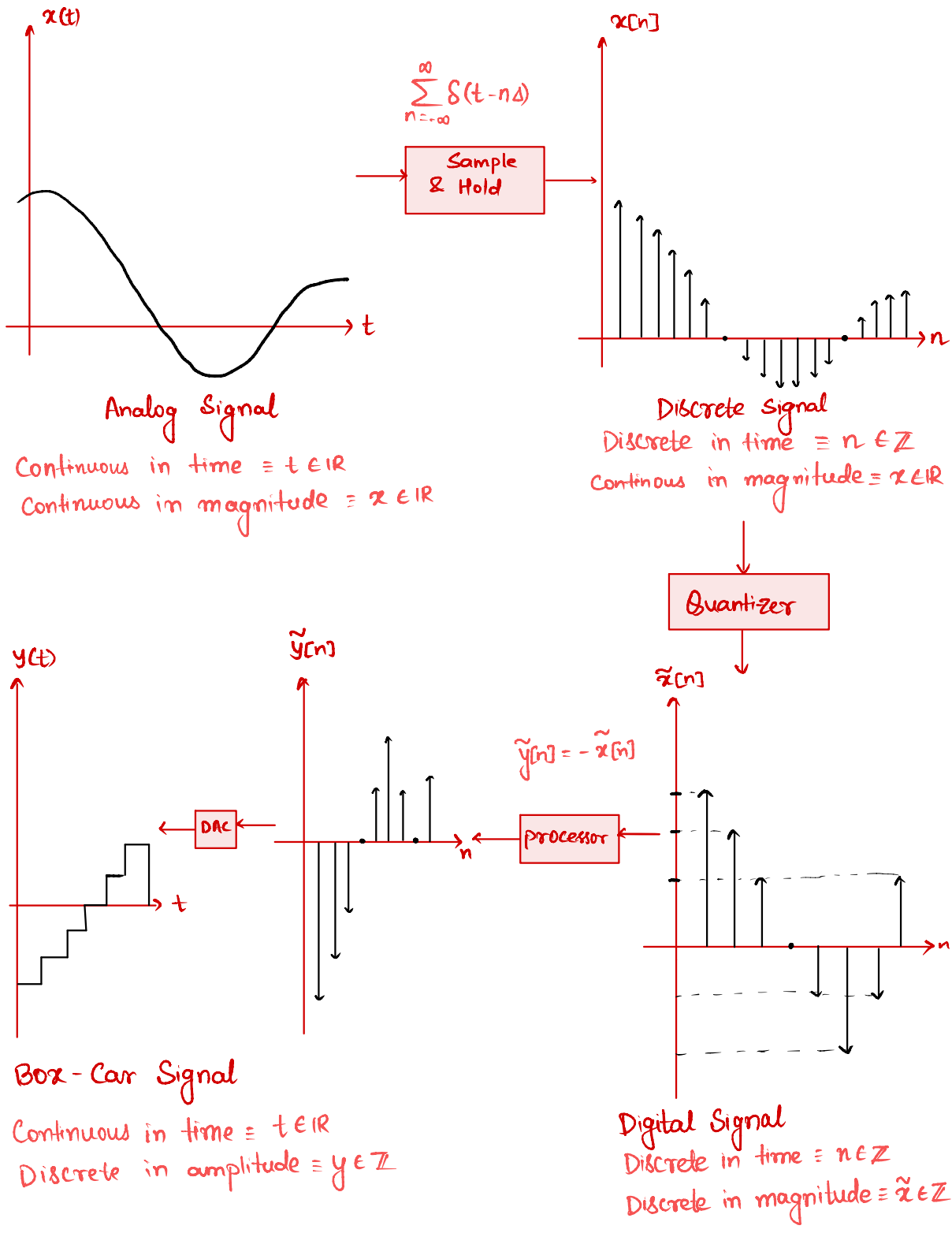

x(t)
t
Analog Signal
Continuous in time ≡ t ∈ IR
Continuous in magnitude ≡ x ∈ IR
$\sum_{n=-\infty}^{\infty} \delta(t-n\Delta)$
Sample & Hold
x[n]
n
Discrete Signal
Discrete in time ≡ n ∈ ℤ
Continous in magnitude ≡ x ∈ IR
Quantizer
$\tilde{x}[n]$
$\tilde{y}[n] = -\tilde{x}[n]$
processor
$\tilde{y}[n]$
n
DAC
y(t)
t
Box-Car Signal
Continuous in time ≡ t ∈ IR
Discrete in amplitude ≡ y ∈ ℤ
Digital Signal
Discrete in time ≡ n ∈ ℤ
Discrete in magnitude ≡ $\tilde{x}$ ∈ ℤ

# Representation Formats:

6 sensors

| Analog | | Digital |
|---|---|---|
| $x^{(0)}(t)$ | | $x^{(0)}[n]$ |
| $x^{(1)}(t)$ | | $x^{(1)}[n]$ |
| $x^{(2)}(t)$ | $\longleftrightarrow$ | $x^{(2)}[n]$ |
| $\vdots$ | | $\vdots$ |
| $x^{(5)}(t)$ | | $x^{(5)}[n]$ |

Sampling Rate = 10 sps = 10 Hz

No. of bits/sample (bps) = 16 bits

Data size in 1s = No. of Sensors × Sampling Rate × Time × bps

$= 6 \times 10 \text{ samples } s^{-1} \times 1\,s \times 16 \text{ bits sample}^{-1}$

$= 960$ bits

$= 120$ bytes

Data in 1 year = $365 \text{ days} \times 24 \text{ hrs day}^{-1} \times 3600\,s\,hr^{-1} \times 120 \text{ bytes } s^{-1}$

$\approx 3.8$ GB

Data per year in Nalanda Complex = $(3.8 \times 96 \times 5)$ GB

$\approx$ 1.8 TB

Sensors

- $CO_x$
- $NO_x$
- $SO_x$
- Temperature
- Humidity

No of cables = $(6 \times 96)$

$\approx 580$

If $K \ll D$,

$K$ → 2, $D$ → 580

## Data Compression :

$\underline{x}[n] = [\, x^{(0)}[n] \,,\, x^{(1)}[n] \,,\, \cdots \,,\, x^{(d)}[n] \,,\, \cdots \,,\, x^{(D-1)}[n] \,]^T \in \mathbb{R}^D$

↳ Vector valued sample at $n^{th}$ instance

$\underline{X} = [\, \underline{x}[0] \,,\, \underline{x}[1] \,,\, \underline{x}[2] \,,\, \cdots \underline{x}[n] \,,\, \cdots \,,\, \underline{x}[N-1] \,]$

$= [\, \underline{x}_0 \,,\, \underline{x}_1 \,,\, \cdots \,,\, \underline{x}_n \,,\, \cdots \,,\, \underline{x}_{N-1} \,]$

$$= \begin{bmatrix} \begin{bmatrix} x^{(0)}[0] \\ x^{(1)}[0] \\ \vdots \\ x^{(d)}[0] \\ \vdots \\ x^{(D-1)}[0] \end{bmatrix} & \begin{bmatrix} x^{(0)}[1] \\ x^{(1)}[1] \\ \vdots \\ x^{(d)}[1] \\ \vdots \\ x^{(D-1)}[1] \end{bmatrix} & \cdots & \begin{bmatrix} x^{(0)}[N-1] \\ x^{(1)}[N-1] \\ \vdots \\ x^{(d)}[N-1] \\ \vdots \\ x^{(D-1)}[N-1] \end{bmatrix} \end{bmatrix}$$

$$= \begin{bmatrix} x_0^{(0)} & x_1^{(0)} & \cdots & x_n^{(0)} & \cdots & x_{N-1}^{(0)} \\ x_0^{(1)} & x_1^{(1)} & & & & \\ x_0^{(2)} & x_1^{(2)} & & & & \\ \vdots & \vdots & & \vdots & & \\ x_0^{(d)} & x_1^{(d)} & & x_n^{(d)} & & \\ \vdots & \vdots & & \vdots & & \\ x_0^{(D-1)} & x_1^{(D-1)} & \cdots & x_n^{(D-1)} & \cdots & x_{N-1}^{(D-1)} \end{bmatrix}$$

$\in \mathbb{R}^{D \times N}$

Number of samp

Dimensions of data samples

$y = A^T x \rightarrow \mathbb{R}^{D \times 1}$

$\mathbb{R}^{K \times D}$

$\in \mathbb{R}^{K \times 1}$

Q] Prove $y = A^T x$ is linear.

$y_1 = A^T x_1$

$y_2 = A^T x_2$

$y = y_1 + y_2$

$= A^T x_1 + A^T x_2$

$y = A^T x$

When $x$ is replaced with $kx$ ($k \in \mathbb{R}$)

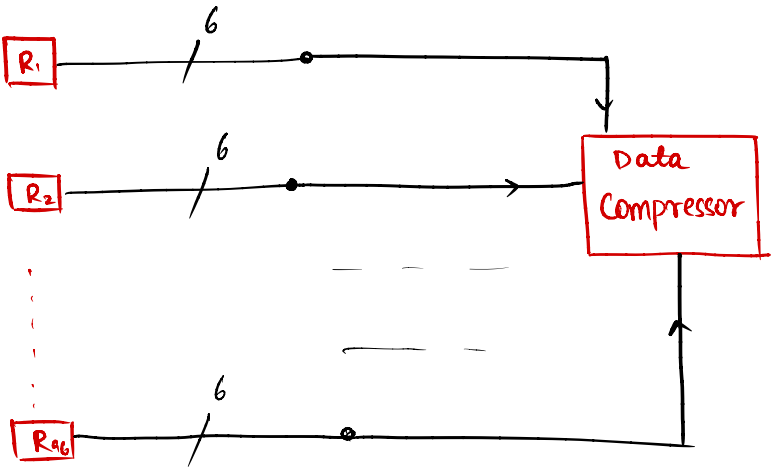
$y' = A^T (kx)$

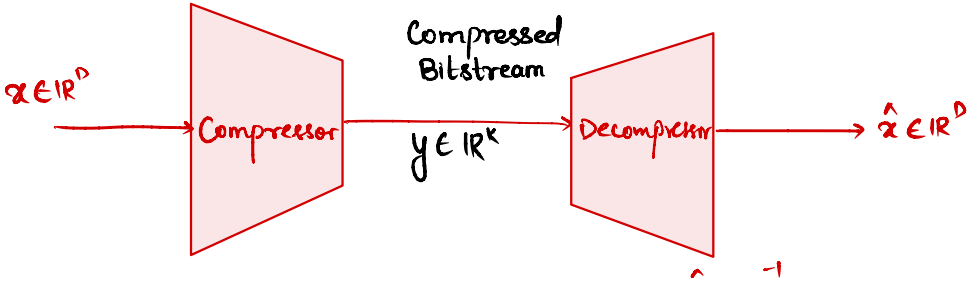
$= k A^T x$

$= ky$

$= A^T (x_1 + x_2)$

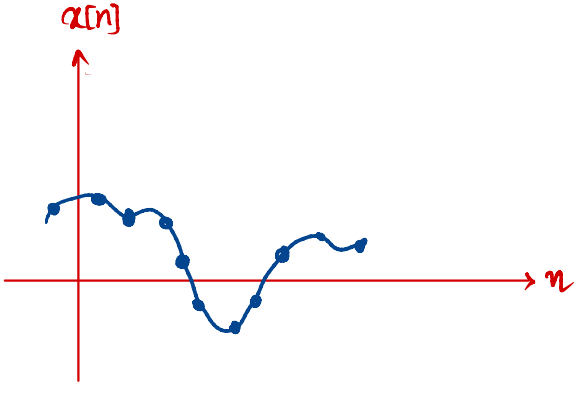
$= A^T x$



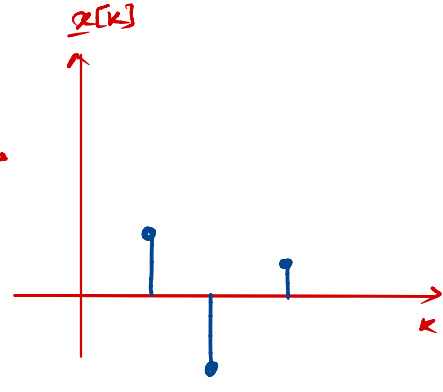




$\underline{y} = \underline{A}^T \underline{x}$

$\underline{\hat{x}} = \underline{A}^{-1} \underline{y}$




$\mathcal{F}$



$\underline{x} = [x[-2], x[-1], x[0], x[1], \cdots, x[N-3]] \in \mathbb{R}^N$

$\downarrow \mathcal{T}$

$\underline{y} = [y[-2], y[-1], y[0], y[1], \cdots, y[N-3]] \in \mathbb{R}^N$

$y[-2] = 0$, $y[-1] \neq 0$, $y[0] = 0$, $y[1] = 0$, $y[N-3] = 0$

$\underline{y} = \mathcal{T}(\underline{x})$ $\quad (K \ll N)$

$\underline{y} \in \mathbb{R}^K$, $\underline{x} \in \mathbb{R}^N$

$\underline{y} = \underline{A}^T \underline{x}$

↳ DFT coefficent matrix

(Analytically computed in case of FFT, DCT, DWT)

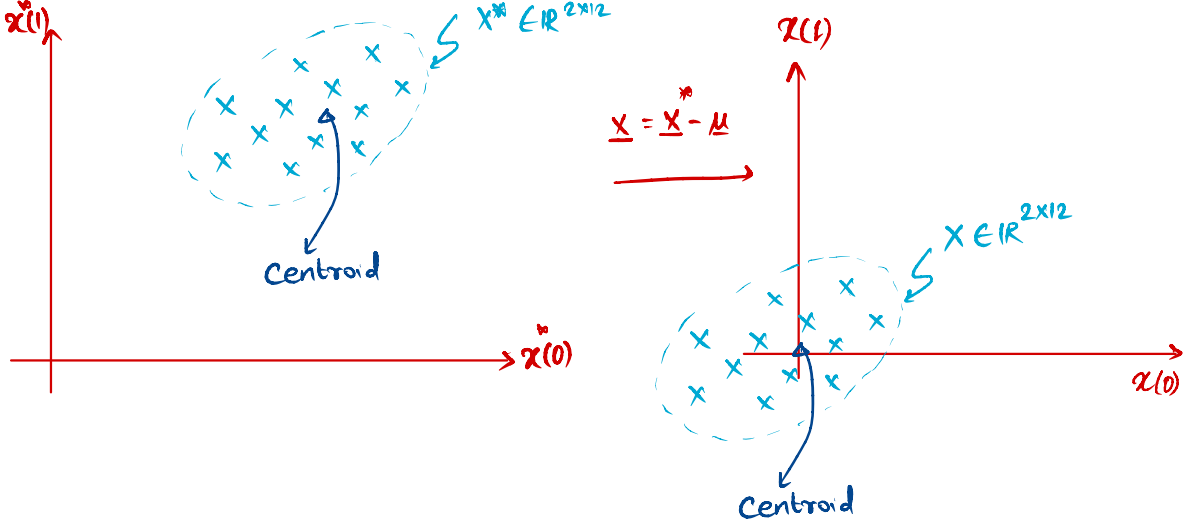
Q] Can we obtain $A^T$ from $\underline{X}$ s.t. $\underline{Y} = \underline{A}^T \underline{X}$, $Y \in \mathbb{R}^{K \times N}$, $X \in \mathbb{R}^{D \times N}$, $K \ll D$.

$D = 2$

$\underline{x}_n = [x_n^{(0)} \; x_n^{(1)}]^T \in \mathbb{R}^2$

$\underline{y}_n = [y_n^{(0)}]^T \in \mathbb{R}$

$X = [\underline{x}_0 \; \underline{x}_1 \; \cdots \; \underline{x}_{N-1}] \in \mathbb{R}^{2 \times N} \longrightarrow \underline{Y} = [\underline{y}_0 \; \underline{y}_1 \; \underline{y}_2, \cdots, \underline{y}_{N-1}] \in \mathbb{R}^{1 \times N}$




$$\underline{X} = \underline{X}^* - \underline{\mu}$$

$$\underline{\mu} = [\mu^{(0)} \quad \mu^{(1)} \quad \cdots \quad \mu^{d} \quad , \quad \cdots \quad , \quad \mu^{(D-1)}]$$

$$\text{s.t } \mu^{(d)} = \frac{1}{N} \sum_{n=0}^{N-1} x_n{}^{(d)}$$

$$\underline{a}^{(0)} = [a_{00}, a_{10}] \in \mathbb{R}^{2\times 1}$$

$$\underline{a}^{(1)} = [a_{10}, a_{11}] \in \mathbb{R}^{2\times 1}$$

$$y_n{}^{(0)} = \underline{a}^{(0)T} \underline{x_n}$$

$$y_n{}^{(1)} = \underline{a}^{(1)T} \underline{x_n}$$

$$\underline{A}^T = \begin{bmatrix} a^{(0)T} \\ a^{(1)T} \end{bmatrix} = \begin{bmatrix} a_{00} & a_{10} \\ a_{01} & a_{11} \end{bmatrix}$$

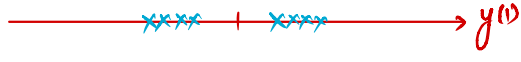
$$\underline{y}_n = \underline{A}^T \underline{x}_n$$

$$\underline{A} = \begin{bmatrix} a_{00} & a_{10} \\ a_{01} & a_{11} \end{bmatrix} = [\underline{a}^{(0)} \quad \underline{a}^{(1)}] \in \mathbb{R}^{2\times 2}$$

$\underline{a}^{(0)} \rightarrow \lambda^{(0)}$, $\underline{a}^{(1)} \rightarrow \lambda^{(1)}$

$\lambda^{(0)}, \lambda^{(1)}$ : Eigenvalues


Higher $\lambda$




Lower $\lambda$

$\underline{x}_n \in \mathbb{R}^D \qquad \underline{\tilde{y}}_n \in \mathbb{R}^k \; ; \; k \ll D$

$\underline{y}_n = \underline{A}^T \underline{x}_n$

$\underline{y}_n \in \mathbb{R}^D$, $\underline{A}^T \in \mathbb{R}^{D\times D}$, $\underline{x}_n \in \mathbb{R}^D$

$$\begin{bmatrix} y^{(0)} \\ y^{(1)} \\ \vdots \\ y^{(d)} \\ \vdots \\ y^{(D-1)} \end{bmatrix} = \begin{bmatrix} a_{00} & a_{10} & \cdots & a_{D-1\,0} \\ a_{01} & a_{11} & & a_{D-1\,1} \\ \vdots & \vdots & & \vdots \\ a_{0d} & a_{1d} & & a_{D-1\,d} \\ \vdots & \vdots & & \vdots \\ a_{0\,D-1} & a_{1\,D-1} & & a_{D-1\,D-1} \end{bmatrix} \begin{bmatrix} x^{(0)} \\ x^{(1)} \\ \vdots \\ x^{(d)} \\ \vdots \\ x^{(D-1)} \end{bmatrix}$$

$$\tilde{y}_n = [\, y^{(0)}, y^{(1)}, \cdots, y^{(d)}, \cdots y^{(D-1)} \,]$$

$$\hat{x}_n = (A^{nT})^{-1} \tilde{y}_n$$

$$\| x_n - \hat{x}_n \| = \epsilon_n$$

$$\epsilon_n^2 = \sum_{j=k}^{N-1} \lambda_j^2$$

## Singular Value Decomposition (SVD):-

PCA: $x_i \rightarrow \tilde{y}_i$

$\tilde{y}_i = \tilde{A}^T x_i$

$x_i \in \mathbb{R}^{D \times 1}$, $\tilde{y}_i \in \mathbb{R}^{K \times 1}$

$X \rightarrow \tilde{Y}$

$\tilde{Y} \rightarrow A^T X$

$X \in \mathbb{R}^{D \times N}$, $Y \in \mathbb{R}^{K \times N}$

$$X = U \begin{bmatrix} \Lambda^{1/2} & O_a \\ O_b & O_c \end{bmatrix} V^H$$

$U \in \mathbb{R}^{D \times D}$, $\Lambda^{1/2} \in \mathbb{R}^{D \times D}$, $V^H \in \mathbb{R}^{N \times N}$

$X \in \mathbb{R}^{D \times N}$

$\Lambda^{1/2} \in \mathbb{R}^{D \times D}$

$U \in \mathbb{R}^{D \times D}$

$V^H \in \mathbb{R}^{D \times N}$

$O_a \in \mathbb{R}^{D \times (N-D)}$

$O_b \in \mathbb{R}^{0 \times D}$

$O_c \in \mathbb{R}^{0 \times (N-D)}$

$$\Lambda^{1/2} = \begin{bmatrix} \sqrt{\lambda_0} & 0 & 0 & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_1} & 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & & & \\ 0 & 0 & 0 & \cdots & \cdots & \sqrt{\lambda_{D-1}} \end{bmatrix}$$

$(|\lambda_0| > |\lambda_1| > |\lambda_2| > \cdots > |\lambda_{D-1}|)$

- $\lambda^d$ is the Eigenvalue of $\underline{X}\,\underline{X}^T$

$\underline{X}^T \in \mathbb{R}^{N \times D}$

$\underline{X} \in \mathbb{R}^{D \times N}$

$\underline{X}\,\underline{X}^T \in \mathbb{R}^{D \times D}$

- $\underline{U}$ is the set of Eigenvectors of $\underline{X}\,\underline{X}^T$ $(\underline{X}\underline{X}^T \in \mathbb{R}^{N \times N})$

$$\underline{U} = [\underline{u}_0 \quad \underline{u}_1 \quad \underline{u}_2 \quad \dots \quad \underline{u}_d \quad \dots \quad \underline{u}_{D-1}]$$

$\underline{u}_d \in \mathbb{R}^{D \times 1}$

- $V^H$ is the set of Eigenvectors of $\underline{X}^T\underline{X}$ $(\underline{X}^T\underline{X} \in \mathbb{R}^{N \times N})$

$$\underline{V}^H = [\underline{v}_0 \quad \underline{v}_1 \quad \underline{v}_2 \quad \dots \quad \underline{v}_{n-1} \quad \dots \quad \underline{v}_{N-1}]$$

$\underline{v}_{n-1} \in \mathbb{R}^{N \times 1}$

P - bytes per number

Double - 8 bytes

$(K \ll D \,,\ K \ll N)$

sizeof $(\underline{X})$ = pDN

sizeof $(\underline{U})$ + sizeof $(\Lambda^{1/2})$ + sizeof $(\underline{V}^H)$

= p (DK + K + KN)

= p (DK + K + KN) < p(DN)

