_i (\beta^T X_i) - \log (1 + e^{\beta^T X_i})$$

Date
Page No.

$$\nabla_\beta f = -\sum_{i=1}^{n} \left(x_i y_i - \left(\frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} \right) x_i \right) = 0$$

$$= -\sum_{i=1}^{n} \left(y_i - \frac{e^{\beta^T x_i}}{1+e^{\beta^T x_i}} \right)$$

Choose β^0

Do until

$$\| \nabla f(\beta) \| \leq \epsilon$$

$$\beta^{t+1} = \beta^t + \eta_i \sum_{i=1}^{n} \left(y_i - \frac{e^{x_i^T \beta^t}}{1+e^{x_i^T \beta^t}} \right) x_i$$

end

$\eta_i = 0.01, 0.001$

$\eta_i = \frac{1}{1+i}$

$\eta_i = \frac{1}{i^2}$

- $\hat{p}(x) = \frac{e^{\beta^T x}}{1+e^{\beta^T x}}$

$$f(x) = \begin{cases} 0 & \hat{p}(x) \leq 0.5 \\ 1 & \text{otherwise} \end{cases}$$

$$\text{logit } \hat{p}(x) = \beta^T x = \log \left(\frac{\hat{p}(x)}{1-\hat{p}(x)} \right)$$

$$f(x) = \begin{cases} 0 & \text{if } \beta^T x \leq 0 \\ 1 & \text{otherwise} \end{cases}$$

$$\beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \beta_0 \leq 0$$

$$\beta^T X = 0$$

$$P(Y=1/X) = \frac{e^{\beta_1^T X}}{1 + \sum_{j=1}^{K-1} e^{\beta_j^T X}}$$

$$P(Y=2/X) = \frac{e^{\beta_2^T X}}{1 + \sum_{j=1}^{K} e^{\beta_j^T X}}$$

$$P(Y=K/X) = \frac{e^{\beta_K^T X}}{1 + \sum_{j=1}^{K} e^{\beta_j^T X}}$$

Weighted Regression technique

Min $f_0(x)$

subject to $f_i(x) \leq 0 \quad , i = 1, 2, \ldots m$

$h_i(x) = a_i^T x - b_i = 0 \quad , i = 1, 2, \ldots p$

Min $3x_1^2 + 4x_2^2$

such that

$4x_1 + 5x_2 \leq 5$

$6x_1 + 6x_2 = 4$

$x_1 \geq 0 \quad , x_2 \geq 0$

$$p^* = \text{Inf} \{ f_0(x) / x \in D \}$$

where, $D = \{ x / f_i(x) < 0 \quad i = 1, 2, \ldots m \}$

Date
Page No.

$h_i(x) = 0 \quad , i = 1, 2, \ldots$

$$f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i$$

For $x \in D$, $\lambda \geq 0$

$$L(x, \lambda, \nu) \leq f_0(x)$$

$$\inf L(x, \lambda, \nu) \leq \inf_{x \in D} L(x, \lambda, \nu)$$

$$\leq \inf_{x \in D} f_0(x) = p^*$$

$$g(\lambda, \nu) = \inf_{x \in R^n} f_0(x) + \sum_{i=1}^{m} \lambda_i f_i(x) + \sum_{i=1}^{p} \nu_i h_i(x)$$

Subject to $\lambda \geq 0$

$$g(\lambda, \nu) \leq p^*$$

Date
Page No.

