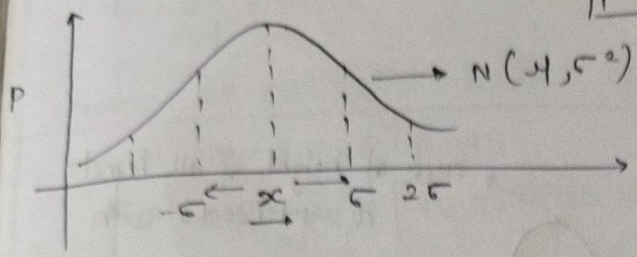
Gaussian / Normal distribution

pdf



$$p(x=x) = \frac{1}{\sqrt{2\pi}\cdot\sigma}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right)$$

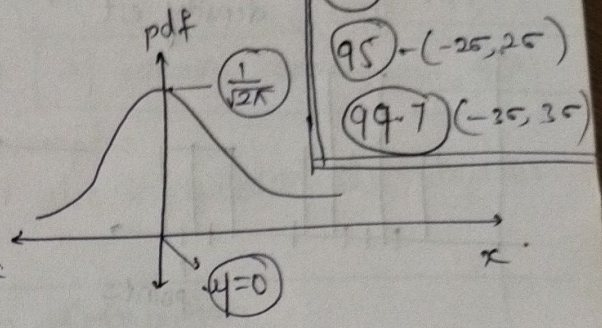
(1)

68% of pts lie in (-σ, σ)

(68) - (-σ, σ)

(95) - (-2σ, 2σ)

(99.7) - (-3σ, 3σ)



let $\mu = 0, \sigma = 1$

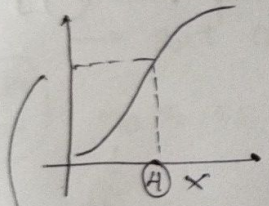
$$p(x) = \frac{1}{\sqrt{2\pi}}\exp\left\{-\frac{1}{2}\cdot x^2\right\}$$

$$p(x) \propto \exp^{-x^2} \propto \frac{1}{e^{x^2}}$$

props ① symetric

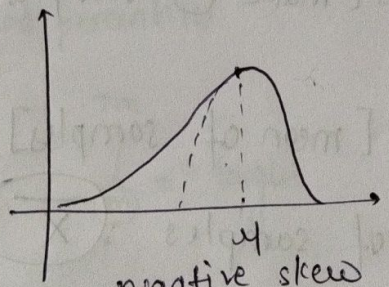
② as x moves away, pdf falls e^{x^2} → square exponential