$\Rightarrow$ (KD + K + KN) < DN

$\underline{U} \in \mathbb{R}^{D \times D}$ $\quad \underline{\Lambda}_K^{1/2} \in \mathbb{R}^{K \times K}$ $\quad \underline{V} \in \mathbb{R}^{N \times N}$

$$\underline{\tilde{X}}_K = \left[\begin{array}{c|c} \underline{U}_K \in \mathbb{R}^{D \times K} & 0^{D \times (D-K)} \end{array}\right] \left[\begin{array}{c|c} \begin{bmatrix} \sqrt{\lambda_0} & 0 & \cdots & 0 \\ 0 & \sqrt{\lambda_1} & \cdots & 0 \\ \vdots & & & \\ 0 & 0 & \cdots & \sqrt{\lambda_{K-1}} \end{bmatrix} & 0^{K \times (N-K)} \\ \hline 0^{(D-K) \times K} & 0^{(D-K) \times (N-K)} \end{array}\right] \left[\begin{array}{c} \underline{V}_K^H \in \mathbb{R}^{K \times N} \\ \hline 0^{(N-K) \times N} \end{array}\right]$$

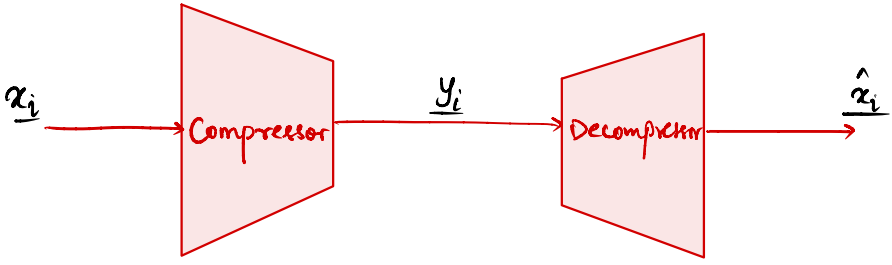
Find $K$ such that $\min \|\underline{X} - \underline{\tilde{X}}\|$ and $\max \left( \dfrac{\text{sizeof}(\underline{X})}{\text{sizeof}(\underline{U}_K, \underline{\Lambda}_K, \underline{V}_K)} \right)$

↓

Compression Factor (CF)

Data Transform

- Learnable Transform
- Fixed Transform




PCA: $\underline{y}_i = \tilde{\underline{A}}^T \underline{x}_i$

$\underline{y}_i \in \mathbb{R}^{K \times 1}$ $\quad$ $\underline{x}_i \in \mathbb{R}^{D \times 1}$

$\hat{\underline{x}}_i = \underline{A}\, \underline{y}_i$

$\hat{\underline{x}}_i \in \mathbb{R}^{D \times 1}$ $\quad$ $\underline{y}_i \in \mathbb{R}^{K \times 1}$

$(K < D)$

$$\text{Compression Factor (CF)} = \frac{\text{sizeof}(X)}{\text{sizeof}(Y)}$$

SVD: $\underline{U}_K \underline{\Lambda}_K^{1/2} \underline{V}_K^H \approx \underline{X}$

$\hat{\underline{X}} = \underline{U}_K \underline{\Lambda}_K^{1/2} \underline{V}_K^H$

$$\text{Compression Factor (CF)} = \frac{\text{sizeof}(X)}{\text{sizeof}(U) + \text{sizeof}(\Lambda_K^{1/2}) + \text{sizeof}(V_K^H)}$$

PDN → precision of number (bits per number)

double → precision floating point number (64 bits)

single → 32 bits

half → 16 bits

quarter → 8 bits

Compression Factor (CF) = $\frac{\text{sizeof}(X)}{\text{sizeof}(Y)} = \frac{PDN}{qDN} = \frac{P}{q}$

Dynamic Range of representing a number

$\underline{X} \in \mathbb{R}^{DXN} \longrightarrow Y \in \mathbb{R}^{DXN}$

## Information Theory :-

$x \in [0, 255] \cap \mathbb{R}$ ⟶ Infinite bits

$\in [0, 255] \cap \mathbb{Z}$ ⟶ 8 bits

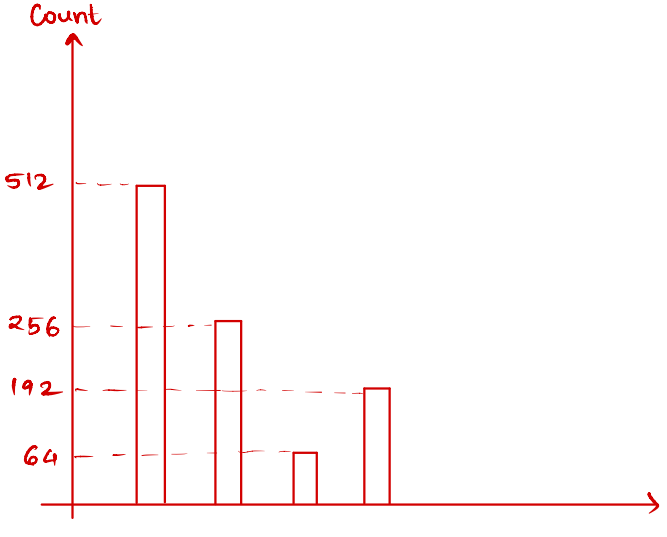
$2^0 - 1 = 1$

$2^1 - 1 = 3$

$\vdots$

$2^8 - 1 = 255$

$X = [x_0 \quad x_1 \quad x_2 \cdots x_{N-1}] \in \mathbb{R}^N$




For representation of this no of bits are required is ____.

| Symbol | Bitcodes |
|---|---|
| 10 | 00 |
| 100 | 01 |
| 190 | 10 |
| 250 | 11 |

$X \rightarrow Y$ (Encoder)

$Y \rightarrow \tilde{X}$ (Decoder)

$|X - \tilde{X}| = 0$

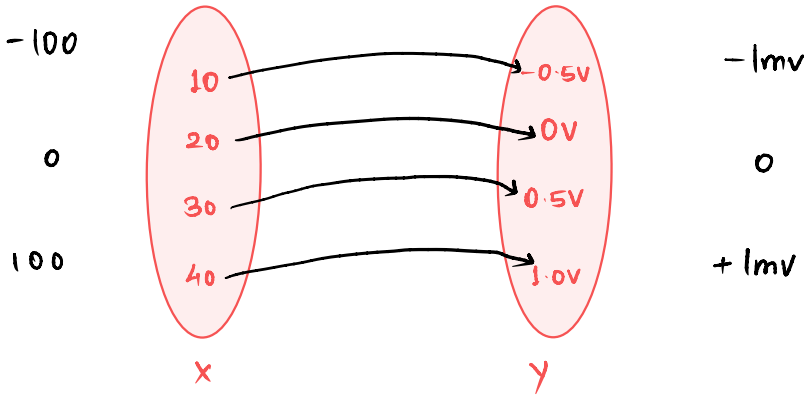
↳ So, loseless system

- In worst case, CF is equal to 1.

## Clande Elevood Shannon –

X —Channel→ Y

($T_x$) ($R_x$)

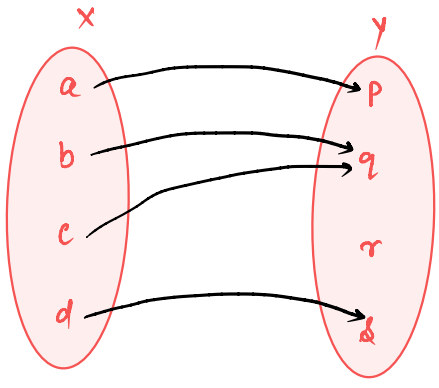



1-1 mapping

Atlantis Communication

↳ Thickness - 3m

Restrained Channel Capacity




$W = H + GS$

Entropy = $\frac{\text{Heat}}{\text{Temp}}$

of gen = $kT$

Boltzmann's constant

| | | | | |
|---|---|---|---|---|
| 10 | — | a | — 512 | $P(a) = 1/2$ |
| 100 | — | b | — 256 | $P(b) = 1/4$ |
| 190 | — | c | — 64 | $P(c) = 1/16$ |
| 250 | — | d | — 192 | $P(d) = 3/16$ |
| | | | $\Sigma = 1024$ | |

- Information carried by a signal,

$$I(a) = \log_r P(a)$$

$r$ = 2 bits

$r$ = e-nats

$$H(X) = E_{s \in S}(I(s)) = -\sum_{s \in S} p(s) \log_2 (p(s)) \geq 0$$

↓ Entropy

**Joint Entropy:** $H(X,Y) = -\sum_{s \in S, r \in T} p(s,r) \log_2 (p(s,r)) \geq 0$

| Y \ X | a | b | c | d | |
|---|---|---|---|---|---|
| p | 1/8 | 1/16 | 1/32 | 1/32 | 1/4 = P(p) |
| q | 1/16 | 1/8 | 1/32 | 1/32 | 1/4 = P(q) |
| r | 1/16 | 1/16 | 1/16 | 1/16 | 1/4 = P(r) |
| s | 1/4 | 0 | 0 | 0 | 1/4 = P(s) |
| | 1/2 = P(a) | 1/4 = P(b) | 1/8 = P(c) | 1/8 = P(d) | |

$$H(X) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{8}\log_2\frac{1}{8} + \frac{1}{8}\log_2\frac{1}{8}\right)$$

$$= 2.25 \text{ bits}$$

$$H(Y) = -\left(\frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4} + \frac{1}{4}\log_2\frac{1}{4}\right)$$

$$= 2 \text{ bits}$$

$$H(X,Y) = \frac{27}{8} \text{ bits}$$

$$H(X,Y) = H(X) + H(Y/X) = H(Y) + H(X/Y)$$

Relative Entropy → Kullback - Leiber Divergence

Divergence → Effective distance

$$D(p||q) = \sum_{x \in X} p(x) \log \frac{p(x)}{q(x)}$$

$$= E_p \log \frac{p(x)}{q(x)}$$

## Mutual Information -

$$E_{p(x,y)} \log \frac{p(x,y)}{p(x)\,p(y)} = H(x) - H(X/Y)$$

$$I(x;y) = D(p(x,y) \,||\, p(x) \cdot p(y))$$

$$I(x,x) = H(X)$$

- For symmetric channel = $H(Y) - H(Y/X)$

## Huffman Compression :-

Coconutcurry

$X = ['c', 'o', 'c', 'o', 'n', 'u', 't', 'c', 'u', 'r', 'r', 'y'] \in \mathbb{B}^{8 \times 12}$

| | Symbol | No of occurence | P(symbol) |
|---|---|---|---|
| 10 | C | 3 | 3/12 |
| 00 | O | 2 | 2/12 |
| 010 | n | 1 | 1/12 |
| 110 | u | 2 | 2/12 |

| | | | |
|---|---|---|---|
| 0110 | t | 1 | 1/12 |
| 111 | r | 2 | 2/12 |
| 0111 | y | 1 | 1/12 |
| | | Σ = 12 | |

↓ Reorder Descending

c - 3/12
o - 2/12
u - 2/12
r - 2/12
n - 1/12
t - 1/12
y - 1/12 } {t,y} - 2/12

→

c - 3/12
o - 2/12
u - 2/12
r - 2/12
{t,y} - 2/12 (n5)
n - 1/12 } {{t,y}, n} - 3/12

↓

c - 3/12
{{t,y}, n} - 3/12 (n4)
o - 2/12
u - 2/12
r - 2/12 } → {u,r} - 4/12

{u,r} - 4/12
c - 3/12
{{t,y}, n} - 3/12
o - 2/12 } {{{t,y},n}, o} - 8/12 ←

{{{t,y},n},o} - 5/12

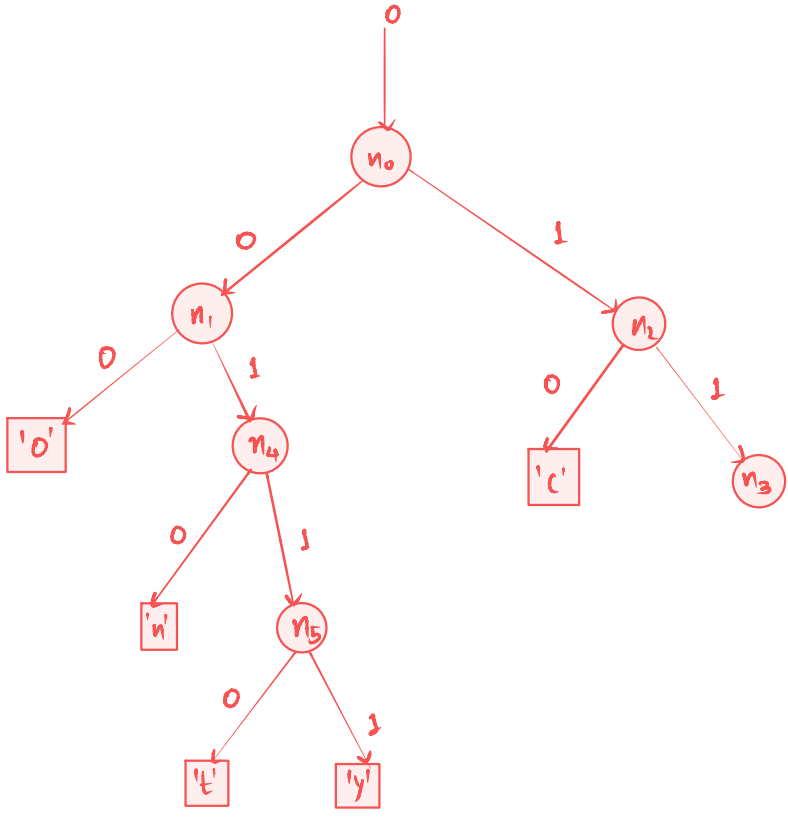
{u,r} — 4/12 $n_3$

c — 3/12

{{u,r},c} - 7/12

→ {{u,r},c} - 7/12 $n_2$

{{{t,y},n},o} - 5/12 $n_1$



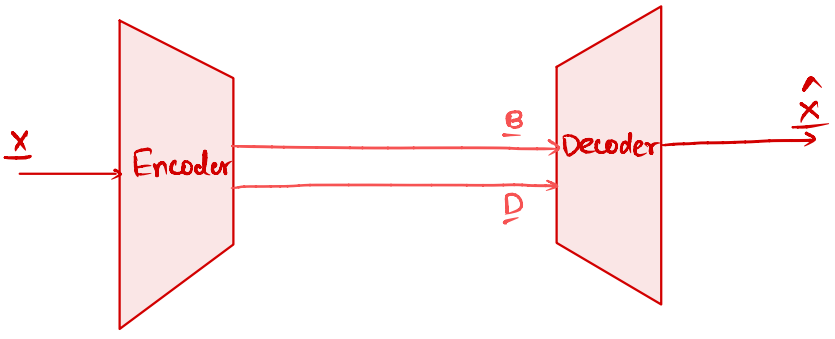

No of bits = 33

Avg no. of bits / symbol = $\frac{33}{12}$

= 2.75

$$H(X) = -\sum_{symb} p(symb) \log_2 p(symb)$$

$$= -\left(\frac{3}{12}\log_2\frac{3}{12} + \frac{6}{12}\log_2\frac{6}{12} + \frac{3}{12}\log\frac{1}{12}\right) = 2.69$$




| Header | Data |
|---|---|
| D | B |
| 9 - bits | 33 - bits |

MSE $(X, \hat{X}) = 0$ ⟶ Lossless

Compression Factor (CF) = $\frac{\text{No of bits in } X}{\text{No of bits in } B}$

$$= \frac{12 \times 8}{33}$$

$$\approx 2.9$$

Max CF = $\frac{\text{Avg. no of bits in } X}{H(X)}$

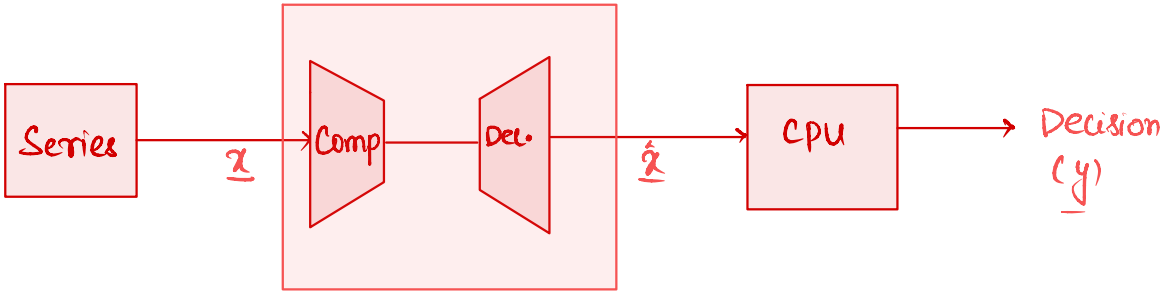
$$\text{Coding Efficiency} = \frac{H(x)}{\text{Avg. bits per symbol}}$$

$$= \frac{2.69}{2.75}$$

$$\approx 97.1$$

## Data Compression

- Lossy
  - PCA
  - SVD
- Lossless → Huffman




## Bayes Decision Theorem:-

$$P(y_j|\underline{x}) = \frac{P(\underline{x}|y_j) \cdot P(y_j)}{P(\underline{x})}$$

- $P(y_j|\underline{x})$ → posterior
- $P(\underline{x}|y_j)$ → Likelihood
- $P(y_j)$ → Prior
- $P(\underline{x})$ → Evidence

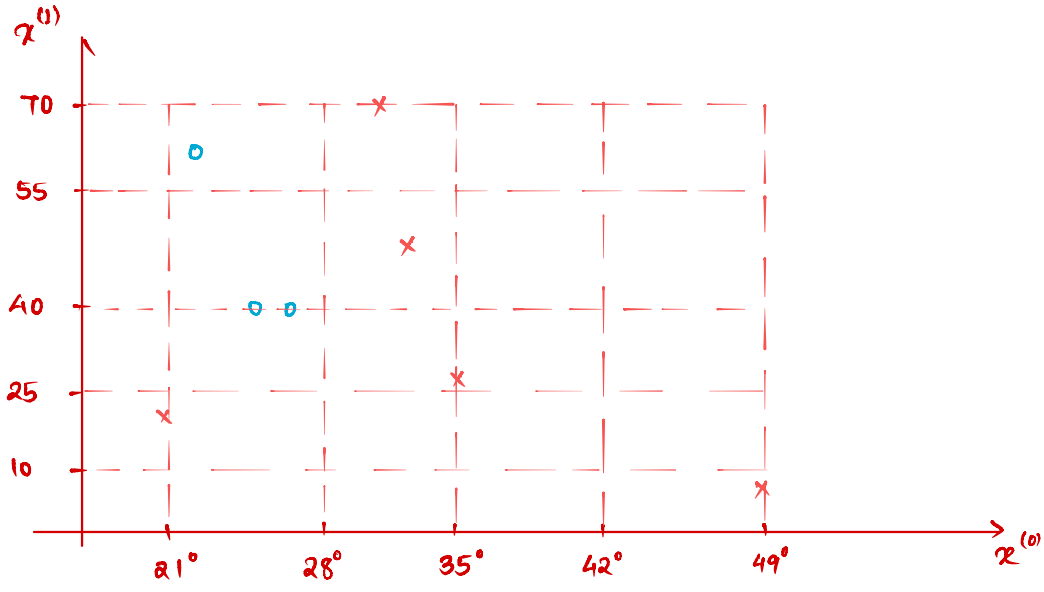
| | $x^{(0)}$ | $x^{(1)}$ | $y^{(0)}$ | $y^{(1)}$ | |
|---|---|---|---|---|---|
| | Temp | Humidity | Comfort | N. Comfort | |
| $\underline{x}_0^T$ | 31°C | 70 % rh | 0 | 1 | $\underline{y}_0$ |
| $\underline{x}_1^T$ | 24°C | 40 % rh | 1 | 0 | $\underline{y}_1$ |
| $\underline{x}_2^T$ | 22°C | 60 % rh | 1 | 0 | $\underline{y}_2$ |
| $\underline{x}_3^T$ | 49°C | 10 % rh | 0 | 1 | $\underline{y}_3$ |
| $\underline{x}_4^T$ | 35°C | 30 % rh | 0 | 1 | $\underline{y}_4$ |
| $\underline{x}_5^T$ | 32°C | 50 % rh | 0 | 1 | $\underline{y}_5$ |
| $\underline{x}_6^T$ | 27°C | 40 % rh | 1 | 0 | $\underline{y}_6$ |
| $\underline{x}_7^T$ | 21°C | 20 % rh | 0 | 1 | $\underline{y}_7$ |

$\underline{x}_i \in \mathbb{R}^{2\times 1}$

$\underline{X} = [\underline{x}_0, \underline{x}_1, \underline{x}_2, \dots, \underline{x}_7] \in \mathbb{R}^{2\times 8}$

$\underline{y}_i \in \mathbb{B}^{2\times 1}$

$\underline{Y} = [\underline{y}_0, \underline{y}_1, \dots, \underline{y}_7] \in \mathbb{B}^{2\times 8}$




$$P(y = \text{comfort} \mid \underline{x}) = \frac{P(\underline{x} \mid y = \text{comfort}) \cdot P(y = \text{comfort})}{P(\underline{x})}$$

$$P(y = \text{N. Comfort} \mid \underline{x}) = \frac{P(\underline{x} \mid y = \text{N. Comfort}) \cdot P(\text{N. Comfort})}{P(\underline{x})}$$

$P(\text{Comfort}) = \frac{3}{8} = (P(y = [1, 0])) = (P(y = \omega_0))$

$P(\text{N. Comfort}) = \frac{5}{8} = (P(y = [0, 1])) = (P(y = \omega_1))$

$\omega_1$: Class Label

$$h(\underline{x}) = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 2 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

$$h(\underline{x} \mid y = \omega_0) = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

$$h(\underline{x} \mid y = \omega_1) = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 1 \end{bmatrix}$$