→ Binary classification through S.V.M. ⇒

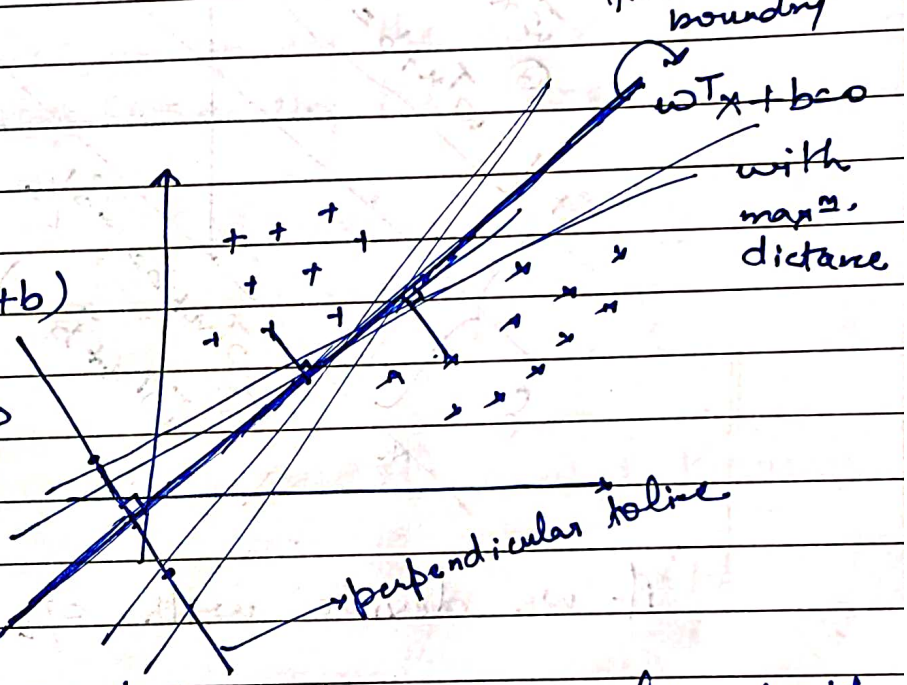
$$T = \{(x_1, y_1), (x_2, y_2) \ldots\ldots (x_L, y_L)\} ; x_i \in R^n$$

$$y_i \in \{-1, 1\}$$

$$i = 1, 2, \ldots L.$$

$$\text{Min } \frac{1}{2}\|w\|_p + C \sum_{i=1}^{L} L(y_i, x_i, w, b)$$

$\frac{1}{2}\|w\|_p$ — we are mostly minimizing → $\frac{1}{2}\|w\|^2$



$$f(w) = \text{Sign}(w^T x + b)$$

But there are ∞ possibility of lines in this region. which one should be take?

→ whose dictance is maximum from both the classes.

⇒ Take projection of pts. on the perpendicular line

$$w^T x + b$$

$$x_+ = \text{Min arg}\,(w^T x_+ + b)\ ,\quad x_+ \in I$$

So, we need to find a dir.^{n} in which → margin.

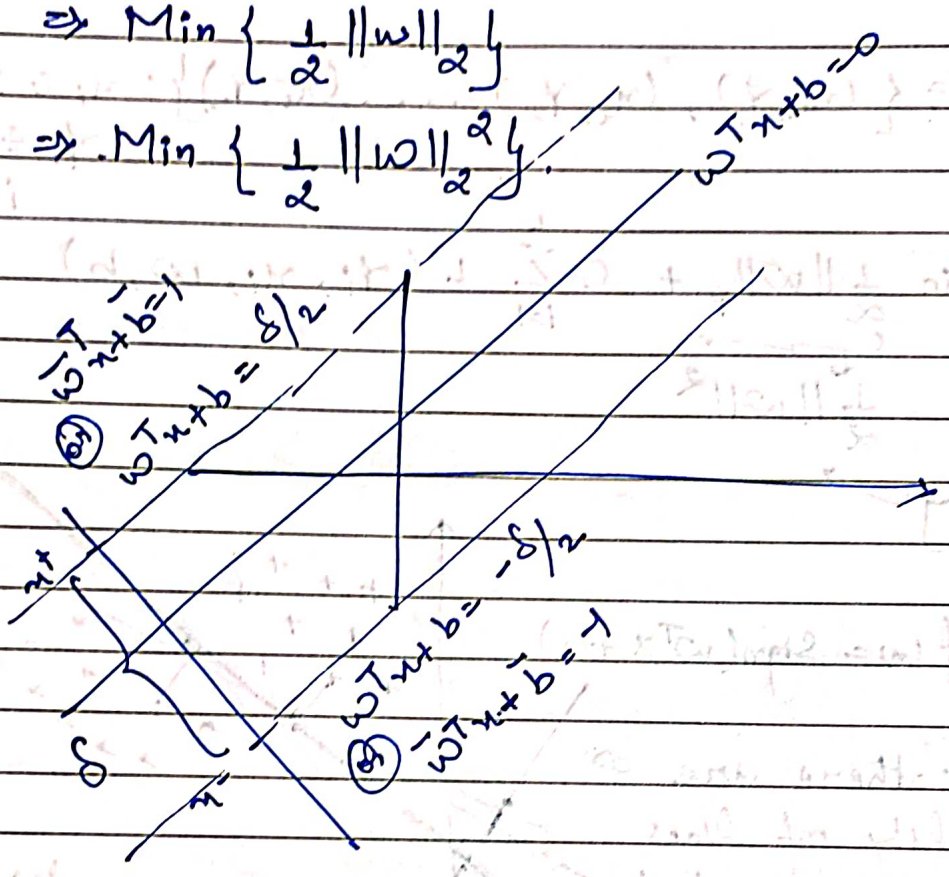
$$\underset{w,b}{\text{Max.}} \quad \frac{(w^T x_+ + b) - (w^T x^- + b)}{\|w\|_2} \geq \frac{1}{\|w\|_2}$$