CDF

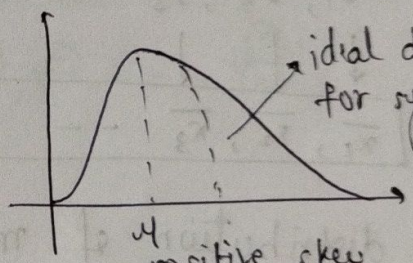


$p(x \le A)$

Skewness



negative skew

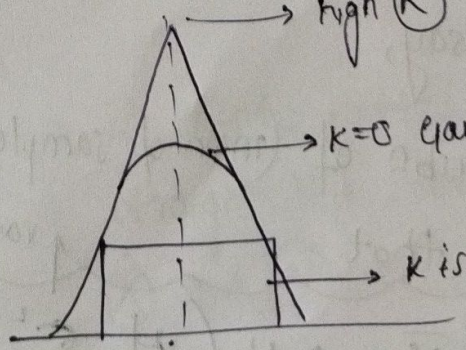


positive skew

Kurtosis

→ How sharp / peaked your distribution is

→ Kurt = 0 for gaussian distrb^n

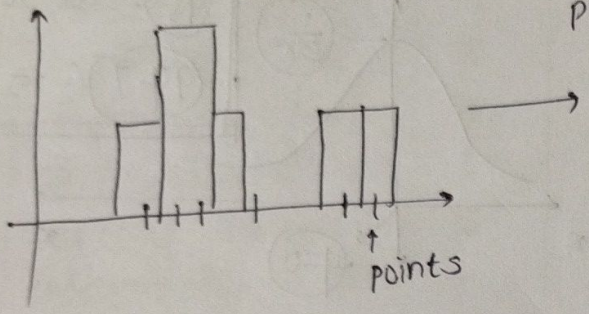


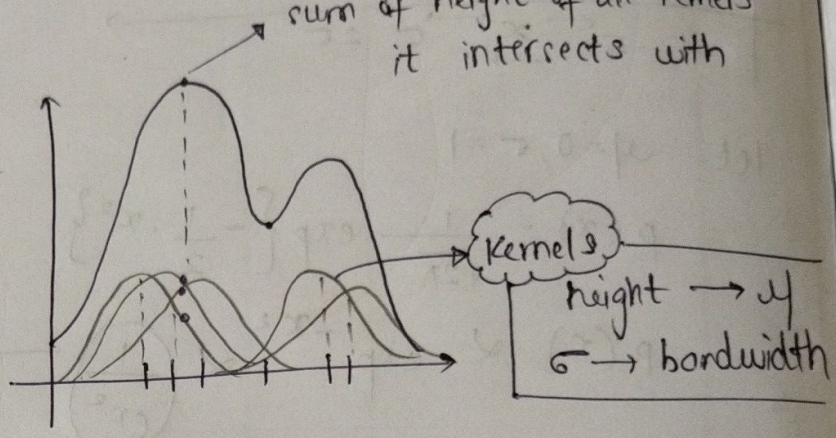
② standard normal variate

$z = \frac{x_i - \mu}{\sigma}$ convert to gausian distribution with $\mu = 0, \sigma = 1$

How pdf?

Histogram $\xrightarrow[\text{density est}]{\text{kernel}}$ pdf





CENTRAL LIMIT THEOREM

* population distrⁿ : $(X) : (\mu, \sigma^2)$ ← can be any type of distribution

* $s_1, s_2, s_3 \; -- \; s_m$ [make (m) samples each of size (n)]

$\bar{x_1}, \bar{x_2}, \bar{x_3} \; -- \; \bar{x_m}$ [mean of samples]

→ distribution of means of samples : $(\bar{X})$

∴ using Central limit theorem,

(i) if X has finite μ and σ^2

we can say,

pop → $x(\mu, \sigma)$

after CLT

sample → $\bar{X}\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

→ distrib⁼ of (mean of samples) follows gausian distribution

such that

$$\bar{X} \rightarrow N\left(\mu, \frac{\sigma^2}{n}\right)$$

(variance)

$n \geq 30$ then it follows gausion

use case

(X) population size → 3 million → may or may not be a (3) gausion distribution
↳ $\bar{y}$ & σ large

asume we plot distribution of heights

let m = 10000, n = 40

→ with just 10,000 × 40 → 40K samples | we can find original (u)

→ $(\bar{y}, \sigma)$ → $(u, \frac{\sigma^2}{2})$ | u remains same | 3mill, 40K

population

distrn of samples
↳ also in gausian distribution

QQ plot Quartile Quartile plot

↳ to check if data follows gaussian distribution

steps

① X: data which is to be checked

(X) $x_1, x_2 \dots x_n$

→ sort and find percentiles

(after sort) $x_{10}\ x_2\ x_3\ x_5 \dots x_m$

(m) m^{th} largest

$x^{(1)}, x^{(2)} \dots x^{(m)}$

↓ first percentile

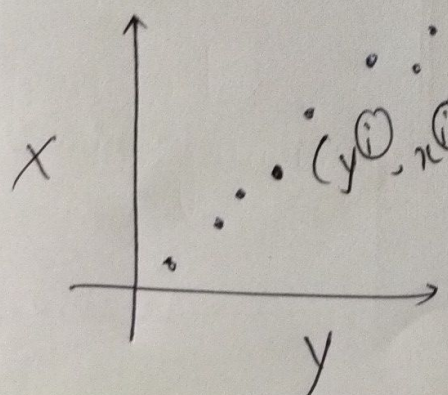
② Y: data which follows gausian distrn

$y: y_1\ y_2 \dots y_n$

→ sort and find percentiles

$y^{(1)}, y^{(2)} \dots y^{(n)}$

③ plot graph of $y^{(i)}, x^{(i)}$



→ if points plotted form a straight line, then

(X) and (Y) are similar

∴ (X) follows gausian distn.