$\underline{x} = [27, 48]$

$$P(\underline{x} \mid y = \omega_0) = \frac{h(\underline{x} \mid y = \omega_0)}{\sum_{\underline{x}} h(\underline{x} \mid y = \omega_0)} = \frac{2}{3}$$

$$P(\underline{x} \mid y = \omega_1) = \frac{h(\underline{x} \mid y = \omega_1)}{\sum_{\underline{x}} h(\underline{x} \mid y = \omega_1)} = \frac{0}{5} = 0$$

$$P(y = \omega_0) = \frac{3}{8}$$

$$P(y = \omega_1) = \frac{5}{8}$$

$$P(\underline{x}) = P(\underline{x} \mid y=\omega_0) \cdot P(y=\omega_0) + P(\underline{x} \mid y=\omega_1) \cdot P(y=\omega_1)$$

$$= \left(\frac{2}{3}\right)\left(\frac{3}{8}\right) + 0 \cdot \frac{5}{8}$$

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

$$P(y=\omega_0 \mid \underline{x}) = \frac{P(\underline{x} \mid y=\omega_0) \cdot P(y=\omega_0)}{P(\underline{x})}$$

$$= \frac{\frac{2}{3} \times \frac{3}{8}}{\frac{1}{4}}$$

$$= 1$$

$$P(y=\omega_1 \mid \underline{x}) = \frac{P(\underline{x} \mid y=\omega_1) \cdot P(y=\omega_1)}{P(\underline{x})}$$

$$= \frac{\frac{0}{5} \cdot \frac{5}{8}}{\frac{1}{4}}$$

$$= 0$$

$$\underline{x} = [38, 47]$$

$$P(\underline{x} \mid y=\omega_0) = \frac{h(\underline{x} \mid y=\omega_0)}{\sum_{\underline{x}} h(\underline{x} \mid y=\omega_0)}$$

$$= \frac{0}{3}$$

$$= 0$$

$$P(\underline{x} \mid y = \omega_1) = \frac{h(\underline{x} \mid y = \omega_1)}{\sum_{\underline{x}} h(\underline{x} \mid y = \omega_1)}$$

$$= 0$$

$$P(\underline{x}) = \sum_{\omega_0, \omega_1} P(\underline{x} \mid y_j) \cdot P(y = \omega_j)$$

$$= P(\underline{x} \mid y = \omega_0) \cdot P(y = \omega_0) + P(\underline{x} \mid y = \omega_1) \cdot P(y = \omega_1)$$

$$= 0 \cdot \frac{3}{8} + 0 \cdot \frac{5}{8}$$

$$= 0$$

$$P(y = \omega_0 \mid \underline{x}) = \frac{P(\underline{x} \mid y = \omega_0) \cdot P(y = \omega_0)}{P(\underline{x})}$$

$$= \frac{0 \cdot \frac{3}{8}}{0}$$

$$= 0/0 \text{ (Undefined)}$$

$$P(y = \omega_1 \mid \underline{x}) = \frac{P(\underline{x} \mid y = \omega_1) \cdot P(y = \omega_1)}{P(\underline{x})}$$

$$= \frac{0 \cdot \frac{5}{8}}{0}$$

$$= 0/0 \text{ (undefined)}$$

## Bayes Decision:-

$$P(y=\omega_j \mid \underline{x}) = \frac{P(\underline{x} \mid y=\omega_j) \cdot P(y=\omega_j)}{P(\underline{x})}$$

Likelihood — Non-zero — Non-zero

## Gaussian Likelihood:-

$$P(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad ; \quad x \in \mathbb{R}, \; D \times 2$$

## Naïve Bayes:-

$$P(\underline{x}) = P(x^{(0)}) \cdot P(x^{(1)}) \cdot P(x^{(2)}) \cdots P(x^{(D-1)})$$

Gaussian, Poisson, Rayleigh

(Assumption is $x^{(d)}$ is independent)

## Multivariate Normal Distribution:-

$$P(\underline{x}) = \frac{1}{(2\pi)^{D/2} \left|\sum\right|^{1/2}} \exp -\frac{1}{2}\left((\underline{x}-\underline{\mu})^T \sum^{-1} (\underline{x}-\underline{\mu})\right)$$

$$\left[\underline{x}\right] \rightarrow \begin{array}{l} \underline{x} \in \mathbb{R}^{D \times 1} \\ \underline{\mu} \in \mathbb{R}^{D \times 1} \\ \sum \in \mathbb{R}^{D \times D} \end{array}$$

$$\mu = \left[\mu^{(0)} \; \mu^{(1)} \cdots \mu^{(d)} \cdots \mu^{(D-1)}\right]$$

$$\sum = \begin{bmatrix} \sigma^{(1)^2} & \sigma^{(1,2)^2} \\ \sigma^{(2,1)^2} & \cdots \\ \vdots & \ddots \end{bmatrix}$$

$$\sigma^{(g,h)} = \frac{1}{N} \sum_{i=0}^{N-1} \left(x_i^{(g)} - \mu^{(g)}\right)\left(x_i^{(h)} - \mu^{(h)}\right)$$

| | $x^{(0)}$ | $x^{(1)}$ | $y$ |
|---|---|---|---|
| $\underline{x}_0$ | $x_0^{(0)}$ | $x_0^{(1)}$ | $\omega_0$ |
| $\underline{x}_1$ | $x_1^{(0)}$ | $x_1^{(1)}$ | $\omega_1$ |
| $\underline{x}_2$ | $x_2^{(0)}$ | $x_2^{(1)}$ | $\omega_2$ |
| ⋮ | ⋮ | ⋮ | ⋮ |
| $\underline{x}_{N-1}$ | $x_{N-1}^{(0)}$ | $x_{N-1}^{(1)}$ | $\omega_{N-1}$ |

## Maximum Likelihood Estimation :-

$$P(\underline{x} \mid y = \omega_j) = \frac{1}{(2\pi)^{D/2} \left| \sum_{y=\omega_0} \right|^{1/2}} \exp -\frac{1}{2} \left( (\underline{x} - \underline{\mu}_{y=\omega_0})^T \underline{\sum}_{y=\omega_0}^{-1} (\underline{x} - \underline{\mu}_{y=\omega_0}) \right)$$

$$\underline{\mu}_{y=\omega_j} = \left[ \mu_{y=\omega_j}^{(0)} \quad \mu_{y=\omega_j}^{(1)} \quad \cdots \quad \mu_{y=\omega_j}^{(d)} \quad \cdots \quad \mu_{y=\omega_j}^{(D-1)} \right]$$

$$\mu_{y=\omega_j}^{(d)} = \frac{1}{N_j} \sum_{\forall y_i \in \omega_j} x_i^{(d)}$$

$$N_j = |\underline{y} = \omega_j|$$

$$\underline{\sum}_{y=\omega_j} = \begin{bmatrix} \sigma_{y=\omega_j}^{(0)\,2} & \sigma_{y=\omega_j}^{(0,1)\,2} & \cdots \\ \sigma_{y=\omega_j}^{(1,0)} & \sigma_{y=\omega_j}^{(1)\,2} & \cdots \\ \vdots & \vdots & \cdots \end{bmatrix}$$

$$\sigma_{y=\omega_j}^{(g,h)} = \frac{1}{N_j} \sum (x_i^{(g)} - \mu_{y=\omega_j}^{(g)}) (x_i^{(h)} - \mu_{y=\omega_j}^{(h)})$$

$$P(y=\omega_j|\underline{x}) = \frac{P(\underline{x}|y=\omega_i)\cdot P(y=\omega_i)}{P(\underline{x})}$$

$P(y=\omega_0|\underline{x})$

$P(y=\omega_1|\underline{x})$

⋮

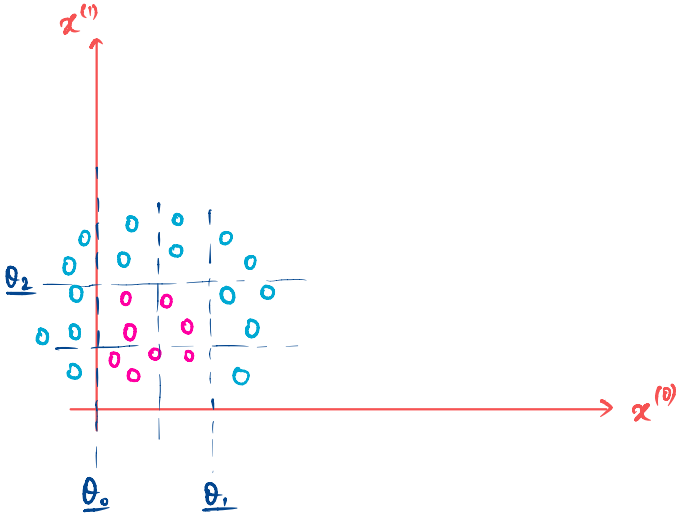
## Maximum Aposteriori (MAP) :-

$$\hat{y} = \underset{\omega_j \in \Omega}{\text{argmax}} \; P(y=\omega_j|\underline{x})$$

$\underline{\Omega} = [\omega_0 \quad \omega_1 \quad \omega_2 \quad \ldots \quad \omega_{K-1}]$

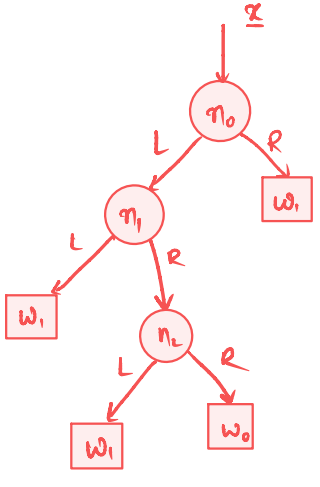
| | $x^{(0)}$ Temp | $x^{(1)}$ Humidity | $y^{(0)}$ Comfort | $y^{(1)}$ N. Comfort | |
|---|---|---|---|---|---|
| $\underline{x}_0^T$ | 31°C | 70 % rh | 0 | 1 | $\underline{y}_0 = \omega_1$ |
| $\underline{x}_1^T$ | 24°C | 40 % rh | 1 | 0 | $\underline{y}_1 = \omega_0$ |
| $\underline{x}_2^T$ | 22°C | 60 % rh | 1 | 0 | $\underline{y}_2 = \omega_0$ |
| $\underline{x}_3^T$ | 49°C | 10 % rh | 0 | 1 | $\underline{y}_3 = \omega_1$ |
| $\underline{x}_4^T$ | 35°C | 30 % rh | 0 | 1 | $\underline{y}_4 = \omega_1$ |
| $\underline{x}_5^T$ | 32°C | 50 % rh | 0 | 1 | $\underline{y}_5 = \omega_1$ |
| $\underline{x}_6^T$ | 27°C | 40 % rh | 1 | 0 | $\underline{y}_6 = \omega_0$ |
| $\underline{x}_7^T$ | 21°C | 20 % rh | 0 | 1 | $\underline{y}_7 = \omega_1$ |

# Decision Trees :-




o - $\omega_0$

o - $\omega_1$



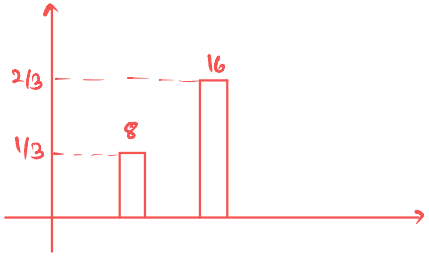

$P(y=\omega_0|x)=0$

$P(y=\omega_1|x)=1$

$n_0$ : if $x^{(0)} > v(\theta_0)$
Goto L-child
else Goto R-child

$n_1$ : if $x^{(1)} > v(\theta_2)$
Goto L-child
else Goto R-child

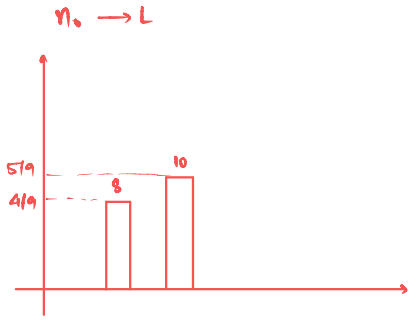
$n_2$ : if $x^{(0)} > v(\theta_1)$
Goto L-child
else Goto R-child

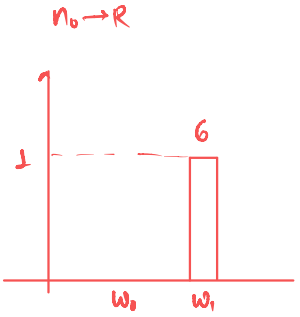



$$P(y=\omega_0|x) = -\frac{|y_j=\omega_0; n_k|}{|n_k|}$$

$$H(n_0) = -\frac{1}{3}\log_2(1/3) - \frac{2}{3}\log_2(2/3)$$

$$= 0.9183$$




$$H(n_0 \rightarrow L) = -\frac{4}{9}\log_2\left(\frac{4}{9}\right) - \frac{5}{9}\log_2\left(\frac{5}{9}\right)$$
$$= 0.9911$$
$$= H(n_1)$$

$$H(n_0 \rightarrow R) = -0\log_2 0 - 1\log_2 1$$
$$= 0$$

$$\text{Information Gain } (IG) = H(n_0) - \left(P(n_0 \rightarrow L)\,H(n_0 \rightarrow L) + P(n_0 \rightarrow R)\,H(n_0 \rightarrow R)\right)$$
$$= 0.9183 - \left(\frac{18}{24} \cdot 0.9911 + \frac{6}{24}(0)\right)$$
$$= 0.2576 \text{ bits}$$

$$(v, \underline{\theta}_0) = \underset{\forall v(\underline{\theta}_j) \in S(n_0)}{\text{argmax}} \left\{ IG(v(\underline{\theta}_j)) \right\}$$

Given a node $n_k$

$$\underline{X}(n_k) = \left\{ (x_0, y_0), (x_1, y_1), \cdots, \right\}$$

| | $x^{(0)}$ | $x^{(1)}$ | $x^{(2)}$ | $\cdots$ | $x^{(D-1)}$ |
|---|---|---|---|---|---|
| $\theta_0$ | $IG(n_0; x^{(0)}, \theta_0^{(0)})$ | | | | |
| $\theta_1$ | $IG(n_0; x^{(0)}, \theta_1^{(0)})$ | | | | |
| $\theta_2$ | | | | | |
| $\vdots$ | | | | | |
| $\theta_{D-1}$ | | | $IG(n_0; x^{(d)}, \theta_2^{(d)})$ | | |

$Q \times D$

$\max(IG)$

$v(\underline{\theta}_2) = \{x^{(d)}, \theta_2^{*(d)}\}$

- Minimum number of samples to split
  - ↳ Minimum leaf Criterion

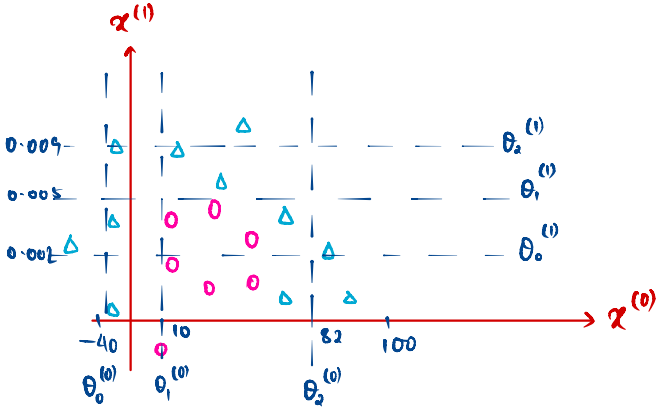
$$P(y = \omega_0 | \underline{x}) = \frac{|y = \omega_0 ; n_k|}{|n_k|}$$

## Class Decision Trees :-

$$\underline{X} = [\underline{x}_0, \underline{x}_1, \cdots, \underline{x}_n, \cdots, \underline{x}_{N-1}] \in \mathbb{R}^{D \times N}$$

$$\underline{Y} = [y_0, y_1, \cdots, y_n, \cdots, y_{N-1}] \in \Omega^N$$

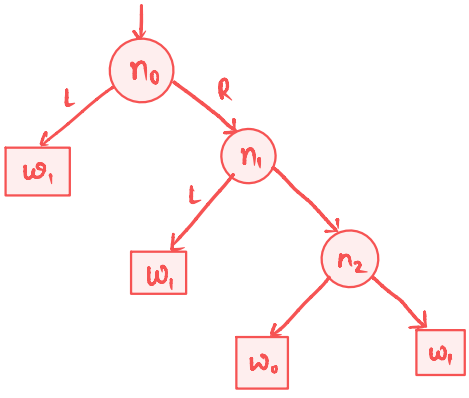
$$\Omega = \{\omega_0, \omega_1\}$$




Axis aligned split -

(a) which axis to split $x^{(0)}$ or $x^{(1)}$

(b) what is the split value $\underline{\theta}_0^x$




$n_0$ : if $x^{(1)} \geq \theta_1^{(1)}$

$\theta_1 = 0.0005$

$n_0 \rightarrow L$

else $n_0 \rightarrow R$

$n_1$ : If $x^{(0)} \geq \theta_1$

$\theta_1 = 68$

$n_1 \rightarrow L$

else $n_1 \rightarrow R$

$n_2$ : If $x^{(0)} \geq \theta_2$

$\theta_2 = 0$

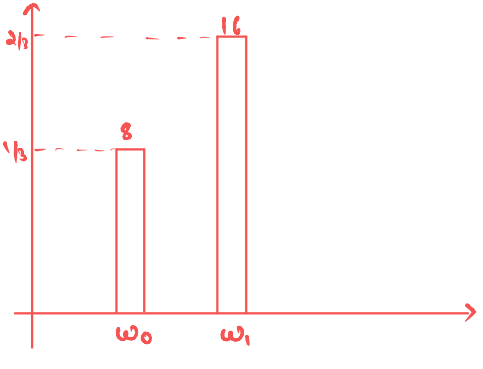
$n_2 \rightarrow L$

else $n_2 \rightarrow R$

| | $x^{(0)}$ | $x^{(1)}$ | |
|---|---|---|---|
| $\theta_1$ | 0.05 | 0.03 | |
| $\theta_2$ | 0.017 | 0.25 | → max |
| $\theta_3$ | 0.07 | 0.05 | |

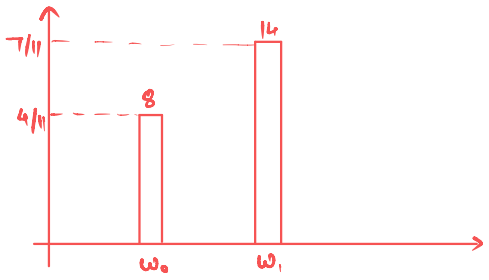
IG $(x^{(d)}, \theta_i^{(d)})$

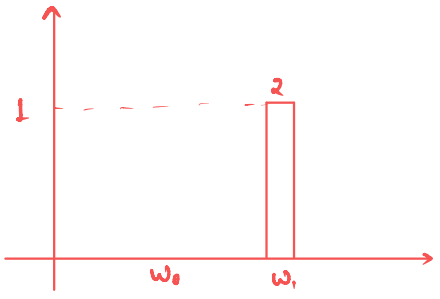
Splitting with $\theta_0^{(0)}$ and $x^0$




$$H(n_0) = -\frac{1}{3}\log_2(1/3) - \frac{2}{3}\log_2(2/3)$$

$$= 0.9183 \text{ bits}$$




$n_0 \rightarrow L$

if $x^{(0)} \geq \theta_0^{(0)}$

$H(n_0 \rightarrow L) = -\frac{4}{11}\log_2(4/11) - \frac{7}{11}\log_2(7/11)$

$= 0.9456$ bits

$n_0 \rightarrow R$

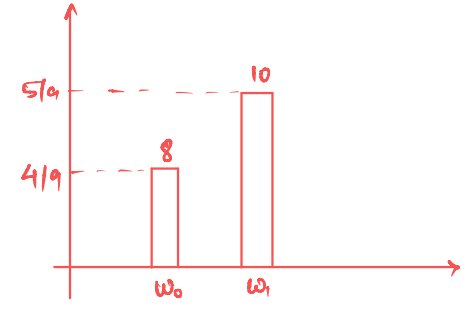
if $x^{(0)} < \theta_1^{(0)}$

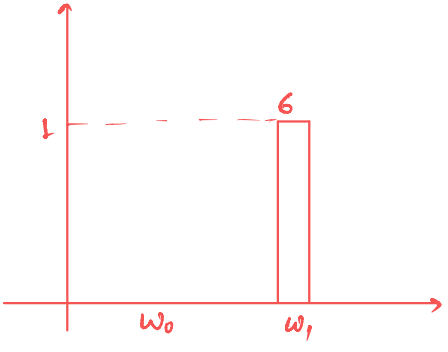
$H(n_0 \rightarrow R) = H(n_0) - (p(n_0 \rightarrow L) H(n_0 \rightarrow L) + p(n_0 \rightarrow R) H(n_0 \rightarrow R))$

$= 0.9183 - \left(\frac{22}{32}(0.9456) + \frac{1}{32}(0)\right)$

$= 0.05$ bits

Splitting with $\theta_1^{(0)}$ on $x^{(0)}$




$n_0 \rightarrow L$

if $x^{(0)} \geq \theta_1^{(0)}$

$H(n_0 \rightarrow L) = 0.9911$

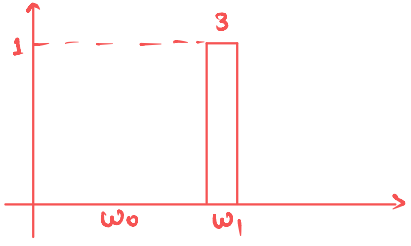
$IG(x^{(0)}, \theta_1^{(0)}) = 0.17$ bits

$n_0 \rightarrow R$

if $x^{(0)} < \theta_1^{(0)}$

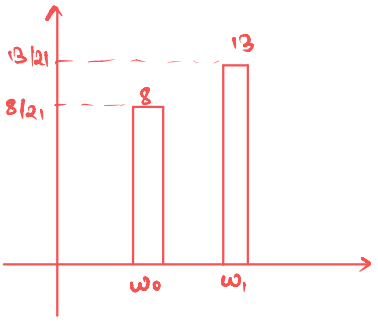
$H(n_0 \rightarrow R) = 0$

Splitting with $\theta_2^{(0)}$ on $x^{(0)}$




$n_0 \rightarrow L$

if $x^{(0)} \gg \theta_2^{(0)}$




$n_0 \rightarrow R$

if $x^{(0)} < \theta_2^{(0)}$

$H(n_0 \rightarrow L) = 0$

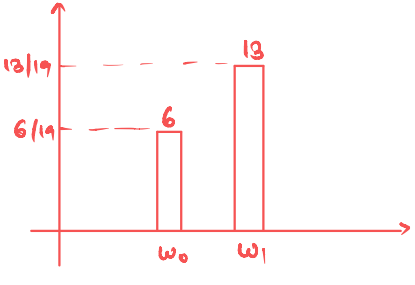
$$H(n_0 \rightarrow R) = -\frac{8}{21}\log(8/21) - \frac{13}{21}\log(13/21)$$

$$= 0.9587 \text{ bits}$$

$$IG(x^{(0)}, \theta_2^{(0)}) = 0.9183 - \left(\frac{3}{24}(0) + \frac{21}{24}(0.2286)\right)$$