margin.

$$\Rightarrow \text{Max}\left\{\frac{\delta}{\|w\|_2}\right\} = \text{Max}\left\{\frac{1}{\|w\|_2}\right\}$$

$$\Rightarrow \text{Min}\left\{\frac{1}{2}\|w\|_2\right\}$$

$$\Rightarrow \text{Min}\left\{\frac{1}{2}\|w\|_2^2\right\}$$

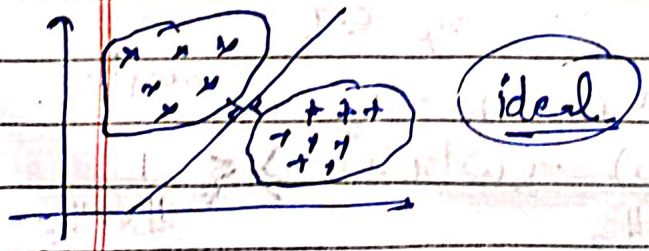


if we deviide with $\delta/2$ $\Rightarrow$

$$\frac{w^T x}{\delta/2} + \frac{b}{\delta/2} = 1 \Rightarrow \bar{w}^T x + \bar{b} = 1$$

sim" $\Rightarrow \bar{w}^T x + \bar{b} = -1$

Que $\rightarrow$ what happens when data is not linearly separable?



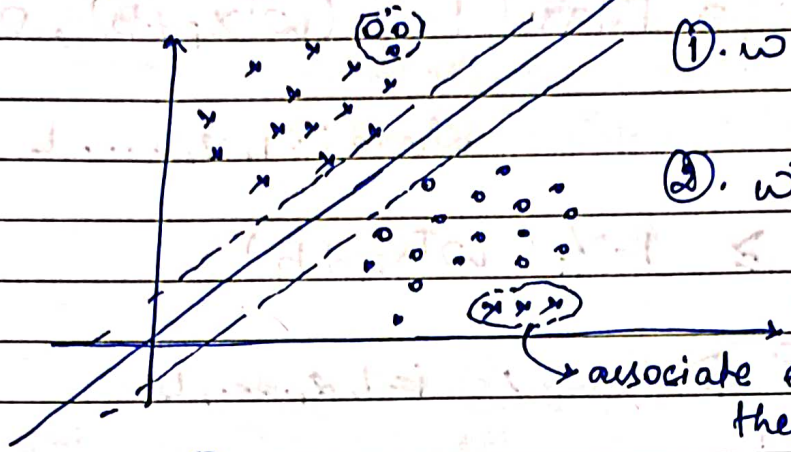
In general → $y_i(w^T x_i + b) \geq 1$

for error to occur ↙

Date
Page No.