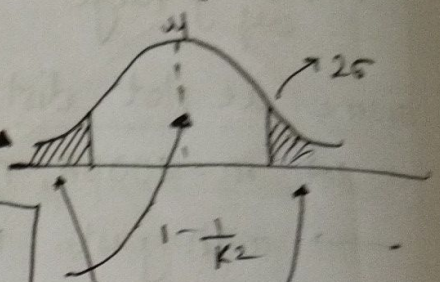
Chebyshevs Inequality

X → dount follow gaussian distribution with μ, σ

$$P(|X-\mu| \geq K\sigma) \leq \frac{1}{K^2}$$

$$P(\mu - K\sigma < X < \mu + K\sigma) > 1 - \frac{1}{K^2}$$

↳ 1, 2 ... n



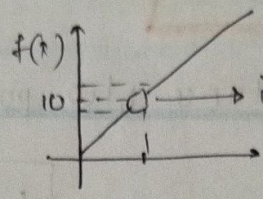
$P(|X-\mu| \geq K\sigma) \rightarrow X \geq \mu + K\sigma$

$X \leq \mu - K\sigma$

Id K=2

Limits

let $f(x) = \begin{cases} 0, x=1 \\ x \cdot 10, x \neq 1 \end{cases}$

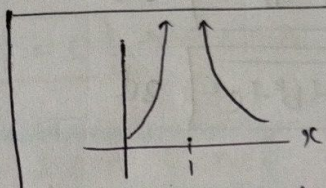


→ indicates $f(x)=0$ / drop

$\therefore$ $\lim_{x \to 1} f(x) = 10$ doesnt care abt $f(x)$ instead what values as you approach $x \to 1$

defn

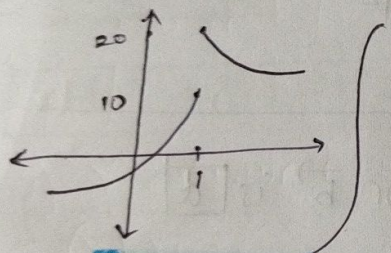
$\lim_{x \to a} f(x) = L$; as x gets closer to a, value of $f(x)$ gets closer to L



$\lim_{x \to 1} = \infty$ / does not exist

left / right limits

$\lim_{x \to 1} f(x)$ is not defined



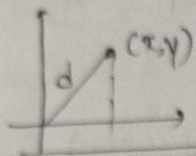
left limit ≠ right limit

$\lim_{x \to 1^-} f(x) = 10$ ← $x \to 1$ from left to right

$\lim_{x \to 1^+} f(x) = 20$ ← $x \to 1$ from right to left

LINEAR ALGEBRA

* dist of pnt from origin

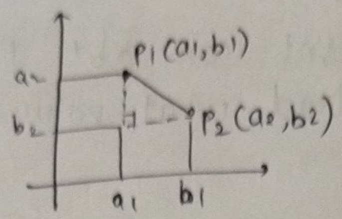


$d = \sqrt{a^2 + b^2}$ 2d

$d = \sqrt{a^2 + b^2 + c^2}$ 3d

* vectors are column vectors if not specified

distance b/w points



$d = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$

n-dimensions → $d = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$

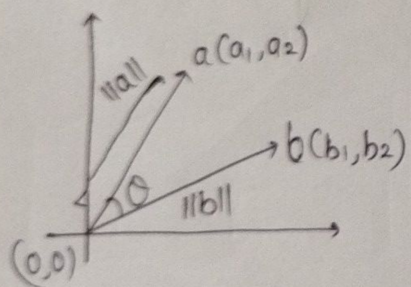
→ $\vec{a} \cdot \vec{b} = \sum_{i=1}^{n} a_i b_i = \|a\|\|b\| \cos\theta$