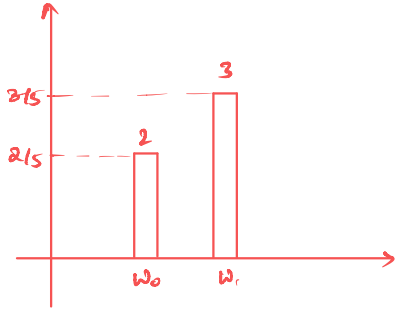
$$= 0.079$$

Splitting with $\theta_0^{(1)}$ on $x^{(1)}$




$n_0 \rightarrow L$

if $x^{(1)} \gg \theta_0^{(1)}$




$n_0 \rightarrow R$

if $x^{(1)} < \theta_0^{(1)}$

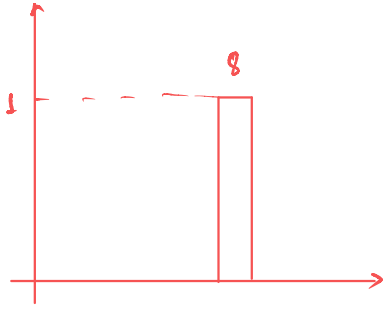
$$H(n_0 \rightarrow L) = -\frac{6}{19}\log_2(6/19) - \frac{13}{19}\log_2(13/19) = 0.899$$

$$H(n_0 \rightarrow R) = -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5) = 0.971$$

$$IG(x^{(1)}, \theta_0^{(1)}) = 0.9183 - \left[\frac{19}{24}(0.899) - \frac{5}{24}(0.971)\right]$$

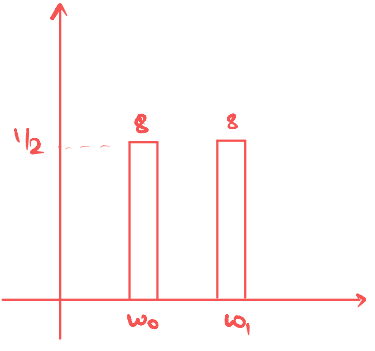
$$= 0.0043$$

Splitting with $\theta_1^{(1)}$ on $x^{(1)}$




$n_0 \rightarrow L$

if $x^{(1)} \geq \theta_1^{(1)}$




$n_0 \rightarrow R$

if $x^{(1)} < \theta_1^{(1)}$

$H(n_0 \rightarrow L) = 0$

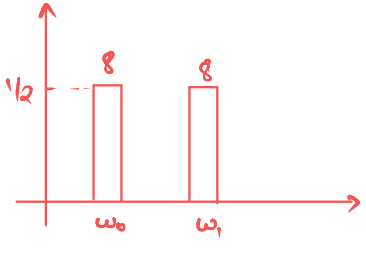
$$H(n_0 \rightarrow R) = -\frac{1}{2}\log_2(1/2) - \frac{1}{2}\log_2(1/2)$$

$$= 1$$

$$IG(x^{(1)}, \theta_1^{(1)}) = 0.9183 - \left(\frac{8}{24}(0) + \frac{16}{24}(1)\right)$$

$$= 0.252$$

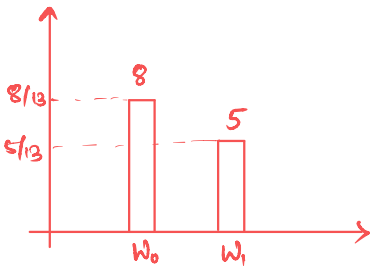
Splitting with $\theta_2^{(1)}$ on $x^{(1)}$




$$H(n_1) = -\frac{1}{2}\log_2(1/2) - \frac{1}{2}\log_2(1/2)$$

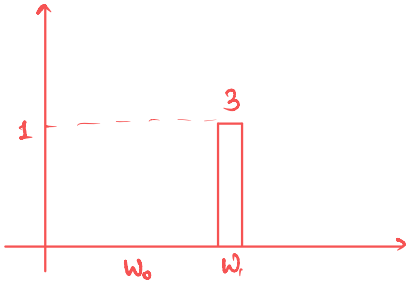
$$= 1 \text{ bit}$$

Splitting with $\theta_0^{(0)}$ and $x^{(0)}$




$n_1 \rightarrow L$

if $x^{(0)} \geq \theta_1^{(0)}$




$n_1 \rightarrow R$

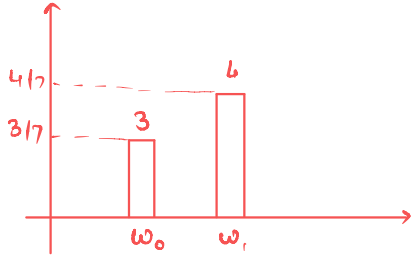
if $x^{(0)} < \theta_1^{(0)}$

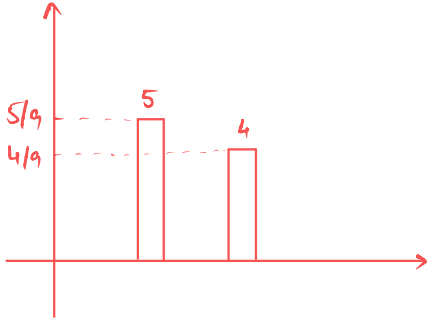
$H(n_1 \rightarrow L) = 0.9612$

$H(n_1 \rightarrow R) = 0$

$IG = 0.21$

Splitting with $\theta_0^{(1)}$ on $x^{(0)}$




$H(n_1 \rightarrow L) = -\frac{3}{7}\log_2(3/7) - \frac{4}{7}\log_2(4/7)$

$= 0.985$

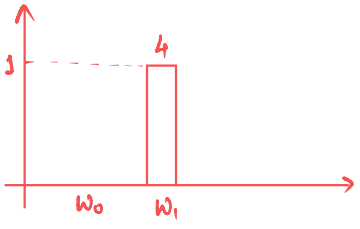
$H(n_1 \rightarrow R) = -\frac{5}{9}\log_2(5/9) - \frac{4}{9}\log_2(4/9)$

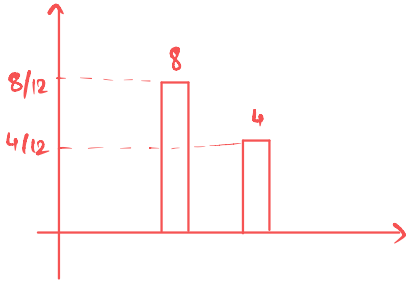
$= 0.991$

$IG = 1 - \left(\frac{7}{16}(0.985) + \frac{9}{16}(0.991)\right)$

$= 0.012$

Splitting with $\theta_0$ on $x^{(0)}$




$H(n_1 \rightarrow L) = 0$

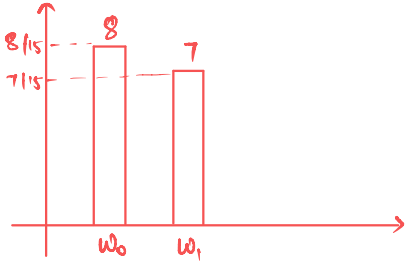
$H(n_1 \rightarrow R) = -\frac{8}{12}\log_2(8/12) - \frac{4}{12}\log_2(4/12)$

$= 0.9183$

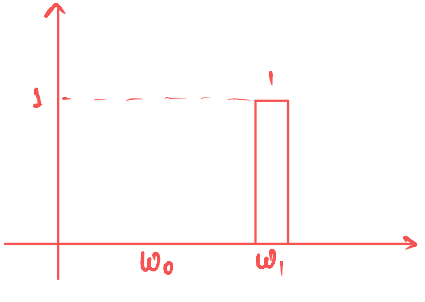
$IG = 1 - \left(0 + \frac{12}{16}(0.9183)\right)$

$= 0.31$

Splitting with $\theta_0^{(1)}$ on $x^{(1)}$




$n_1 \rightarrow L$




$n_1 \rightarrow R$

$H(n_1 \rightarrow L) = -\frac{8}{15}\log_2(8/15) - \frac{7}{15}\log_2(7/15)$

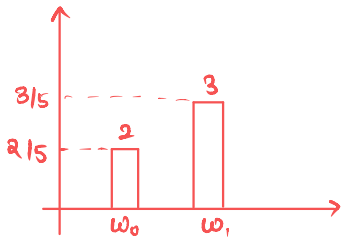
$= 0.997$

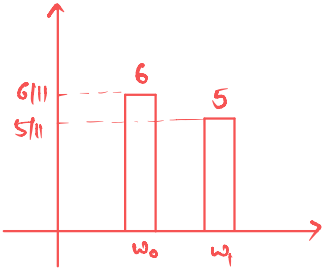
$H(n_1 \rightarrow R) = 0$

$$IG = 1 - \left(\frac{15}{16}(0.997) + 0\right)$$

$$= 0.065$$

Splitting with $\theta_1^{(0)}$ on $x^{(1)}$




$$H(n_1 \to L) = -\frac{2}{5}\log_2(2/5) - \frac{3}{5}\log_2(3/5)$$

$$= 0.9709$$

$$H(n_1 \to R) = -\frac{6}{11}\log_2(6/11) - \frac{5}{11}\log_2(5/11)$$

$$= 0.9940$$

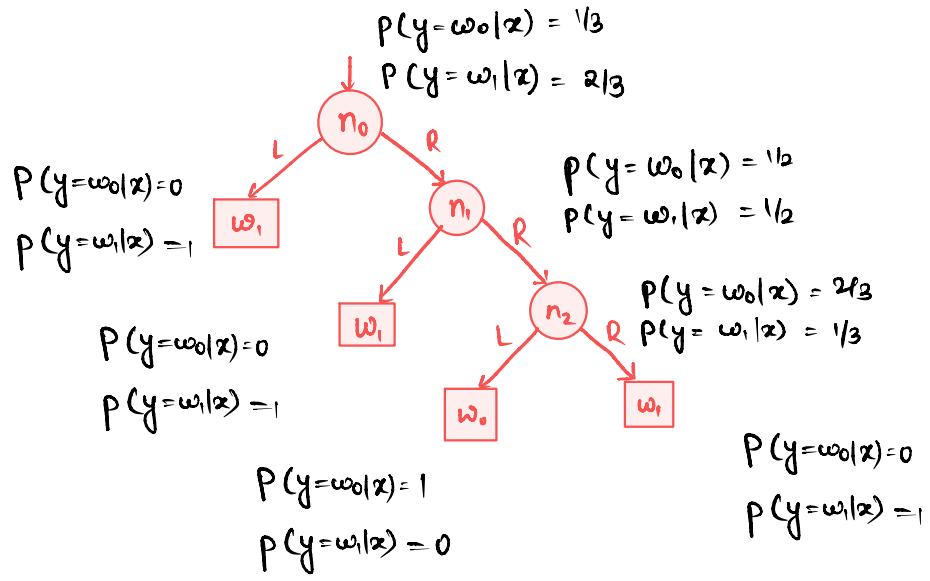
$$IG = 0$$

| | $x^{(0)}$ | $x^{(1)}$ |
|---|---|---|
| $\theta_1$ | 0.21 | 0.06 |
| $\theta_2$ | 0.01 | 0 |
| $\theta_3$ | 0.31 | 0.06 |

Min leaf = 15

Min. no of samples required such that node doesn't become a leaf node.

$N_0 = 24$




## Inference with a Decision Tree:

$\underline{x} = (83, \ 0.0068)$

$x^{(0)}$ $x^{(1)}$

$$\begin{bmatrix} P(y=\omega_0|x) = 0 \\ P(y=\omega_1|x) = 1 \end{bmatrix}$$

$\underline{x} = (5, \ 0.0023)$

$x^{(0)}$ $x^{(1)}$

$$\begin{bmatrix} P(y=\omega_0|x) = 2/3 \\ P(y=\omega_1|x) = 1/3 \end{bmatrix}$$

## Training - grow with the tree (DT):

# Compute $\propto$ No. of nodes in a DT

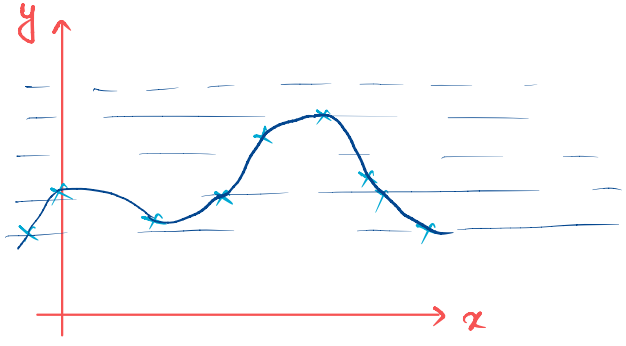
# Space $\propto$ No. of nodes in a DT.

## Inference - Infer with a DT:

# Compute $\propto$ Height of DT

# Space $\propto$ Nodes in a DT

## Regression problems:-




$y \rightarrow \hat{y}$

$y = \text{freq}(x)$

$$\underline{X} = [\underline{x_0}, \underline{x_1}, \cdots, \underline{x_n}, \cdots, \underline{x_{N-1}}] \in \mathbb{R}^{D \times N}$$

$$\underline{Y} = [y_0, y_1, \cdots, \underline{y_n}, \cdots, \underline{y_{N-1}}] \in \mathbb{R}^{B \times N}$$

$$\hat{\underline{Y}} = [\hat{y}_0, \hat{y}_1, \cdots, \hat{\underline{y}}_n, \cdots, \hat{\underline{y}}_{N-1}] \in \Omega^N$$

$$\Omega = \{w_0, w_1, \cdots, w_j\}$$

$$w_j = [y_0^{(0)} - y_1^{(0)}, y_1^{(1)} - y_2^{(1)}, \cdots]$$

# Random Forest :-

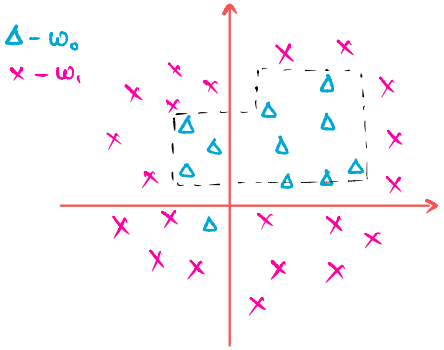
i) Ensemble ML

ii) Classifier Combination

iii) Distribution Consensus

$\underline{X} = \{\underline{x_0}, \underline{x_1}, \cdots, \underline{x_n}, \cdots, \underline{x_{N-1}}\}$ $\quad x_n \in \mathbb{R}^{D\times 1}$

$Y = \{y_0, y_1, \cdots, \underline{y_n}, \cdots, \underline{y_{N-1}}\}$ $\quad y_n \in \Omega = \{\omega_0, \omega_1, \cdots, \omega_{k-1}\}$




$P(y=\omega_0|x)$

$\hat{y} = \underset{\forall \omega_k \in \Omega}{\text{argmax}} \{P(y=\omega_k|x)\}$

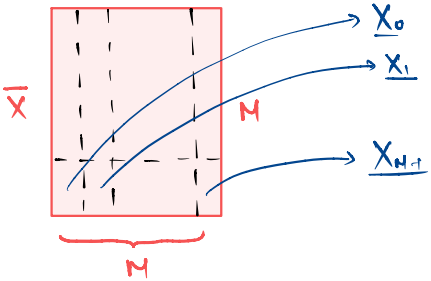
$\underline{X} = \underline{X_0} \cup \underline{X_1} \cup \underline{X_2} \cup \cdots \cup X_{m-1}$

such that $\underline{X_m} \cap \underline{X_0} = \phi$, $m \neq 0$

$\updownarrow$

$Y_m$

$$P(y=\omega_k)\Big|_{\forall y \in Y_m} \cong P(y=\omega_k)\Big|_{\forall y \in Y_0} \cong P(y=\omega_k)\Big|_{\forall y \in Y}$$




$T_m = DT(\underline{X_m}, \underline{Y_m})$

$\{x_m, y_m\} \xrightarrow[DT]{Train} T_m$

$x \xrightarrow{Inference} T_m(x) = \left[P(y=\omega_k|x)_m\right]_{\forall \omega_k \in \Omega}$

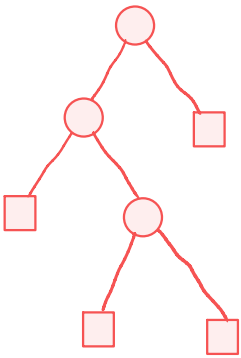
$P(y=\omega_0|x)_m$

$P(y=\omega_1|x)_m$

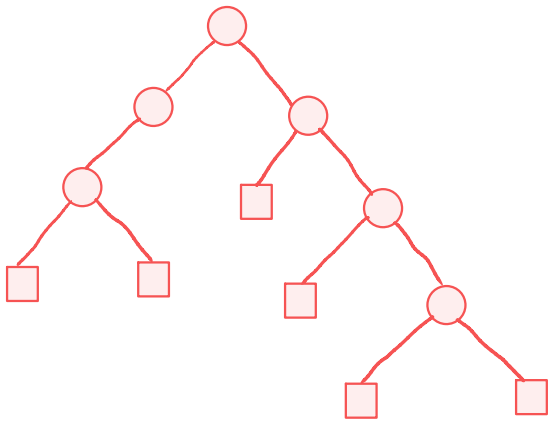
---

$P_{RF}(y=\omega_k|x) = \frac{1}{M} \sum_{m=0}^{M-1} P(y=\omega_k|x)_m$

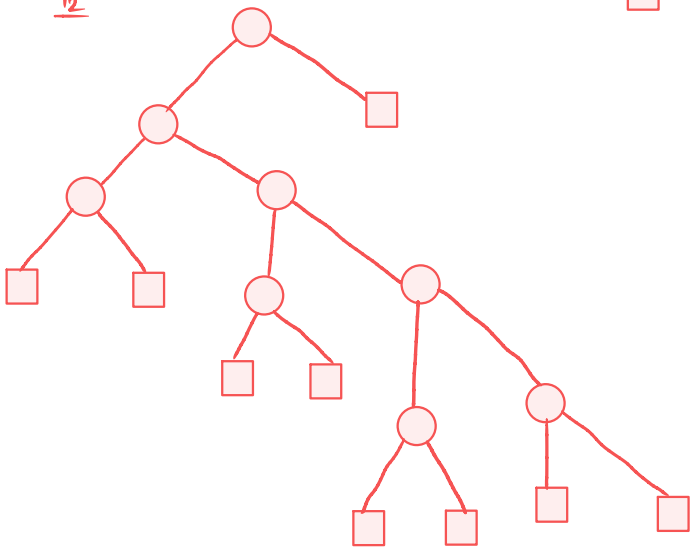
$T_0$



$T_1$



$T_2$



Bootstrap Sampling Aggregation → Bagging

# Mixture Model

Likelihood, Prior

$$P(y=\omega_k \mid \underline{x}) = \frac{P(\underline{x} \mid y=\omega_k) \cdot P(y=\omega_k)}{P(\underline{x})}$$

Posterior, Evidence

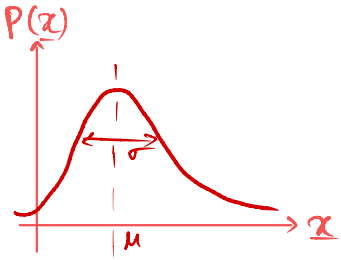
$$P(\underline{x}) = \sum_{\forall \omega_k \in \Omega} P(\underline{x} \mid y=\omega_k) \cdot P(y=\omega_k)$$

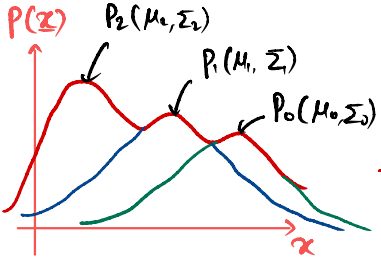
$$\hat{y} = \underset{\forall \omega_k \in \Omega}{\operatorname{argmax}} \{P(y=\omega_k \mid \underline{x})\}$$

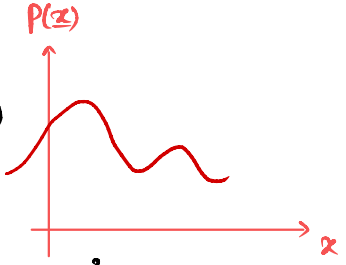
$$P(\underline{x} \mid y=\omega_k) \sim \text{Multivariate Normal Distribution}$$

$$\sim \frac{1}{(2\pi)^{D/2} |\Sigma|^{1/2}} \exp\left(\frac{-1}{2} (\underline{x}-\underline{\mu})^T \Sigma^{-1} (\underline{x}-\underline{\mu})\right)$$

Assumption is that $P(\underline{x} \mid y=\omega_k)$ is unimodal



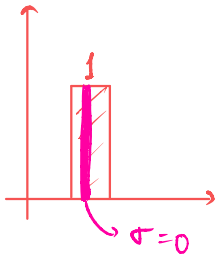




$$P(\underline{x}) = \sum_{j=0}^{2} P_j \mathcal{N}(\mu_j, \Sigma_j)$$

$$\int_{\underline{x}} P(\underline{x}) d\underline{x} = 1, \quad \sum_{j=0}^{2} P(\underline{x}) = 1$$

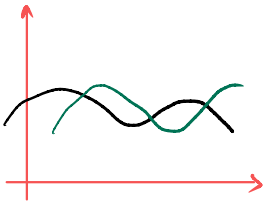
- You have only samples $\underline{X}, Y = (\underline{x}, y)$

- You have to model $P(\underline{x}|y=\omega_k)$




$\underline{\mu}_0, \underline{\Sigma}_0, P_0$
$\underline{\mu}_1, \underline{\Sigma}_1, P_1$
$\underline{\mu}_2, \underline{\Sigma}_2, P_2$

$\longrightarrow \Theta$ (Parameter Vector of the GMM)



B. $\underline{x} \in \mathbb{R}^D$.

Each dim has 10 bins if $D=10$ then I have $10^{10}$ bins.

$P(\Theta) \longrightarrow \mathcal{Q}(\Theta)$

↳ property denoting the goodness of the GMM.

$\frac{\partial \mathcal{Q}(\Theta)}{\partial \Theta} = 0$

Gaussian Mixture Model (GMM) $\xrightarrow{\text{Solve}}$ Expectation - Maximization (EM) algorithm.

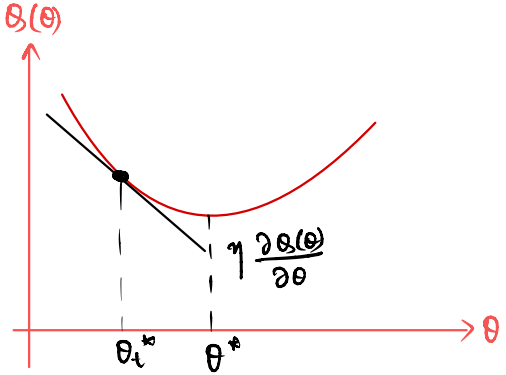
E Step: Random Guess of $\Theta$.

$$\mathcal{Q}(\underline{\Theta}, \underline{\Theta}_t) = E\left(-\sum_n \ln\left(P(\underline{x}_n; \underline{\Theta}|\Theta_t)\right)\right)$$

$$\underline{X} = [\underline{x}_0, \underline{x}_1, \cdots, \underline{x}_n, \cdots, \underline{x}_{N-1}]$$

M Step: $\frac{\partial \mathcal{Q}(\underline{\Theta}, \underline{\Theta}_t)}{\partial \Theta} = 0 \;\Big|\; \underline{\Theta}_{t+1} = \Theta_t - \eta \frac{\partial \mathcal{Q}(\underline{\Theta}; \underline{\Theta}_t)}{\partial \Theta}$

↳ Learning Rate | Gradient Control




$$\underline{X} = [\underline{x_0}, \underline{x_1}, \cdots, \underline{x_n}, \cdots, \underline{x_{N-1}}] \in \mathbb{R}^{D \times N}$$

$$Y = [y_0, y_1, \cdots, y_n, \cdots, y_{N-1}] \in \Omega^N$$

$$\Omega = \{\omega_0, \omega_1, \cdots, \omega_{K-1}\}$$

S-I : $\underline{X}|_{\omega_0}, \underline{X}|_{\omega_1}, \cdots \underline{X}|_{\omega_{K-1}}$

$= [x_0 | y_0 = \omega_0, x_1 | y_1 = \omega_1, \cdots, x_n | y_n = \omega_n]$

S-II : Estimate $p(\underline{x} | y = \omega_k)$ using GMM.