Real word data →

error case →



① . $w^T x_i + b \geq -1$;
$x_i \in$ class ②
if $y_i = -1$;

② . $w^T x_i + b \leq 1$;
$x_i \in$ class ①.
if $y_i = 1$

associate error with them

final optimization problem

① → Try minimising the error.
② → Try maximising the margin

$$\Rightarrow \underset{w,b}{\text{Min}} \; \frac{1}{2} w^T w + C \cdot \sum_{i=1}^{n} \xi_i \quad \xi_i = \text{Max}(1 - y_i(w^T x_i + b), 0)$$

So, for error →

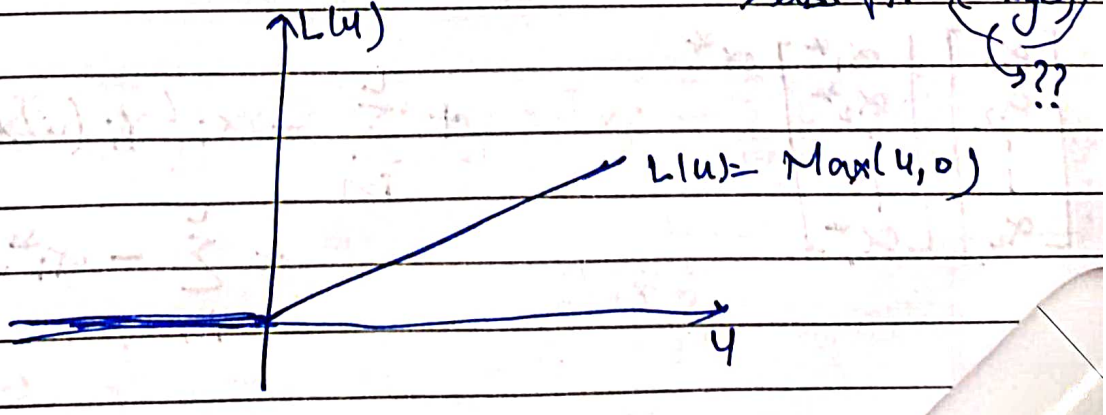
$$1 - y_i(w^T x_i + b) \geq 0 \rightarrow \text{error.}$$

$u \quad < 0 \rightarrow$ No error.

$$\Rightarrow \underset{w,b}{\text{Min}} \; \frac{1}{2} \|w\|^2 + C \cdot \sum_{i=1}^{n} L \cdot (1 - y_i(w^T x_i + b))$$

↳ loss fn. (Hinge)

??



Date
Page No.

$$\xi_i = \text{Max}\left(1 - y_i(w^T x_i + b), 0\right)$$

where $i = 1, 2, \ldots, L$

$$\xi_i \geq 1 - y_i(w^T x_i + b)$$

$$\xi_i \geq 0 \quad ; \quad i = 1, 2, \ldots, L$$

$$\Rightarrow \underset{w, b, \xi_i}{\text{Min}} \ \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{L} \xi_i$$

(CPP) → (convex programming problem).

Final optimization problem is reduced to →

$$\underset{w, b, \xi_i}{\text{Min}} \ \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{L} \xi_i$$

s.t →

$$y_i(w^T x_i + b) \geq 1 - \xi_i,$$

$$\xi_i \geq 0 \ ; \ i = 1, 2, \ldots, L.$$

Lagrangian fn. →

$$L(w, b, \xi_i, \alpha, \alpha^*) = \frac{1}{2}\|w\|^2 + C \cdot \sum_{i=1}^{L} \xi_i$$

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \\ \vdots \\ \alpha_L \end{bmatrix} \quad \begin{bmatrix} \alpha_1^* \\ \alpha_2^* \\ \vdots \\ \alpha_L^* \end{bmatrix} \quad \alpha^*$$

$$+ \sum_{i=1}^{L} -\alpha_i \left(y_i(w^T x_i + b) - 1 + \xi_i \right)$$

$$+ \sum_{i=1}^{L} -\alpha_i^* \cdot \xi_i$$

So →

$$L(w, b, \xi_i, \alpha, \alpha^*) = \frac{1}{2} w^T w + C \cdot \sum_{i=1}^{L} \xi_i - \sum_{i=1}^{L} \alpha_i \cdot (y_i (w^T x_i + b) - 1 + \xi_i) - \sum_{i=1}^{L} \xi_i \cdot \alpha_i^*$$

$$\frac{\partial L}{\partial w} = 0 \Rightarrow w = \sum_{i=1}^{L} \alpha_i \cdot y_i \cdot x_i$$

$$\frac{\partial L}{\partial b} = 0 \Rightarrow \sum_{i=1}^{L} \alpha_i \cdot y_i = 0$$