Select the no. of components : $j = 3$

Set initial $\theta_{t=0}|_{\omega_k} = \{\underline{\mu_0}, \underline{\Sigma_0}, p_0, \underline{\mu_1}, \underline{\Sigma_1}, p_1, \underline{\mu_2}, \underline{\Sigma_2}, p_2\}$

Solve EM algorithm and get $\theta^*|_{\omega_k}$

S-III : $p(\underline{x} | y = \omega_k) \simeq GMM(\theta^*|_{y=\omega_k})$

$$p(y = \omega_k), p(\underline{x}) = \sum_{\forall \omega_k \in \Omega} p(\underline{x} | y = \omega_k) \cdot p(y = \omega_k)$$

$$\hat{y} = \underset{\forall \omega_k \in \Omega}{\text{argmax}} \left( p(y = \omega_k | \underline{x}) \right)$$

where $p(y = \omega_k | \underline{x}) = \dfrac{p(\underline{x} | y = \omega_k) \cdot p(y = \omega_k)}{p(\underline{x})}$

There is a crucial assumption for j.

↳ Random guess of $J_s = 0$.

E-step: for $J_s$

- E-step for $\mathcal{Q}(\theta_t)$
- M-step for $\theta_{t+1}$

M-Step for $J_{s+1}$

Unsupervised Learning for Finite Mixture Models

↳ Figuerdo and Jain - IEEE PAMI 24(3), 2002

## Feature Selection :-

$\underline{x} \in \mathbb{R}^{D \times 1}$

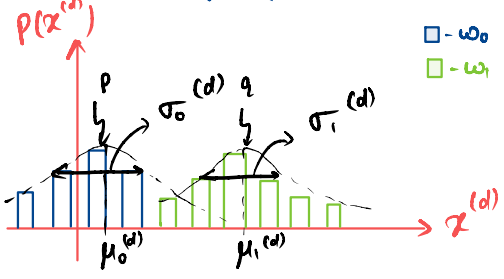
$y \in \Omega^{K \times 1}$

$\Omega = \{\omega_0, \omega_1, \cdots, \omega_{K-1}\}$

$$\underline{x} = [x^{(0)}, x^{(1)}, \cdots, x^{(d)}, \cdots, x^{(D-1)}]$$

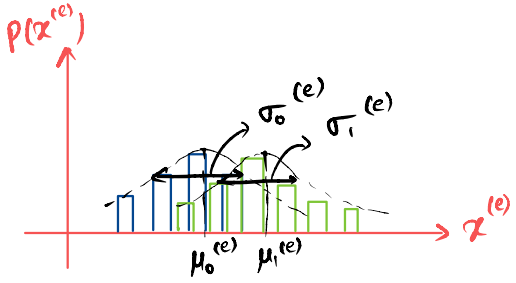
- Each feature is not equivocally relevant for a classification task.

$\varphi(x^{(d)})$ → Relevance of a feature $x^{(d)}$ to solve a classification task.




$$\varphi(x^{(d)}) = \frac{|\mu_0^{(d)} - \mu_1^{(d)}|^2}{\sigma_0^{(d)2} + \sigma_1^{(d)2}}$$

$$= \frac{\text{Inter Class Variance}}{\text{Intra Class Variance}}$$




$$\phi(x^{(e)}) = \frac{|\mu_0^{(e)} - \mu_1^{(e)}|^2}{\sigma_0^{(e)2} + \sigma_1^{(e)2}}$$

- Relevance of $x^{(e)}$ is lesser than relevance of $x^{(d)}$ since $|\mu_0^{(e)} - \mu_1^{(e)}| < |\mu_0^{(d)} - \mu_1^{(d)}|$.

## Kulback Leiber Divergence:

$$d_{KL}(P||q) = \sum_{\forall i} P_i \log\left(\frac{P_i}{q_i}\right)$$

$$d_{KL}(P||q) \neq d_{KL}(q||P)$$

## Jensen-Shannon Divergence:

$$d_{JS}(P||q) = \sum_{\forall i} \left(\frac{P_i}{c_i} \log_2\left(\frac{P_i}{c_i}\right) + \frac{q_i}{c_i} \log_2\left(\frac{q_i}{c_i}\right)\right); \quad c_i = \frac{P_i + q_i}{2}$$

## Filter Model of Feature Selection:

- for $d = 0, \ldots, (D-1)$

  $\phi(x^{(d)}) = \phi_{JS}(P||q)$ ;

  $P \sim h(x^{(d)}|_{y=\omega_0})$ ;

  $q \sim h(x^{(d)}|_{y=\omega_1})$ ;

  ↖ histogram

  end

$$\underline{x}^* = [x^{*(0)} \quad x^{*(1)} \quad x^{*(2)}, \ldots, x^{*(d)}, \ldots, x^{*(D-1)}]$$

such that $\phi(x^{*(0)}) \geq \phi(x^{*(1)}) \geq \cdots \geq \phi(x^{*(d)}) \geq \cdots \geq \phi(x^{*(D-1)})$

## Wrapper Model of Feature Selection:

$\hat{y} = F(x)$, $F : \mathbb{R}^{D\times 1} \longmapsto \Omega^{k\times 1}$

$x^{(0)}, x^{(1)}, \ldots, x^{(d)}, \ldots, x^{(D-1)}$

↳ pick any one

$\tilde{x}^{(j)} \longrightarrow F_j(\cdot) \xrightarrow{\text{Accuracy}} a_j$

$F_0(x^{(0)}) \longrightarrow a_0$

$F_1(x^{(1)}) \longrightarrow a_1$

$\vdots$

$F_{D-1}(x^{(D-1)}) \longrightarrow a_{D-1}$

$\xrightarrow{\text{argmax} = x^{(7)}} \tilde{x}_{(0)} = [x^{(7)}]$

↳ pick any one not present $\tilde{x}_{(0)}$

$F_0(x^{(0)}) \longrightarrow a_0$

$F_1(x^{(1)}) \longrightarrow a_1$

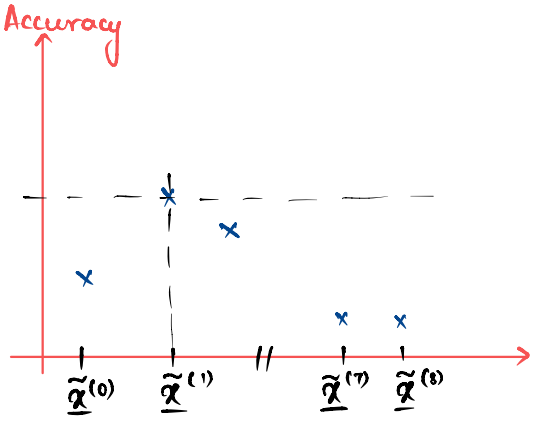
$\vdots$

$F_{D-1}(x^{(D-1)}) \longrightarrow a_{D-1}$

↳ Accuracy

$\xrightarrow{\text{argmax} = x^{(0)}} \tilde{x}_{(0)} = [x^{(7)}]$

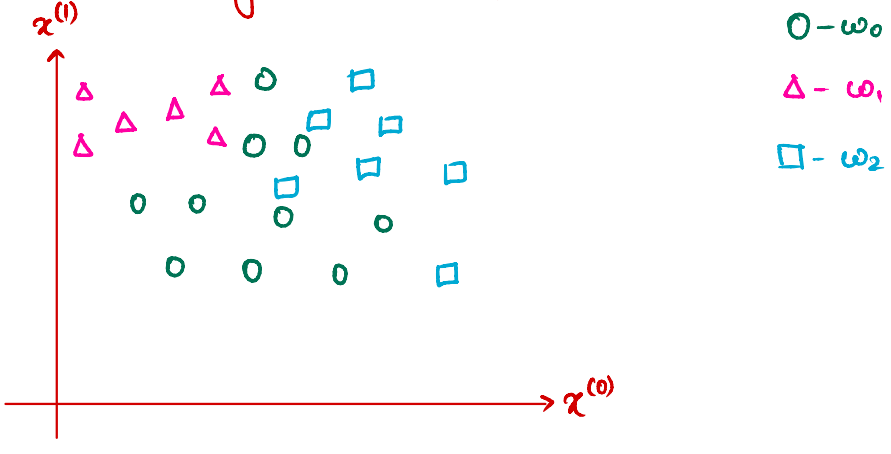
$\tilde{x}_{(1)} = [x^{(7)}, x^{(1)}]$




Optimal set of features relevant to F.

Sequential Forward Search (SFS)

## K- Nearest Neighbour Classification (KNN):-




$$\underline{X} = [\underline{x}_0, \underline{x}_1, \cdots, \underline{x}_n, \cdots, \underline{x}_{N-1}] \in \mathbb{R}^{D \times N}$$

Such that

$$\underline{Y} = [y_0, y_1, \cdots, \underline{y_n}, \cdots, y_{N-1}] \in \Omega^N$$

Such that $\Omega = \{\omega_0, \omega_1, \omega_2\}$

$$d(\underline{x}_n, \underline{x}_j) = \| \underline{x}_n - \underline{x}_j \|_2$$

- Sort based on $d(\underline{x}_n, \underline{x}_j)$

$\tilde{X} = [\tilde{\underline{x}}_0, \tilde{\underline{x}}_1, \cdots, \tilde{\underline{x}}_n, \cdots, \tilde{\underline{x}}_{N-1}] \in \mathbb{R}^{D \times N}$

such that $\tilde{\underline{x}}_n \in \mathbb{R}^{D \times 1}$, $d(\tilde{\underline{x}}_0, \underline{x}_j) \leq d(\tilde{\underline{x}}_1, \underline{x}_j) \leq d(\tilde{\underline{x}}_2, \underline{x}_j) \cdots \leq d(\tilde{\underline{x}}_n, \underline{x}_j)$
$\cdots \leq d(\tilde{\underline{x}}_{N-1}, \underline{x}_j)$

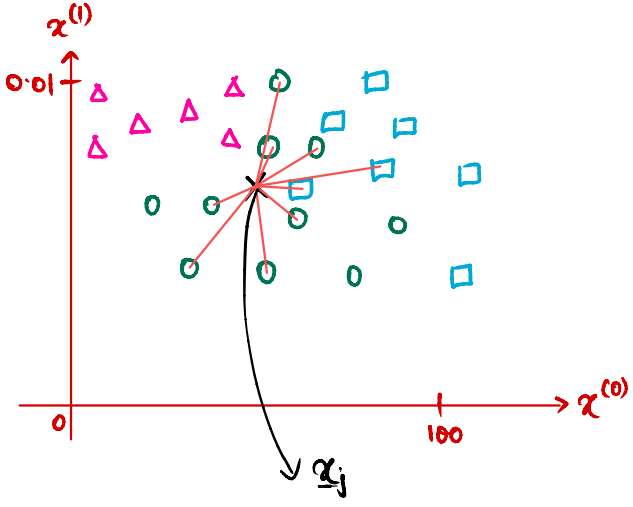
$\tilde{Y} = [\tilde{y}_0, \tilde{y}_1, \cdots, \tilde{y}_n, \cdots, \tilde{y}_{N-1}] \in \Omega^N$

such that $\Omega = \{\omega_0, \omega_1, \omega_2\}$

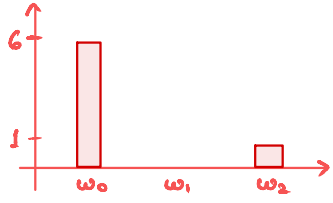
- $\tilde{\underline{X}}_K \subset \tilde{\underline{X}}$

$\tilde{X}_K = [\tilde{x}_0, \tilde{x}_1, \cdots \tilde{x}_{K-1}]$

max votes $\left(\tilde{Y}_K = [\tilde{y}_0, \tilde{y}_1, \cdots, \tilde{y}_{K-1}]\right) = \omega_0$




K=8
↳ Issue of selection of K



- $\cos\theta = \frac{|\underline{x}_n \cdot \underline{x}_j|}{\|\underline{x}_n\|_2 \cdot \|\underline{x}_j\|}$

→ Cosine similarity distance

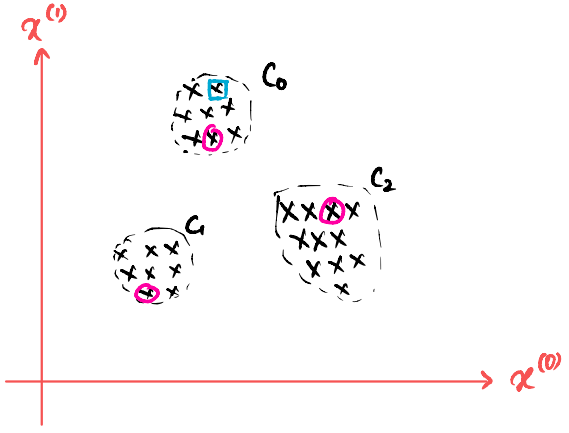
## Clustering :-

- Unsupervised method

$\underline{X} = [\underline{x}_0, \underline{x}_1, \cdots, \underline{x}_n, \cdots, \underline{x}_{N-1}] \in \mathbb{R}^{D \times N}$

such that $\underline{x}_n \in \mathbb{R}^{D \times 1}$

$\underline{Y}$ = Not Available




# Clusters = $K = 3$

$\tilde{\underline{Y}} = [y_0, y_1, \cdots, y_n, \cdots, y_{N-1}] \in C^N$

$C = \{C_0, C_1, \cdots, C_k, \cdots C_{K-1}\}$

$\underline{\mu}^{(t)}$ - State of cluster centroid at $t^{th}$ iteration.

$\underline{\mu}^{(t=0)} = [\underline{\mu}_0^{(t=0)}, \underline{\mu}_1^{(t=0)}, \underline{\mu}_2^{(t=0)}] \in \mathbb{R}^{D \times K} \sim \underline{X}$

such that $\mu_k^{(t)} \in \mathbb{R}^{D \times 1}$

$\underline{X} = \{\underline{x}\} - \{\underline{\mu}^{(t=0)}\}$

for $t = 1 : N-K-1$

$\underline{x}^{(t)} \xrightarrow[\text{rep}]{\text{w/o}} \{\tilde{\underline{X}}\}$

for $k = 0 : (K-1)$

$d_k^{(t)} = \| \mu_k^{(t)} - x^{(t)} \|$

end

$\tilde{y}^{(t)} = \text{argmin} \{d_k^{(t)}\}$

$$\mu_{\tilde{y}^{(t)}}^{(t+1)} = \frac{1}{|\tilde{y}^{(t)}|} \left( (|\tilde{y}^{(t)}| - 1) \mu_{\tilde{y}^{(t)}}^{(t)} + \underline{x}^{(t)} \right)$$

$|C_0|_{t=0} = 1$ $\quad |C_0|_{t=1} = 2$ $\quad |C_1|_{t=1} = 1$ $\quad |C_2|_{t=1} = 1$

- Hyperparameters
  - K - No. of clusters
  - Initial Seed

K - Means - Clustering

- Iterative Agglomerative Clustering.

$\theta = 0$

$\mu^{(\theta=0)} \sim \{\underline{x}\}$ (K-pts)

do

$\underline{\mu}^{(t=0)} = \underline{\mu}^{(\theta=0)}$

Call : K-Means clustering

$\underline{\mu}^{(\theta+1)} = \underline{\mu}^{(t=N-K)}$, $\theta = \theta + 1$

while $\| \underline{\mu}^{(\theta)} - \underline{\mu}^{(\theta+1)} \| > \epsilon$

↳ Hyperparameter / Tolerance

- E-Step:

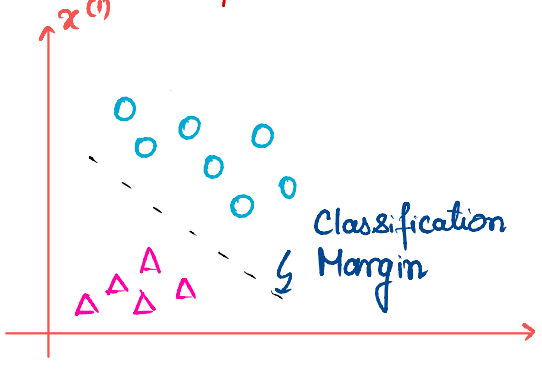
  Guess K

  M-Step:

  Solve GoF with K

- Cluster Compactness Index (CCI) = $\frac{\text{Intra Cluster Index}}{\text{Inter Cluster Index}}$

## Domain Adaptation and Transfer Learning :-



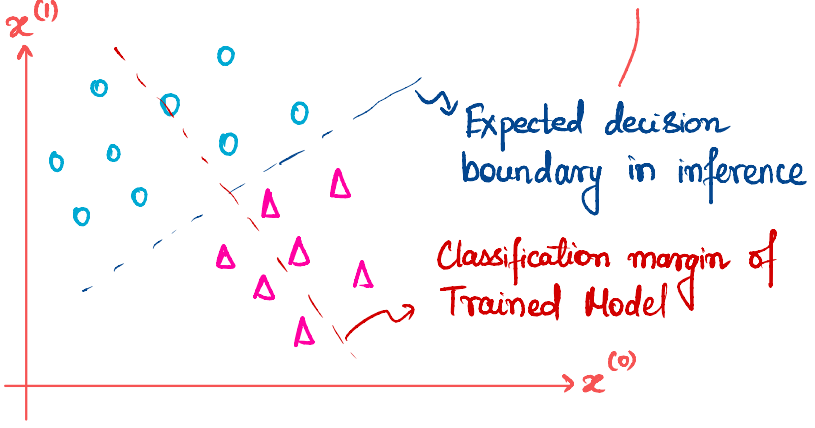

$\Delta - w_0$

$O - w_1$

Training

$E(\underline{\phi}_{Trg}(\underline{X}_{Inf})) \longrightarrow \infty$

$T(0)$




$\underline{x}_i \in \underline{X}_{Inf}$

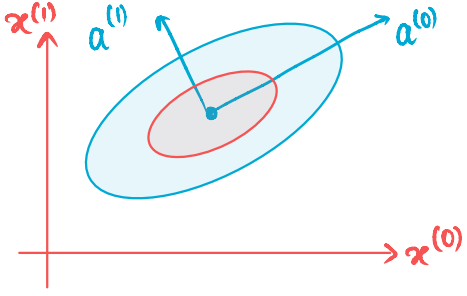
Domain Adaptation

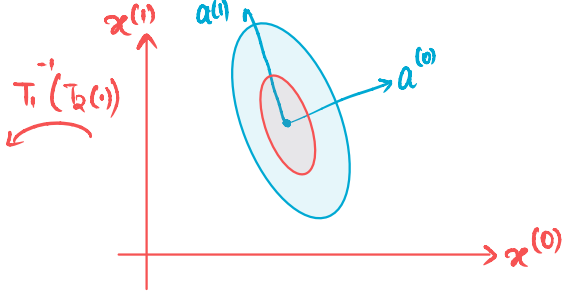
- (A) Feature Transformation
- (B) Model Adaptation

$\underline{\phi}$ : Mathematical Representation of the Classifier model.

$\underline{\phi}_{Trg}$ : $\underline{\phi}$ trained on $\underline{X}_{Trg}$

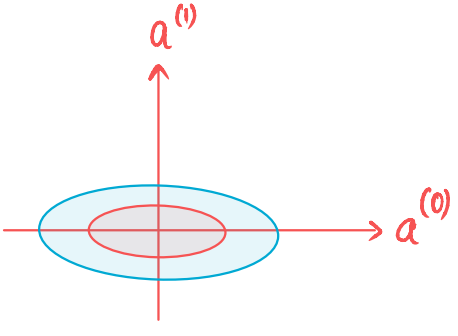
$\underline{\phi}_{Inf}$ : $\underline{\phi}$ trained on $\underline{X}_{Inf}$

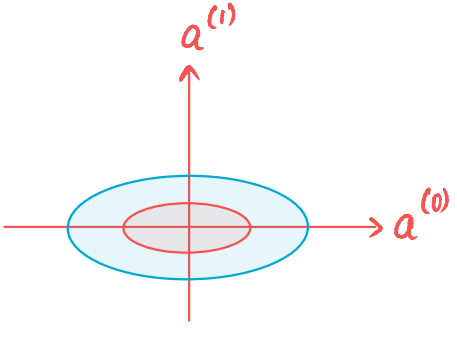



$T_3(\cdot)$ PCA projection

$T_2(\cdot)$



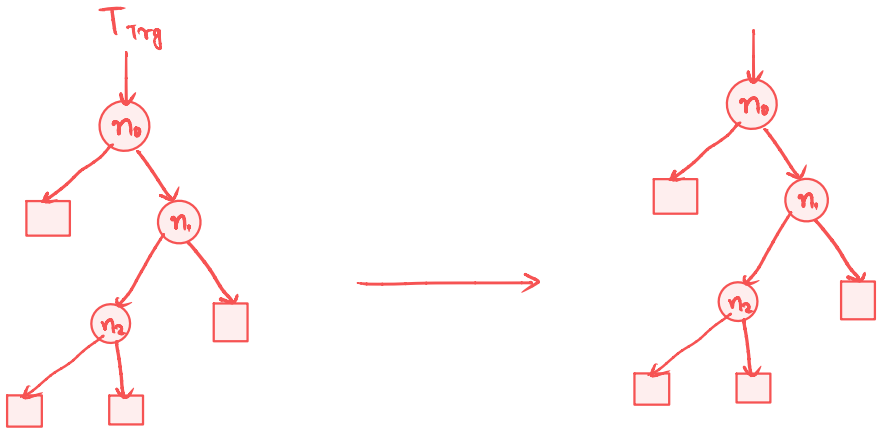

$\underline{X} \neq \underline{X}_{Inf}$

$\underline{\Phi}_{Trg} \neq \underline{\Phi}_{Inf}$

TL

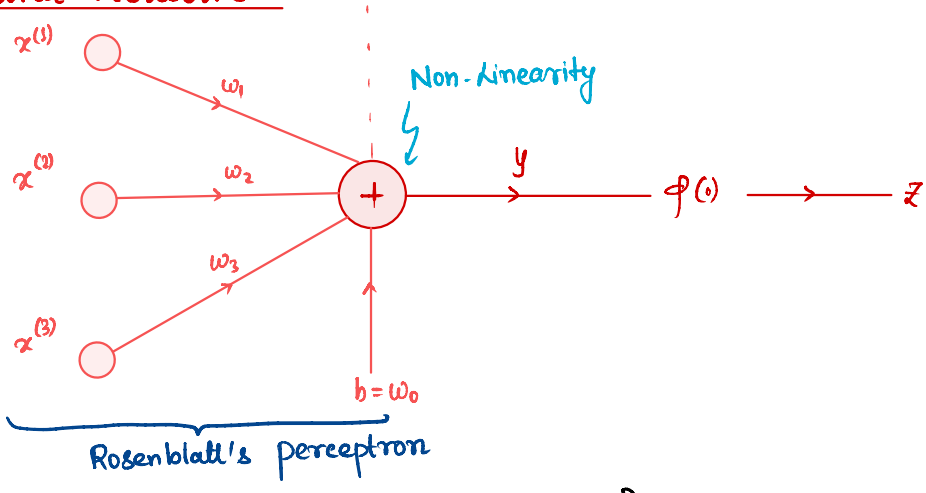
$$\underline{\Phi}_{Trg \to Inf} \simeq \underline{\Phi}_{Inf} \equiv E\left(\underline{\Phi}_{Trg \to Inf}\left(\underline{X}_{Inf}\right)\right) = E\left(\underline{\Phi}_{Inf}\left(\underline{X}_{Inf}\right)\right) + \epsilon$$

such that $\epsilon \to 0$.




$$n_0 := f_{n_0}(x) = \begin{cases} 1 & \text{if } x^{(d)} \geq \theta_{n_0}^{(d)} \\ 0 & \text{otherwise} \end{cases}$$

## Neural Networks :-




$$\underline{X} = [x^{(0)} \quad x^{(1)} \quad \dots \quad x^{(d)} \quad \dots \quad x^{(D-1)}] \in \mathbb{R}^D \qquad \text{(Not for NN)}$$

$$\underline{X} = [x^{(1)} \quad x^{(2)} \quad \dots \quad x^{(d)} \quad \dots \quad x^{(D)}] \in \mathbb{R}^D \qquad \text{(For NN)}$$

$$y = \omega_1 x^{(1)} + \omega_2 x^{(2)} + \omega_3 x^{(3)} + b$$

Bias

$$= \begin{bmatrix} w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} x^{(1)} \\ x^{(2)} \\ x^{(3)} \end{bmatrix} + b$$

Weights

Tunable / Variable factors

$\because b = w_0$,

$$y = \begin{bmatrix} w_0 & w_1 & w_2 & w_3 \end{bmatrix} \begin{bmatrix} 1 \\ x^{(1)} \\ x^{(2)} \\ x^{(3)} \end{bmatrix}$$

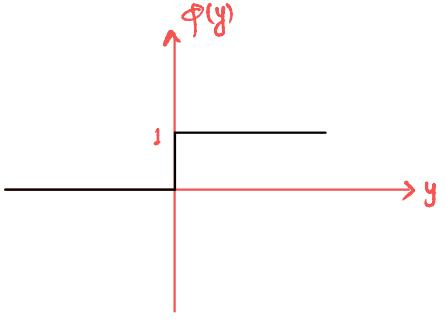
$$= \underline{W}\,\underline{X} + b$$

$$= \underline{W}^*\,\underline{X}^*$$

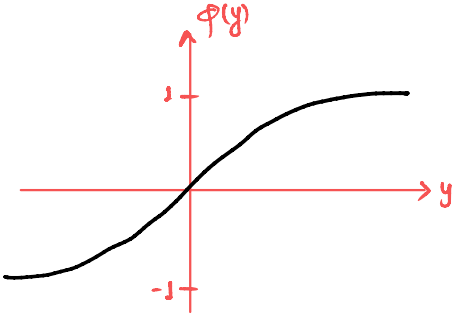
Such that $\underline{W}^* = [w_0 \; w_1 \; w_2 \; w_3]$ and $\underline{X}^* = [1 \quad \underline{X}]^T$

$z = \varphi(y)$

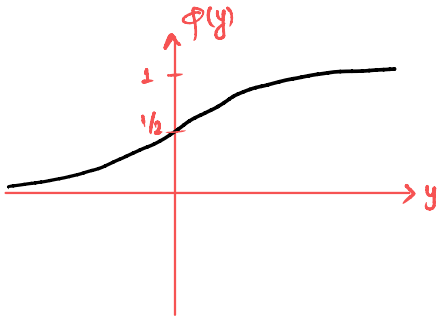
$$\varphi(y) = \begin{cases} 0 & y \le 0 \\ 1 & \text{otherwise} \end{cases}$$

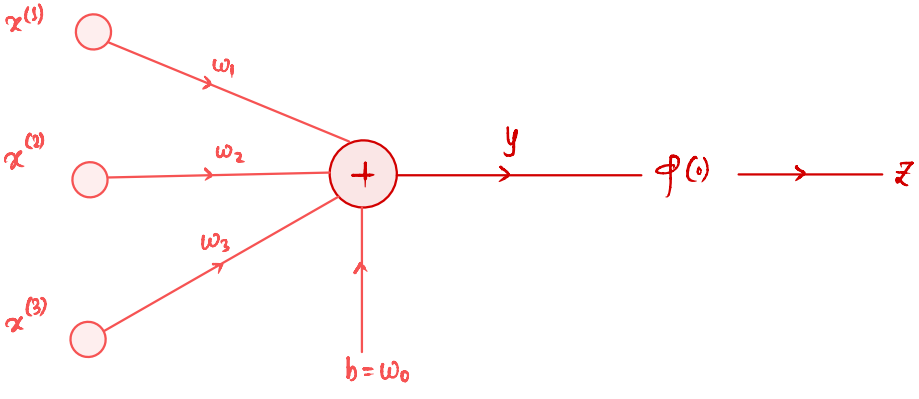



$$\varphi(y) = \tanh y = \frac{e^{y} - e^{-y}}{e^{y} + e^{-y}}$$

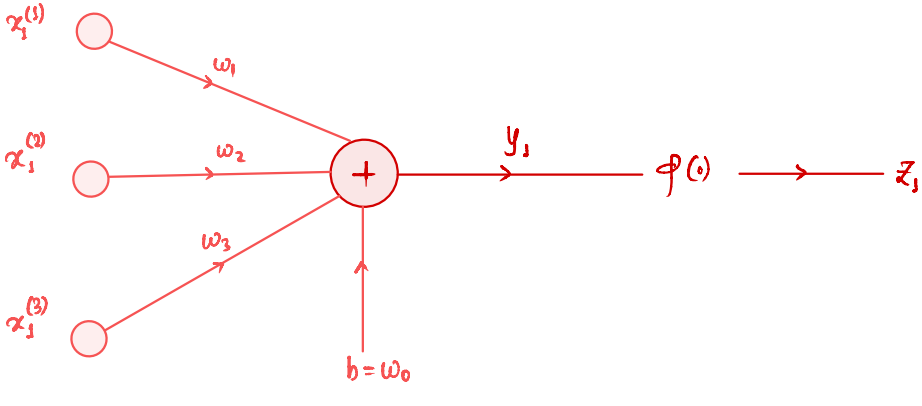


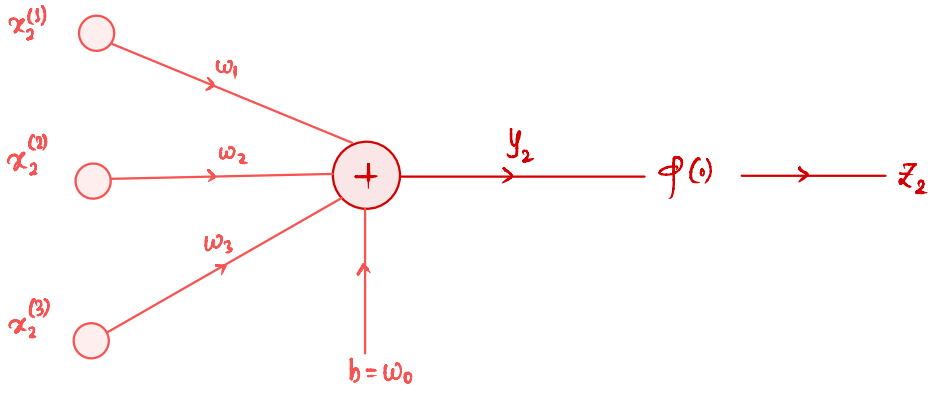
$\varphi(y) = \text{sigmoid}(y) = \frac{1}{1+e^{-y}}$



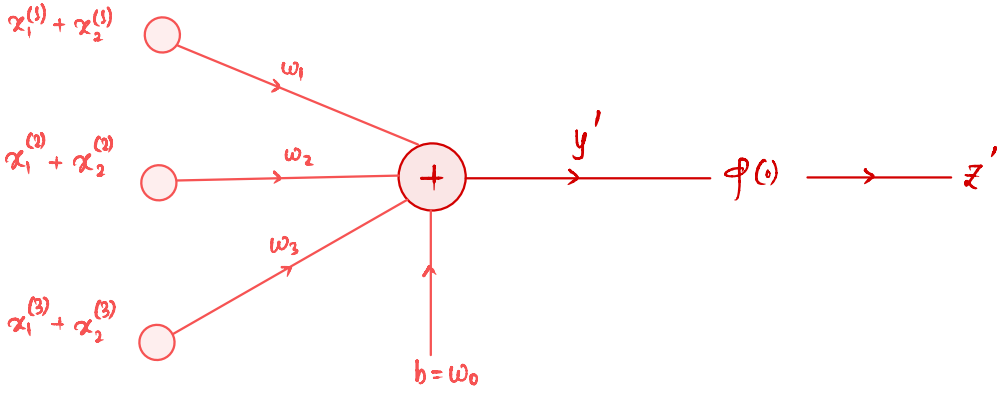

Is this system linear ? prove it.






$$y_1 = w_0 + w_1 x_1^{(1)} + w_2 x_1^{(2)} + w_3 x_1^{(3)}$$

$$y_2 = w_0 + w_1 x_2^{(1)} + w_2 x_2^{(2)} + w_3 x_2^{(3)}$$

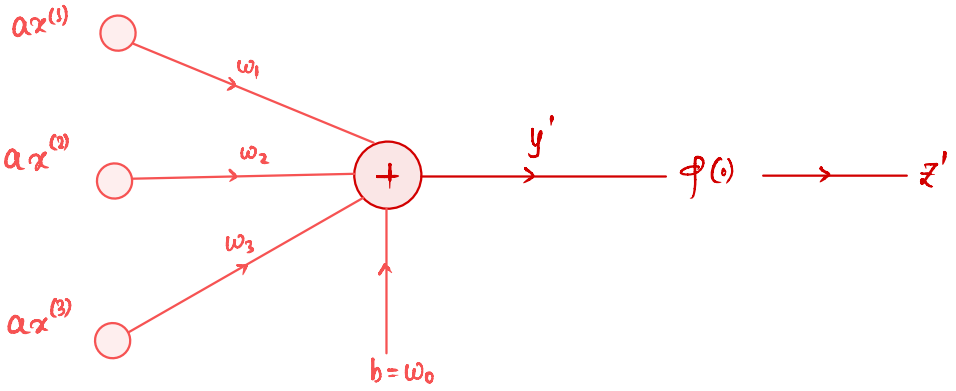



$$y' = w_1 (x_1^{(1)} + x_2^{(1)}) + w_2 (x_1^{(2)} + x_2^{(2)}) + w_3 (x_1^{(3)} + x_2^{(3)}) + w_0$$

$$= w_1 x_1^{(1)} + w_2 x_1^{(2)} + w_3 x_1^{(3)} + w_1 x_2^{(1)} + w_2 x_2^{(2)} + w_3 x_2^{(3)}$$

$$+ w_0 + w_0 - w_0$$

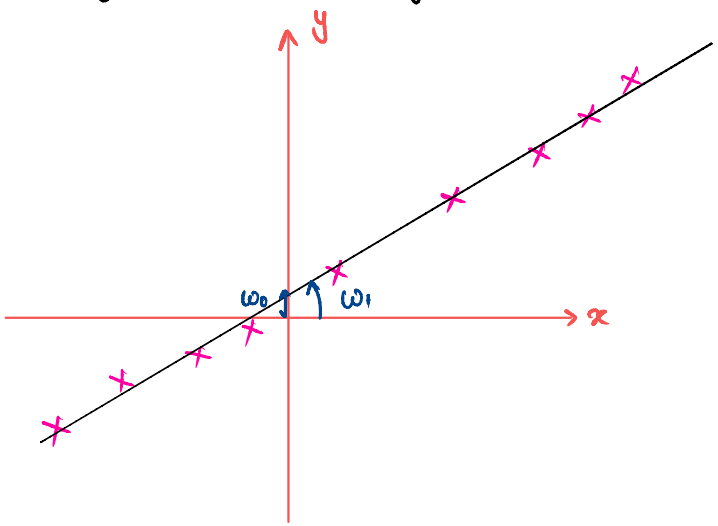
$$= y_1 + y_2 - w_0 \neq y_1 + y_2$$




$$y' = w_1 \cdot ax^{(1)} + w_2 \cdot ax^{(2)} + w_3 \cdot ax^{(3)} + w_0 + aw_0 - aw_0$$

$$= a\left(w_1 x^{(1)} + w_2 x^{(2)} + w_3 x^{(3)} + w_0\right) + (1-a)w_0$$

$$= ay + (1-a)w_0 \neq ay$$




| i | x | y |
|---|---|---|
| 0 | $x_0$ | $y_0$ |
| 1 | $x_1$ | $y_1$ |
| ⋮ | ⋮ | ⋮ |
| N-1 | $x_{N-1}$ | $y_{N-1}$ |

$y = f(x)$

↓

Some function implemented with NN

$f(x) = w_0 + w_1 x = y$

$y = mx + c$

$\| \quad \|$

$w_1 \quad w_0$

↓ ↓

Slope y-axis intercept

Find $w_0, w_1$ such that MSE $(y, f(x))$ is minimum

MSE : Loss

$$\mathcal{L}(\cdot) = J(\cdot) = \text{MSE}(y, \hat{y}) = \frac{1}{N} \sum_{i=0}^{N-1} (y_i - f(x_i))^2$$

$$\{w_0, w_1\}^* = \underset{\{w_0, w_1\}}{\text{argmin}} \; J(\cdot)$$

$$(w_0, w_1)^{(t=0)} \sim \text{Random}, \; t = 1, \; e = \infty$$

while $e > \epsilon$

- $e = 0$
- for $i = 0, 1, \cdots, N-1$
  - $\hat{y}_i = \phi(x_i)$
  - $e = e + (\hat{y}_i - y_i)^2$
- end
- $J^{t}(\cdot) = e$
- $(w_0, w_1)^{(t)} = (w_0, w_1)^{(t-1)} - \eta \frac{\partial}{\partial (w_0, w_1)} J^{(t)}(\cdot)$
  - $\eta$ ↳ Learning Rate
- $t = t + 1$

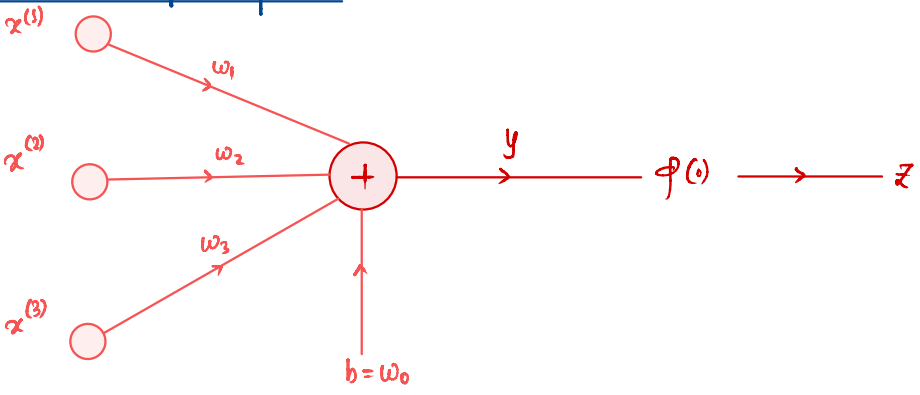
end

$$\hat{y} = w_0 + w_1 x$$

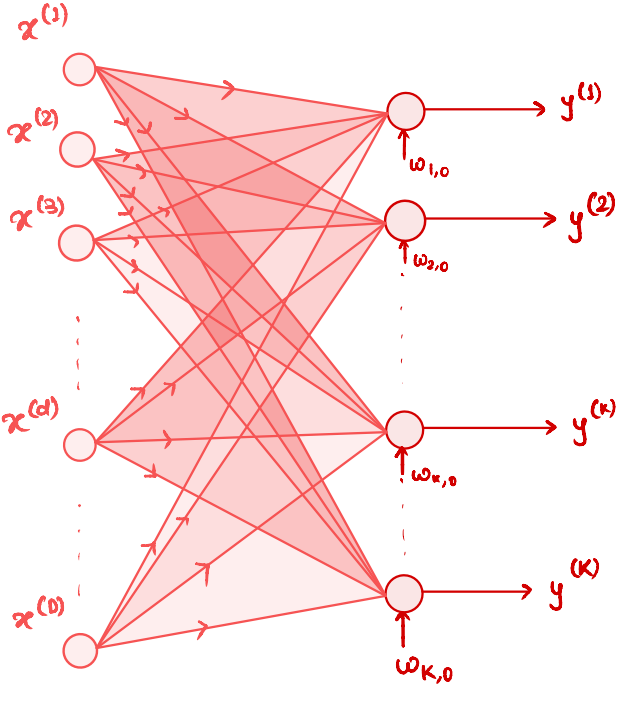
$$\frac{\partial}{\partial w_0} J(\cdot) = \frac{\partial}{\partial \hat{y}} J(\cdot) \cdot \frac{\partial}{\partial w_0} \hat{y} = -\frac{2}{N} \sum_{i=0}^{N-1} (y_i - f(x_i))$$

$$\frac{\partial}{\partial w_1} J(\cdot) = \frac{\partial}{\partial \hat{y}} J(\cdot) \; \frac{\partial}{\partial w_1} \hat{y} = -\frac{2}{N} \sum_{i=0}^{N-1} (y_i - f(x_i)) \, x_i$$

## Rosenblatt's perceptron




## Fully Connected Network




net $(\cdot)$

$$y^{(1)} = w_{1,0} + w_{1,1}x^{(1)} + w_{1,2}x^{(2)} + \cdots + w_{1,d}x^{(d)} + \cdots + w_{1,D}x^{(D)}$$

$$y^{(2)} = w_{2,0} + w_{2,1}x^{(1)} + w_{2,2}x^{(2)} + \cdots + w_{2,d}x^{(d)} + \cdots + w_{2,D}x^{(d)}$$

$$\vdots$$

$$y^{(K)} = w_{K,0} + w_{K,1} x^{(1)} + w_{K,2} x^{(2)} + \cdots + w_{K,d} x^{(d)} + \cdots + w_{K,D} x^{(D)}$$

Output Neuron Index

Output Neuron Index

Input Neuron Index

Input Neuron Index

$$y^{(K)} = w_{K,0} + w_{K,1} x^{(1)} + w_{K,2} x^{(2)} + \cdots w_{K,d} x^{(d)} + \cdots + w_{K,D} x^{(D)}$$

$$\begin{bmatrix} y^{(1)} \\ y^{(2)} \\ \vdots \\ y^{(K)} \\ \vdots \\ y^{(K)} \end{bmatrix} = \begin{bmatrix} w_{1,0} & w_{1,1} & \cdots & w_{1,d} & \cdots & w_{1,D} \\ w_{2,0} & w_{2,1} & \cdots & w_{2,d} & \cdots & w_{2,D} \\ \vdots & & & & & \\ w_{K,0} & w_{K,1} & \cdots & w_{K,d} & \cdots & w_{K,D} \\ \vdots & \vdots & & & & \\ w_{K,0} & w_{K,1} & \cdots & w_{K,d} & \cdots & w_{K,D} \end{bmatrix} \begin{bmatrix} 1 \\ x^{(1)} \\ \vdots \\ x^{(K)} \\ \vdots \\ x^{(K)} \end{bmatrix}$$

$$\underline{Y} = \underline{W}\,\underline{X} + \underline{b}$$

$\underline{Y} \in \mathbb{R}^{K \times J}$ : Output Activation

$\underline{W} \in \mathbb{R}^{K \times D}$ : Weights

$\underline{X} \in \mathbb{R}^{D \times J}$ : Input Activation

$\underline{b} \in \mathbb{R}^{K \times J}$ : Bias

$$\underline{W}^{(t+1)} = \underline{W}^{(t)} - \eta \frac{\partial}{\partial \underline{W}^{(t)}} J^{(t)}(\cdot)$$

$$\Rightarrow \underline{W}^{(t+1)} = \underline{W}^{(t)} - \eta \nabla \underline{W}^{(t)}$$

$$\frac{\partial}{\partial \underline{W}^{(t)}} J^{(t)}(\cdot) = \underbrace{\frac{\partial}{\partial \underline{Y}^{(t)}} J^{(t)}(\cdot)}_{\nabla \text{Loss}(\cdot)} \cdot \underbrace{\frac{\partial}{\partial \underline{W}^{(t)}} \underline{Y}^{(t)}}_{\nabla \text{net}(\cdot)}$$

$\nabla \text{Loss}(\cdot)$: Gradient of loss function

$\nabla \text{net}(\cdot)$: Gradient of the network

$$\underline{Y} = \underline{W}\,\underline{X} + \underline{b}$$

$\underline{W} \in \mathbb{R}^{K \times D}$, $\underline{X} \in \mathbb{R}^{D \times 1}$, $\underline{Y} \in \mathbb{R}^{K \times 1}$

$$\nabla \underline{W} = \frac{\partial}{\partial \underline{W}} J(\cdot) = \nabla \underline{\hat{Y}}\, \underline{X}^T$$

$$\nabla \underline{b} = \frac{\partial}{\partial \underline{b}} J(\cdot) = \nabla \underline{\hat{Y}} \in \mathbb{R}^{K \times 1}$$

$\underline{Q} = \underline{R}\,\underline{S} + \underline{T}$ ($\underline{Q} \leftarrow \underline{Y}$, $\underline{R} \leftarrow \underline{W}$, $\underline{S} \leftarrow \underline{X}$, $\underline{T} \leftarrow \underline{b}$)

$\underline{R} \in \mathbb{R}^{a \times b}$

$\underline{S} \in \mathbb{R}^{b \times c}$

$\underline{T} \in \mathbb{R}^{a \times c}$

$\underline{Q} \in \mathbb{R}^{a \times c}$

$$\frac{\partial}{\partial \underline{R}} J(\cdot) = \nabla \underline{R} = \nabla \underline{Q}\, \underline{S}^T$$

$$\frac{\partial}{\partial \underline{T}} J(\cdot) = \nabla \underline{T} = \nabla \underline{\theta}$$

$$\frac{\partial}{\partial \underline{S}} J(\cdot) = \nabla \underline{S} = \underline{R}^T = \nabla \underline{\theta}$$

$$J(\cdot) = \frac{1}{N} \sum_{i=0}^{N-1} (\underline{y}_i - \underline{\hat{y}}_i)^2$$

$$\frac{\partial}{\partial \underline{\hat{y}}} J(\cdot) = \left[ \frac{\partial}{\partial \hat{y}_1} J(\cdot) \quad \frac{\partial}{\partial \hat{y}_2} J(\cdot) \quad \cdots \quad \frac{\partial}{\partial \hat{y}_k} J(\cdot) \quad \cdots \quad \frac{\partial}{\partial \hat{y}_K} J(\cdot) \right]^T \in \mathbb{R}^{K \times 1}$$

$$= -\frac{2}{N} \sum_{i=0}^{N-1} (y_i^{(k)} - \hat{y}_i^{(k)})$$

Update Rule:

$$\underset{\mathbb{R}^{K \times D}}{\underline{W}^{(t+1)}} = \underset{\mathbb{R}^{K \times D}}{\underline{W}^{(t)}} - \eta \nabla \underset{\mathbb{R}^{K \times D}}{\underline{W}^{(t)}}$$

$$\underset{\mathbb{R}^{K \times 1}}{\underline{b}^{(t+1)}} = \underline{b}^{(t)} - \eta \nabla W^{(t)}$$

$\underline{X} \in \mathbb{R}^{D \times 1}$ : Inputs

$\underline{Y} \in \mathbb{R}^{D \times 1}$ : Outputs

$\underline{W} \in \mathbb{R}^{K \times D}$ : Weights

$\underline{b} \in \mathbb{R}^{K \times 1}$ : Biases

} Parameters / Learnable parameters.