$$\frac{\partial L}{\partial \xi_i} = 0 \Rightarrow C - \alpha_i - \alpha_i^* = 0 \quad ; \text{ for } i = 1, 2, \ldots L.$$

$$\alpha_i (y_i (w^T x_i + b) - 1 + \xi_i) = 0 ;$$

$$\alpha_i^* \xi_i = 0 \quad ; \quad i = 1, 2, 3, \ldots L.$$

$$y_i (w^T x_i + b) \geq 1 - \xi_i \quad ; \quad i = 1, 2, \ldots, L$$

$$\xi_i \geq 0 \quad ; \quad i = 1, 2, \ldots L.$$

$$\Rightarrow \text{Max} \sum_{i=1}^{L} \sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j x_i^T x_j - \sum_{i=1}^{L} \alpha_i y_i \left(\sum_{j=1}^{L} \alpha_j y_j x_j \right)^T x_i + \sum_{i=1}^{L} \alpha_i$$

$$- \sum_{i=1}^{L} \sum_{j=1}^{L} \ldots \ldots -$$

Subject to → $\sum_{i=1}^{L} \alpha_i y_i = 0$

$C - \alpha_i - \alpha_i^* \quad ; \quad i = 1, 2, \ldots, L \; ; \; \alpha_i \leq C$

$\alpha_i = C - \alpha_i^*$

Date
Page No.

Min

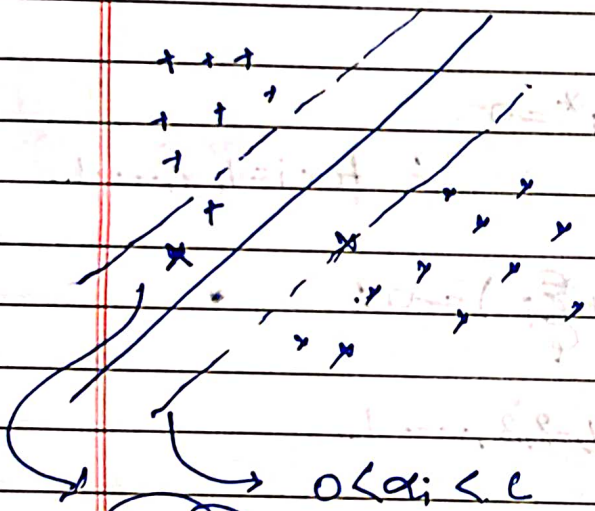
$$\Rightarrow \underset{(\alpha)}{\text{Max}} \; +\frac{1}{2}\sum_{i=1}^{L}\sum_{j=1}^{L} \alpha_i \alpha_j y_i y_j (x_i^T x_j) + \sum_{i=1}^{L} \alpha_i$$

Subject to →

$$\sum_{i=1}^{L} \alpha_i y_i = 0$$

$$0 \le \alpha_i \le C \; ; \; i = 1, 2, \ldots, L.$$

Geometrical Representation →



$$w = \sum_{i=1}^{L} \alpha_i y_i x_i$$

$0 < \alpha_i < C$

$\alpha_i = C$

→ they are contributing in w.

If $\alpha_i \in (0, C) \rightarrow \alpha_i^* > 0$

then $\xi_i = 0$

then $y_i * (w^T x_i + b) - 1 = 0$

$$y_i b = 1 - y_i (w^T x_i)$$

$$b = \frac{1}{y_i} - w^T x_i \quad \text{if } 0 < \alpha_i < C$$

$$b = y_i - w^T x_i$$

Date
Page No.

(or).

$$\left\{ b = y_i - \sum_{j=1}^{L} \alpha_j \cdot y_j \cdot x_j^T \cdot x_i \right\}.$$

final value of b is mean of all values of b.

for non-linearity ⟶

(changes).

$$x_i \longrightarrow \phi(x_i)$$

$$x_i^T x_j \longrightarrow \phi(x_i)^T \cdot \phi(x_j)$$

$$w \longrightarrow \sum_{i=1}^{L} \alpha_i y_i \phi(x_i)$$

$w^T x =$ —

$x_i^T x_j =$ Kernel $\Rightarrow K(x_i, x_j)$