$\underline{t} \in \mathbb{R}^{K \times 1}$ : Target

**Forward:** (Input → Output port)

$$\underline{y} = \underline{W}\,\underline{x} + \underline{b}$$

**Backward:** (Output port → Input)

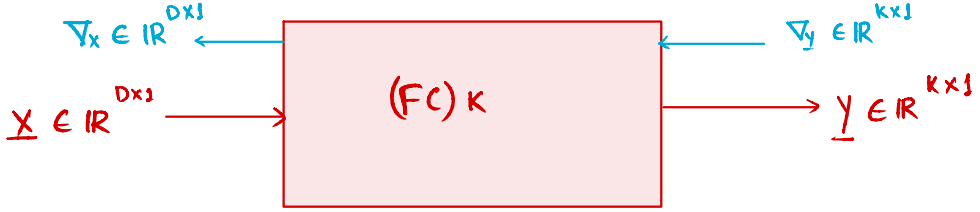
$$\nabla_{\underline{W}} = \nabla_{\underline{y}}\,\underline{x}^{T}$$

$$\nabla_{\underline{b}} = \nabla_{\underline{y}}$$

$$\nabla_{\underline{x}} = \underline{W}^{T}\,\nabla_{\underline{y}}$$

$\underline{W} \in \mathbb{R}^{K \times D}$ $\quad \nabla_{\underline{W}} \in \mathbb{R}^{K \times D}$

$\underline{b} \in \mathbb{R}^{K \times 1}$ $\quad \nabla_{\underline{b}} \in \mathbb{R}^{K \times 1}$




(FC) := Symbolic Notation

Forward

$$\underline{y}_i = \underline{W}\,\underline{x}_i + \underline{b} \quad : \text{Forward : net}(\cdot)$$

$$J(\cdot) = fcn(t_i, t_i) \quad : \text{Forward : Loss}(\cdot)$$

$$\nabla_{\underline{y}} = \frac{\partial}{\partial \underline{y}} J(\cdot) \quad : \text{Backward : Loss}(\cdot)$$

Backward

$$\nabla_W = \frac{\partial}{\partial W} J(\cdot) = \frac{\partial}{\partial W} Y \cdot \frac{\partial}{\partial Y} J(\cdot) = \nabla_y X^T \quad : \text{Backward : Weight}$$

$$W^{\tau+1} = W^{\tau} - \eta \nabla_W(\tau) \quad : \text{Update : weights}$$

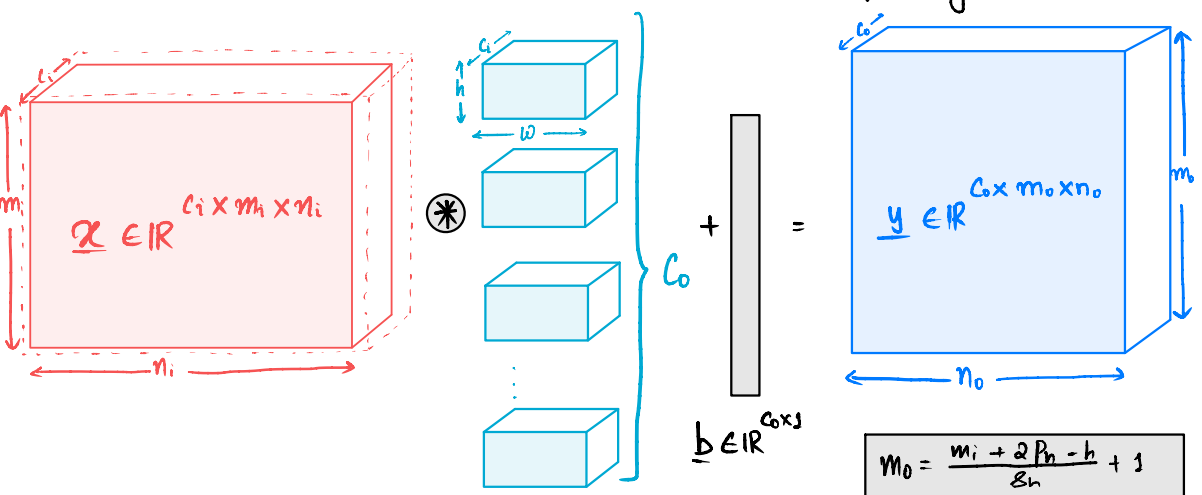
$$\nabla_b = \frac{\partial}{\partial b} J(\cdot) = \frac{\partial}{\partial b} Y \cdot \frac{\partial}{\partial Y} J(\cdot) \quad : \text{Backward : Biases}$$

$$b^{\tau+1} = b^{\tau} - \eta \nabla_b(\tau) \quad : \text{Update : Biases}$$

## Convolution Neuron :-

Stride = $\delta_h, \delta_w$

Padding = $P_h, P_w$




$$m_o = \frac{m_i + 2P_h - h}{\delta_h} + 1$$

$$n_o = \frac{n_i + 2P_w - w}{\delta_w} + 1$$

$$\begin{matrix} x_{11} & x_{12} & x_{13} & \cdots \\ x_{21} & x_{22} & x_{23} & \cdots \\ x_{31} & x_{32} & x_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{matrix} \circledast \begin{matrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{matrix} = \begin{matrix} y_{11} & y_{12} & y_{13} & \cdots \\ y_{21} & y_{22} & y_{23} & \cdots \\ y_{31} & y_{32} & y_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{matrix}$$

180° rotated version of Convolution Kernel

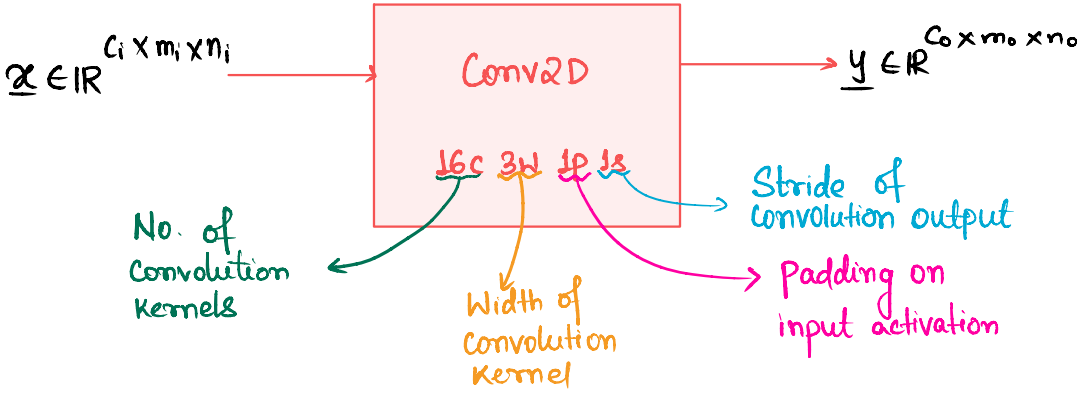
$$y_{11} = x_{11} w_{11} + x_{12} w_{12} + x_{13} w_{13} + x_{21} w_{21} + x_{22} w_{22} + x_{23} w_{23} + x_{31} w_{31} + x_{32} w_{32} + x_{33} w_{33}$$

$$y_{12} = x_{12} w_{12} + x_{13} w_{13} + x_{14} w_{14} + x_{22} w_{22} + x_{23} w_{23} + x_{24} w_{24} + x_{32} w_{32} + x_{33} w_{33} + x_{34} w_{34}$$

Size of Convolution kernel : $w, h = 3, 3$

Padding : $P_w, P_h = 1$

Stride : $S_w, S_h = 1$




$P_w, P_h = 33, 33$

$S_w, S_h = 3, 3$

(Conv2D) : $\mathbb{R}^{1 \times 32 \times 32} \longmapsto \mathbb{R}^{16 \times 32 \times 32}$

$\underline{x} \longrightarrow$ (Conv2D) 16c 5w 1s 0p $\longrightarrow \underline{y}$

$\underline{x} \in \mathbb{R}^{1 \times 32 \times 32}$

$\underline{y} \in \mathbb{R}^{16 \times 32 \times 32}$

If $\underline{y} \in \mathbb{R}^{16 \times 32 \times 32}$, then find when $w \times h = 5 \times 5$,

$C_o = 16$ (Conv 2D): $\mathbb{R}^{1 \times 32 \times 32} \longmapsto \mathbb{R}^{16 \times 32 \times 32}$

$P_w, P_h = 2, 2$

$\delta_w, \delta_h = 1, 1$

| $P_w, P_h$ | 2, 2 | 33, 33 |
|---|---|---|
| $\delta_w, \delta_h$ | 1, 1 | 3, 3 |
| Proportion (%) | 87% | 31% |

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & \cdots \\ x_{21} & x_{22} & x_{23} & \cdots \\ x_{31} & x_{32} & x_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix} \circledast \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \\ w_{31} & w_{32} & w_{33} \end{bmatrix} + b_o = \begin{bmatrix} y_{11} & y_{12} & y_{13} & \cdots \\ y_{21} & y_{22} & y_{23} & \cdots \\ y_{31} & y_{32} & y_{33} & \cdots \\ \vdots & \vdots & \vdots & \end{bmatrix}$$

$$y_{11} = x_{11} w_{11} + x_{12} w_{12} + x_{13} w_{13} + x_{21} w_{21} + x_{22} w_{22} + x_{23} w_{23} + x_{31} w_{31} + x_{32} w_{32} + x_{33} + w_{33} + b_o$$

$$\begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{21} & x_{22} & x_{23} & x_{31} & x_{32} & x_{33} \\ x_{12} & x_{13} & x_{14} & x_{22} & x_{23} & x_{24} & x_{32} & x_{33} & x_{34} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ x_{21} & x_{22} & x_{23} & x_{31} & x_{32} & x_{33} & x_{41} & x_{42} & x_{43} \\ x_{22} & x_{23} & x_{24} & x_{32} & x_{33} & x_{34} & x_{42} & x_{43} & x_{44} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} w_{11} \\ w_{12} \\ w_{13} \\ w_{21} \\ w_{22} \\ w_{23} \\ w_{31} \\ w_{32} \\ w_{33} \end{bmatrix} + b_o = \begin{bmatrix} y_{11} \\ y_{12} \\ y_{13} \\ \vdots \\ y_{21} \\ y_{22} \\ \vdots \end{bmatrix}$$

$\underline{B} \in \mathbb{R}^{1 \times C_o}$

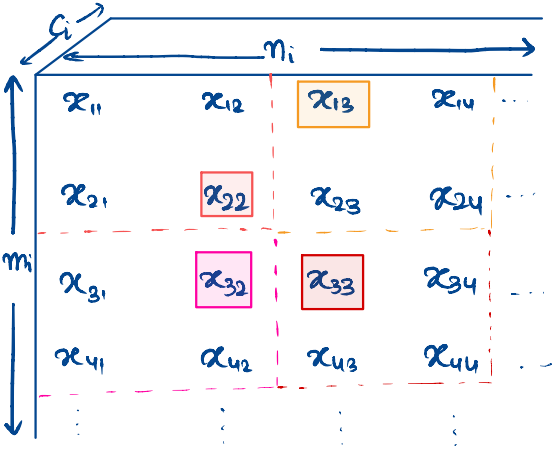
$\underline{X} \in \mathbb{R}^{m_0 n_0 \times C_i w_h}$ $\underline{W} \in \mathbb{R}^{C_i w_h \times C_o}$ $\underline{Y} \in \mathbb{R}^{m_0 n_0 \times C_o}$

$\nabla_w = ?$

$\nabla_b = ?$

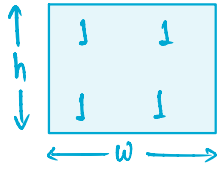
## Pooling :-

### Max - Pooling -

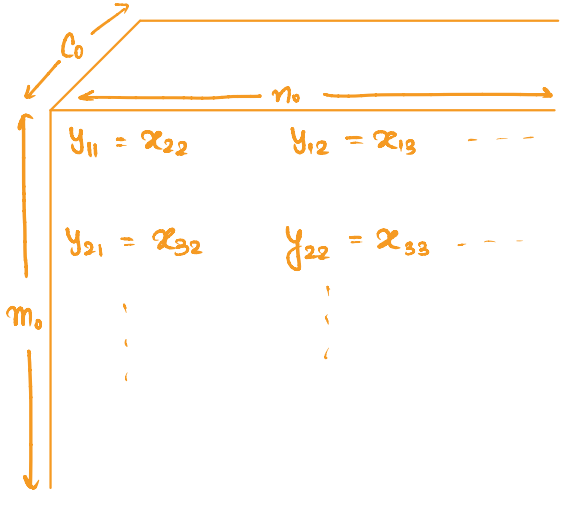



**Pooling Kernel**

Kernel Size : $w, h = 2, 2$




Stride : $s_w, s_h = 2, 2$

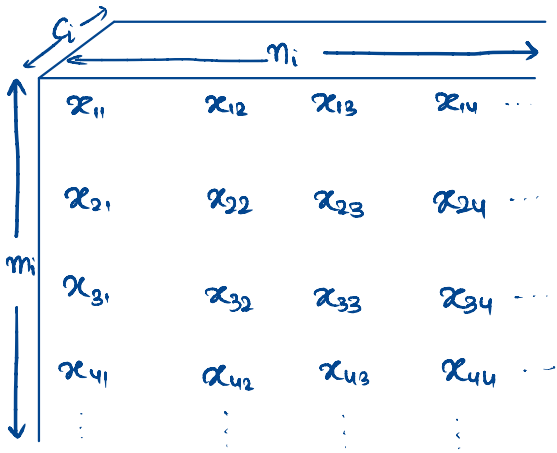



$$m_o = \frac{m_i - h}{s_h} + 1$$

$$n_o = \frac{n_i - w}{s_w} + 1$$

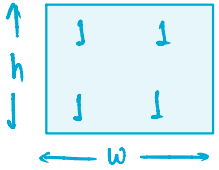
$$C_o = C_i$$

# Avg - pool 2D -




**Pooling Kernel**

Kernel Size : $w, h = 2, 2$




$\delta_w, \delta_h = 2, 2$

$$y_{11} = \frac{1}{4}(x_{11} + x_{12} + x_{21} + x_{22})$$

$$y_{12} = \frac{1}{4}(x_{13} + x_{14} + x_{23} + x_{24})$$

$$y_{21} = \frac{1}{4}(x_{31} + x_{32} + x_{41} + x_{42})$$

$\delta_w, \delta_h = 1, 1$

$$y_{11} = \frac{1}{4}(x_{11} + x_{12} + x_{21} + x_{22})$$

$$y_{12} = \frac{1}{4}(x_{12} + x_{13} + x_{22} + x_{23})$$

$$y_{22} = \frac{1}{4}(x_{22} + x_{23} + x_{32} + x_{33})$$

$$\underline{x} \in \mathbb{R}^{C_i \times m_i \times n_i}$$

$$\underline{y} \in \mathbb{R}^{C_o \times m_o \times n_o}$$

$$C_o = C_i$$

$$m_o = \frac{m_i - h}{\delta_h} + 1$$

$$n_o = \frac{n_i - w}{\delta_w} + 1$$

$$\nabla \underline{x} \in \mathbb{R}^{C_i \times m_i \times n_i}$$

$$\nabla y \in \mathbb{R}^{C_o \times m_o \times n_o}$$

LeNet-5 → 5 layers with trainable parameters

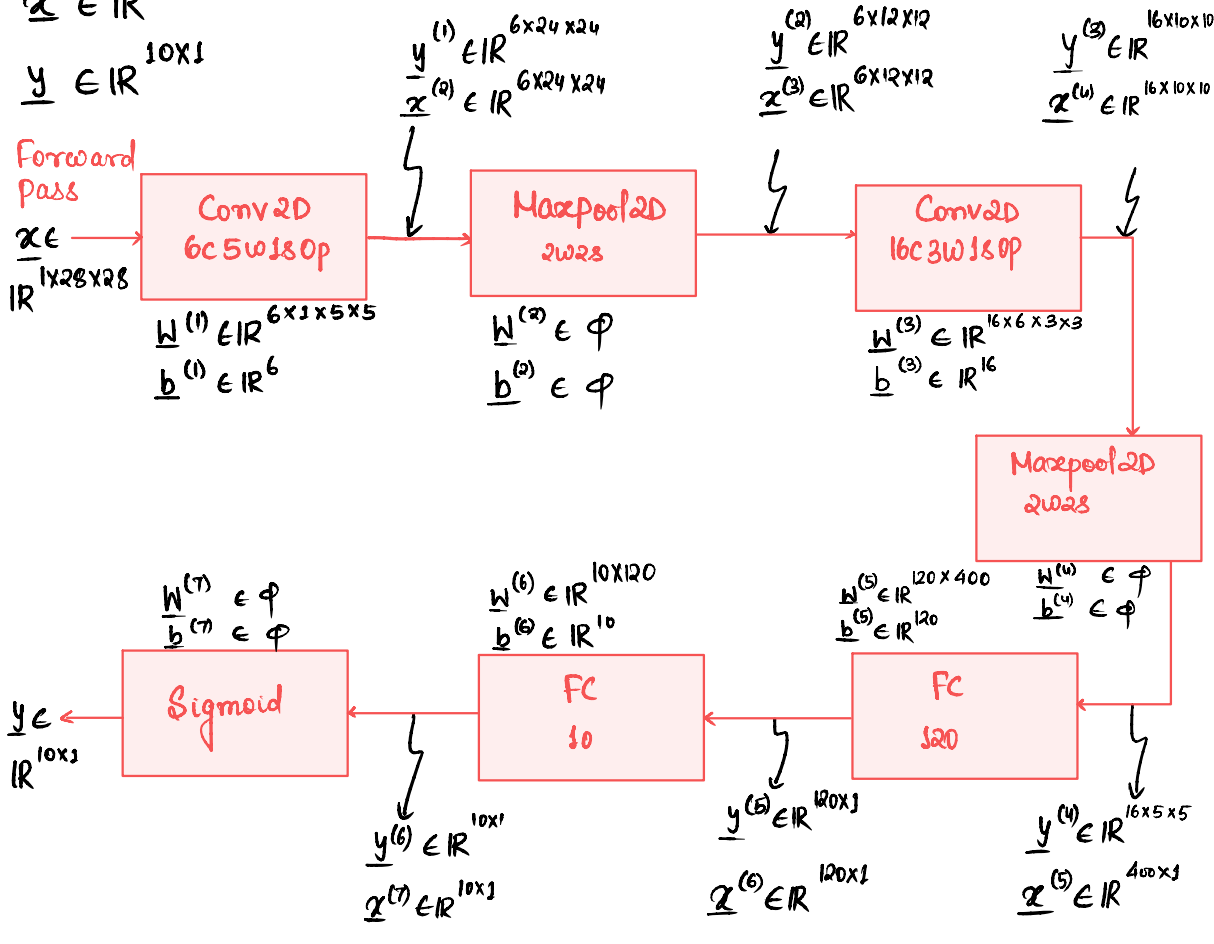
Yann LeCunn — NN

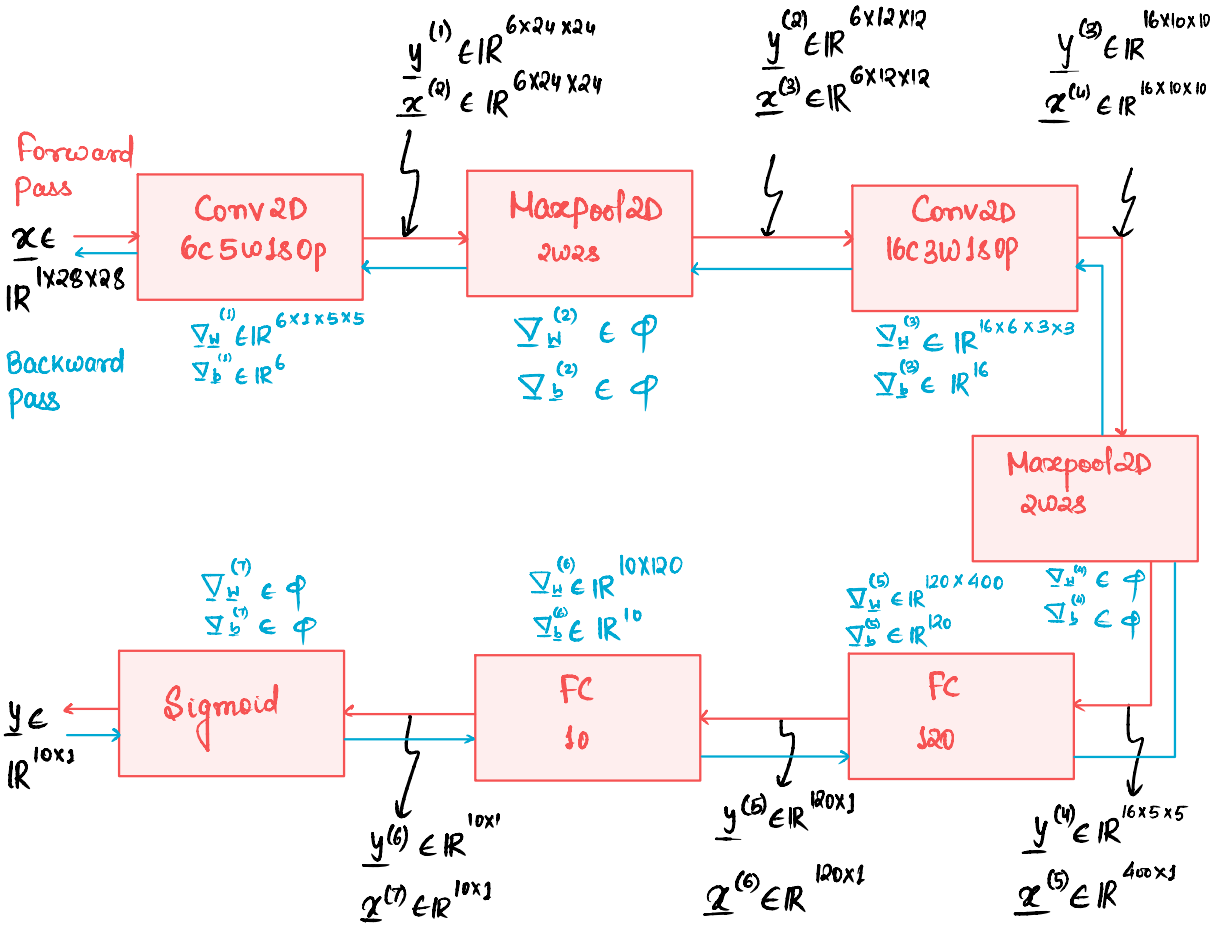
net (·) ⟼ (1: Conv2D) 6c 5w 1s 0p ⟶ (2: maxpool2D) 2w 2s
⟶ (3: Conv2D) 16c 3w 1s 0p ⟶ (4: maxpool2D) 2w 2s
⟶ (5: FC) 120 ⟶ (6: FC) 10 ⟶ (7: Sigmoid)

$\underline{y} = net(\underline{x})$

$\underline{x} \in \mathbb{R}^{1\times28\times28}$

$\underline{y} \in \mathbb{R}^{10\times1}$






| Layer Number: | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Total |
|---|---|---|---|---|---|---|---|---|
| No. of parameters: | 156 | 0 | 80 | 0 | 48,120 | 1210 | 0 | 50366 |

$(|W^{(1)}| + |b^{(1)}|)$

$|\circ|$ → Set cardinality operator

sizeof(net(·)) = Total number of parameters · precision

Float 32 = 50366 × 4 Bytes
≈ 200 KBytes

Double / Float 64 = 50366 × 8 Bytes
≈ 400 KBytes

No. of multiplications = $n_o m_o C_o C_i wh$

No. of additions = $C_o m_o n_o \left( (C_i wh - 1) + 1 \right)$
= $n_o m_o C_o C_i wh$

No. of operations = $2 n_o m_o C_o C_i wh$

| Operations | 172800 | 20736 | 172,800 | 25600 | 9600 | 2400 | 90 | 490426 |
|---|---|---|---|---|---|---|---|---|

≈ 0.5 MegaFlops

No of operations per second = $\frac{10 \times 10^3 \text{ MegaFlops } s^{-1}}{0.5 \text{ MegaFlops}}$

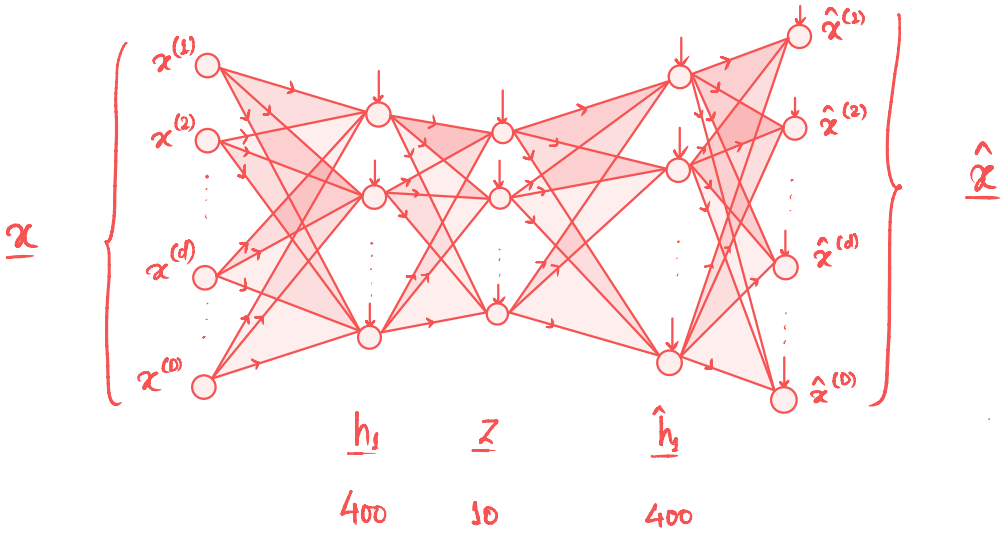
= $20{,}000\ s^{-1}$

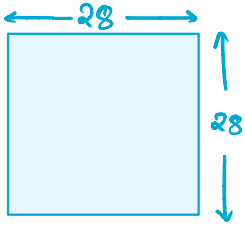
## Generative Models :-

$p(\Omega = \omega | \underline{x})$ :: Discriminative Model

$p(\underline{y}, \underline{x})$ :: Generative Model

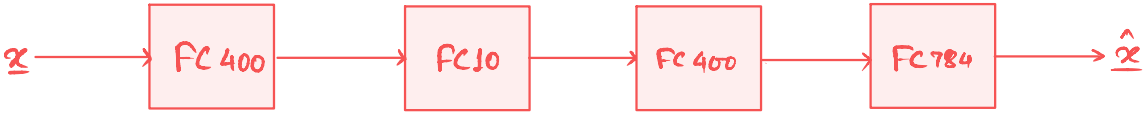
# AutoEncoder-




$\mathcal{I} \in \mathbb{R}^{1 \times 28 \times 28}$

$\underline{x} \in \mathbb{R}^{784 \times 1}$




$net_{AE}(\cdot) = net_{Dec}(net_{Enc}(\cdot))$

$\hat{x}_{AE} = net_{AE}(x)$

$$\underline{z} = net_{Enc}(\underline{x})$$

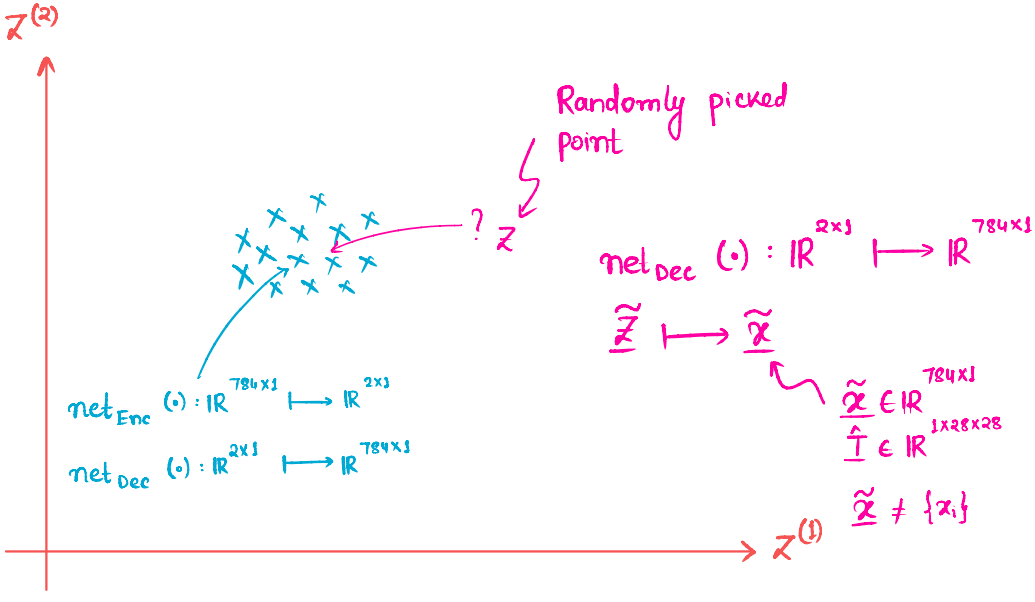
$$\hat{\underline{x}} = net_{Dec}(\underline{z})$$

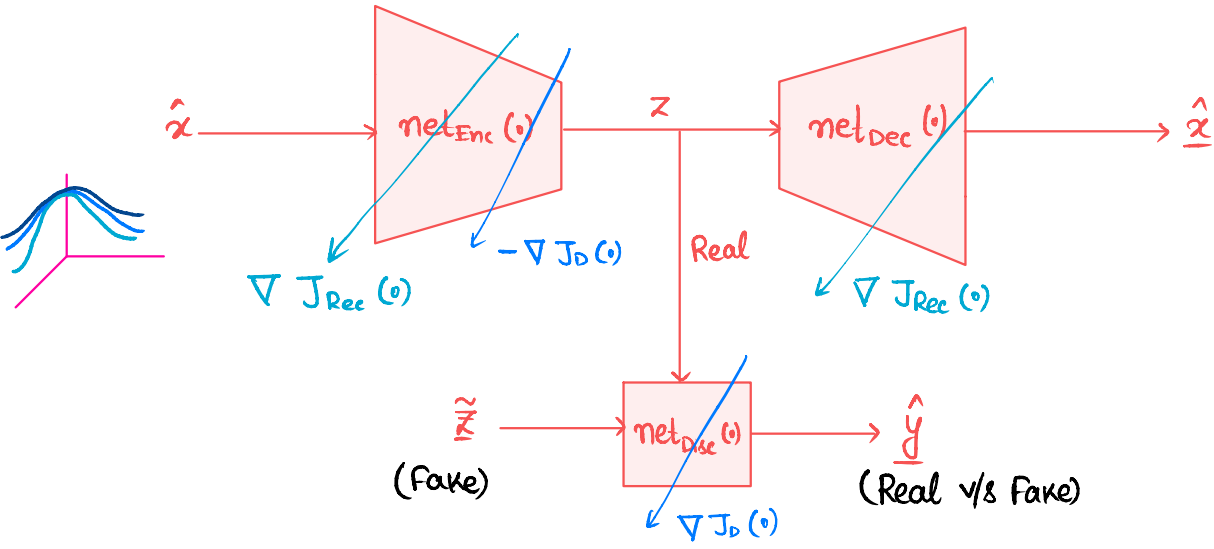
$$J(\cdot) = \frac{1}{N} \sum_{i=0}^{N-1} \| \underline{x}_i - \hat{\underline{x}}_i \|_2^2$$

$$\{\underline{W}^*, \underline{b}^*\} = \underset{\{\underline{W}, b\}}{\operatorname{argmin}} J(\cdot)$$

$net_{Enc}(\cdot) \longmapsto$ (1: FC) 400 $\longrightarrow$ (2: FC) 10 $\Rightarrow \mathbb{R}^{784\times 1} \longmapsto \mathbb{R}^{10\times 1}$

$net_{Dec}(\cdot) \longmapsto$ (1: FC) 400 $\longrightarrow$ (2: FC) 784 $\Rightarrow \mathbb{R}^{10\times 1} \longmapsto \mathbb{R}^{784\times 1}$






$$J_{Rec}(\cdot) = MSE(x_i, \hat{x}_i)$$

$$\hat{y} \in \{Real, Fake\}$$

$$J_D(\cdot) = BCE(y, \hat{y})$$

- $net_{Enc}(\cdot)$, $net_{Dec}(\cdot)$, $net_{Disc}(\cdot)$ → Randomly initialized

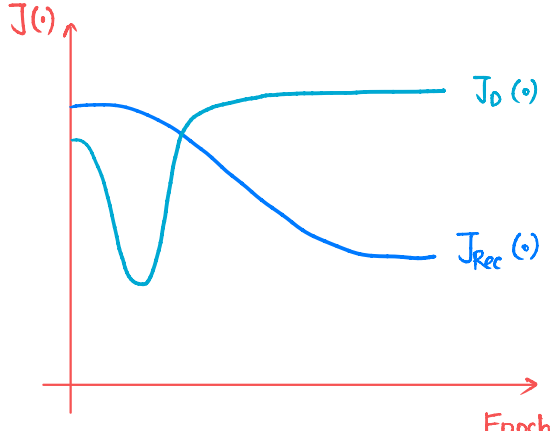
for $i = 0, 1, \ldots, N-1$

$$z_i = net_{Enc}(x_i)$$

$$\hat{x}_i = net_{Dec}(z_i)$$

$$e_i = \| \hat{x}_i - x_i \|_2^2$$

$$s_i = Random\{Fake, Real\}$$



$$\underline{h}_i = \underline{x}_i \odot \underline{s}_i + \tilde{\underline{x}}_i \cdot (1 - \underline{s}_i) \qquad \{\underline{W}_{Enc}, \underline{b}_{Enc}\} \longleftrightarrow net_{Enc}(\cdot)$$

$$\hat{\underline{y}}_i = net_{Dise}(\underline{h}_i) \qquad \{\underline{W}_{Dec}, \underline{b}_{Dec}\} \longleftrightarrow net_{Dec}(\cdot)$$

$$d_i = BCE(\hat{\underline{y}}_i, \underline{s}_i) \qquad \{\underline{W}_{Dise}, \underline{b}_{Dise}\} \longleftrightarrow net_{Dise}(\cdot)$$

$$J_{Rec}(\cdot) = J_{Rec}(\cdot) + e_i$$

$$J_D(\cdot) = J_D(\cdot) + d_i$$

end

$$\underline{W}_{Dec}^{(t+1)} = \underline{W}_{Dec}^{(t)} - \eta \frac{\partial}{\partial \underline{W}_{Dec}^{(t)}} J_{Rec}^{(t)}(\cdot)$$

$$\underline{b}_{Dec}^{(t+1)} = \underline{b}_{Dec}^{(t)} - \eta \frac{\partial}{\partial \underline{b}_{Dec}^{(t)}} J_{Rec}^{(t)}(\cdot)$$

$$\underline{W}_{Enc}^{(t+1)} = \underline{W}_{Enc}^{(t)} - \eta \frac{\partial}{\partial \underline{W}_{Enc}^{(t)}} J_{Rec}^{(t)}(\cdot) + \eta_2 \frac{\partial}{\partial \underline{W}_{Enc}} J_D^{(t)}(\cdot)$$

$$\underline{b}_{Enc}^{(t+1)} = \underline{b}_{Enc}^{(t)} - \eta \frac{\partial}{\partial \underline{b}_{Enc}^{(t)}} J_{Rec}^{(t)}(\cdot) + \eta_2 \frac{\partial}{\partial \underline{b}_{Enc}} J_D^{(t)}(\cdot)$$

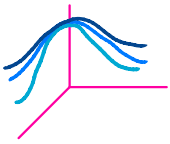
$$\underline{W}_{Dise}^{(t+1)} = \underline{W}_{Dise}^{(t)} - \eta \frac{\partial}{\partial \underline{W}_{Dise}^{(t)}} J_D^{(t)}(\cdot)$$

$$\underline{b}_{Disc}^{(t+1)} = \underline{b}_{Dise}^{(t)} - \eta \frac{\partial}{\partial \underline{b}_{Dise}^{(t)}} J_D^{(t)}(\cdot)$$

$\nabla J_{Rec}(\cdot)$

$\nabla J_D(\cdot)$

# Adversarial Autoencoder (AAE)-



$\underline{x}$ → $net_{Enc}(\cdot)$ → $net_{Dec}(\cdot)$ → $\underline{\hat{x}}$

$\nabla J_{Rec}(\cdot)$

$-\nabla J_D(\cdot)$

$\nabla J_{Rec}(\cdot)$

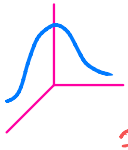
$J_{Rec}(\cdot) = \|\hat{x} - x\|_2^2$

$\underline{\tilde{z}}$ → $net_{Dise}(\cdot)$ → $\underline{y} \in \{Real, Fake\}$

$\nabla J_D(\cdot)$

$J_D(\cdot) = BCE(\underline{y}, \underline{s})$

# Generative Adversial Network-



Multi-Class Object classifier (Non-Trainable)

$\underline{\tilde{z}}$ → $net_G(\cdot)$ → $\underline{\tilde{x}}$ → $net_C(\cdot)$ → $\underline{\tilde{\Omega}}$

$\nabla J_{CLS}(\cdot)$

$-\nabla J_D(\cdot)$

$J_{CLS}(\cdot) = BCE(\underline{\tilde{\Omega}}, \underline{\Omega})$

Fake

$\{x\}$ → $\underline{x}$ → $net_D(\cdot)$ → $y$

$J_D(\cdot)$

Database of Real samples

$J_D(\cdot) = BCE(\underline{y}, \underline{s})$

## AAE

$net_{Enc}(\cdot) : \mathbb{R}^{C \times N \times N} \longmapsto \mathbb{R}^{q}$

$net_{Dec}(\cdot) : \mathbb{R}^{q} \longmapsto \mathbb{R}^{C \times N \times N}$

$net_{Disc}(\cdot) : \mathbb{R}^{q} \longmapsto \mathbb{R}^{2}$

## GAN

$net_{G}(\cdot) : \mathbb{R}^{q} \longmapsto \mathbb{R}^{C \times N \times N}$

$net_{D}(\cdot) : \mathbb{R}^{C \times N \times N} \longmapsto \mathbb{R}^{2}$

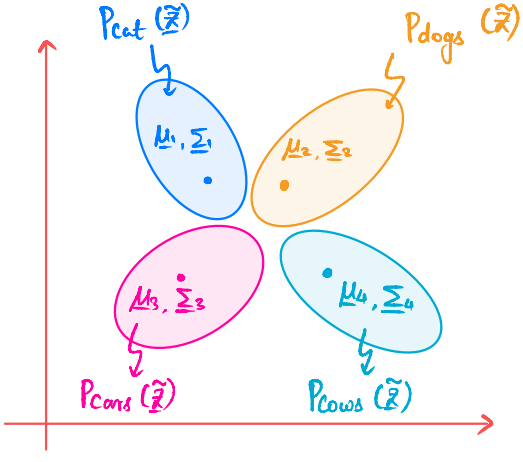
$net_{c}(\cdot) : \mathbb{R}^{C \times N \times N} \longmapsto \mathbb{R}^{k}$

## Conditional Generation :-

$P_{cats}(\underline{\tilde{z}}) \neq P_{dogs}(\underline{\tilde{z}}) \neq P_{cows}(\underline{\tilde{z}}) \neq P_{cars}(\underline{\tilde{z}})$

$q = 2$

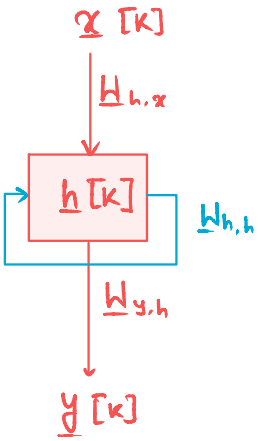
$P(\underline{\tilde{z}}) \sim \mathcal{N}(\mu, \Sigma)$




$\underline{\mu_1}, \underline{\mu_2}, \underline{\mu_3}, \underline{\mu_4} \in \mathbb{R}^{q}$

$\underline{\Sigma_1}, \underline{\Sigma_2}, \underline{\Sigma_3}, \underline{\Sigma_4} \in \mathbb{R}^{q \times q}$

# Recurrent Neural Networks (RNN):-




$$\underline{y}[k] = \underline{W}_{y,h}\ \underline{h}[k] + \underline{b}_y$$

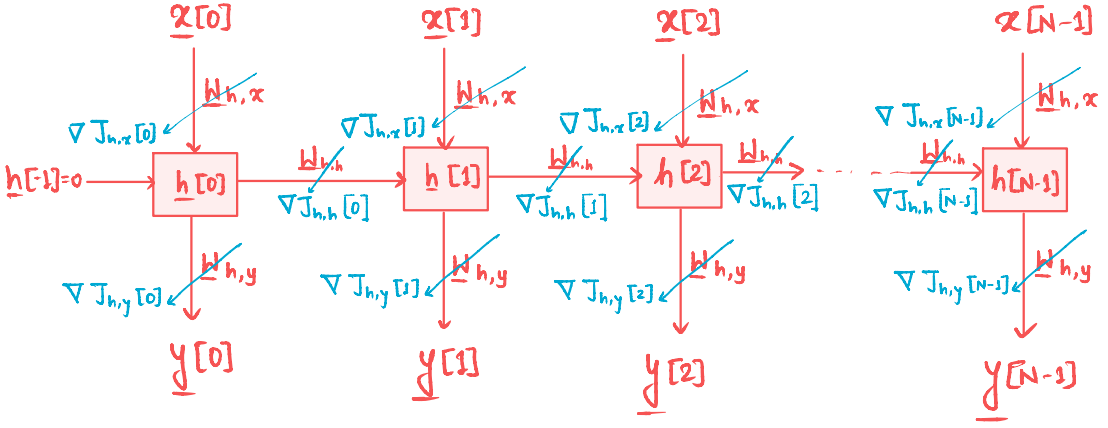
$$\underline{h}[k] = \underline{W}_{h,x}\ \underline{x}[k] + \underline{b}_{h,x} + \underline{W}_{h,h}\ \underline{h}[k-1]$$

---

Many to Many $\Rightarrow \mathbb{R}^{D\times N} \longmapsto \mathbb{R}^{K\times N}$

Many to One $\Rightarrow \mathbb{R}^{D\times N} \longmapsto \mathbb{R}^{K\times 1} \longrightarrow \underline{y}[N-1]$

One to Many $\Rightarrow \mathbb{R}^{D\times 1} \longmapsto \mathbb{R}^{K\times N} \longrightarrow \underline{x}[0]$




$$\underline{X} = \left[\underline{x}[0],\ \underline{x}[1],\ \underline{x}[2],\ \cdots,\ \underline{x}[n],\ \cdots,\ \underline{x}[N-1]\right]$$

Say, $\underline{x}[n] \in \mathbb{R}^D$, then $\underline{X} \in \mathbb{R}^{D\times N}$

$$\underline{Y} = \left[\underline{y}[0],\ \underline{y}[1],\ \underline{y}[2],\ \cdots,\ \underline{y}[n],\ \cdots,\ \underline{y}[N-1]\right]$$

Say, $\underline{y}[n] \in \mathbb{R}^k$, then $\underline{Y} \in \mathbb{R}^{K\times N}$