↳ n-dimensional vector

→ $\vec{a} \cdot \vec{a} = \|a\|^2$

dot product

→ $\vec{a} \cdot \vec{b} = a_1 b_1 + a_2 b_2 \ldots$



$a \cdot b = a_1 b_1 + a_2 b_2$

$= \|a\| \|b\| \cos\theta$

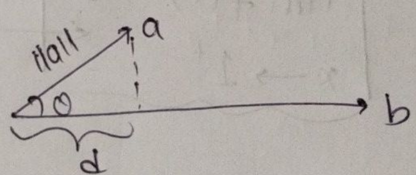
$a \cdot b = \|a\|\|b\| \cos\theta$

$\theta = \cos^{-1}\left(\frac{a \cdot b}{\|a\|\|b\|}\right)$

if $a \cdot b = 0 \Rightarrow \vec{a} \perp \vec{b}$

Projection

→ projection of $\vec{a}$ on $\vec{b}$ is d



unit vector $\hat{a}$

1. same dir as $\vec{a}$
2. $\|a\| = 1$

* proj of a on b → $\frac{d}{\|a\|} = \cos\theta$

$d = \|a\| \cos\theta$ → ①

→ also, $a \cdot b = \sum a_i b_i = \|a\|\|b\| \cos\theta$

$d = \frac{a \cdot b}{\|b\|}$ → same as ①

↓

$\frac{a \cdot b}{\|b\|} \rightarrow \frac{|a| \cancel{|b|} \cos\theta}{\cancel{|b|}}$

↓

$|a| \cos\theta$

$\|\vec{x}\|_2 = \sqrt{\sum_{x_i \in x} x_i^2}$

Equation of line, plane, hyperplane

① Line – 2d

$y = mx + c$

$ax + by + c = 0$

$y = \left(\frac{-c}{b}\right) + \left(\frac{-a}{b}\right)x$ (c, m)

→ 2d : $ax_1 + bx_2 + c = 0$

$\Rightarrow \boxed{w_1x_1 + w_2x_2 + w_0 = 0}$

② plane – 3d

$\boxed{w_1x_1 + w_2x_2 + w_3x_3 + w_0 = 0}$

③ hyperplane – nd

dot product

→ $w_0 + \boxed{\sum_{i=1}^{n} w_i x_i} \Rightarrow w \cdot x$

$$[w_1, w_2 \dots w_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$$

$w_0 + w^T \cdot x = 0$

$\boxed{w_0 + w^T x = 0}$ → Equation of plane (π)

$w \rightarrow 1 \times n$

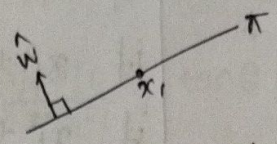
$x \rightarrow n \times 1$

if plane passes origin, $w_0 = 0$

$\boxed{w^T x = 0}$

what is $\hat{w}$

$w = \begin{bmatrix} w_1 \\ \vdots \\ w_n \end{bmatrix}$ $x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$



* w is unit vector ⊥ to π

$w \cdot x = w^T x = \|w\| \|x\| \cos\theta$

as $\theta = 90$, $\cos\theta = 0$

$\therefore \boxed{w^T \cdot x_i = 0} \quad \forall x_i \in \pi$ if $w \perp \pi$

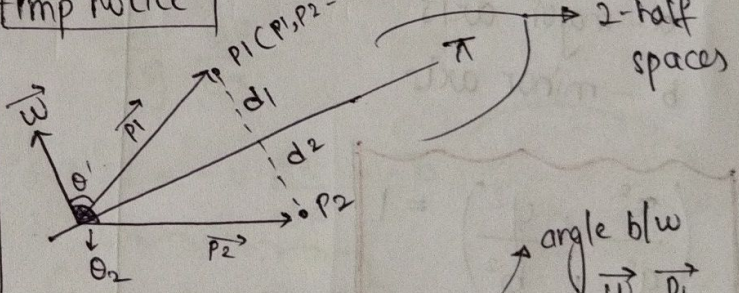
dist of pnt from plane

point $P(p_1, p_2 \dots p_n)$

dot product

$$\boxed{d = \frac{w^T x}{\|w\|} = \frac{w^T x p}{\|w\|} = \frac{w \cdot p}{\|w\|}}$$

Imp notice



$d = \frac{w \cdot p}{\|w\|} \rightarrow \frac{\|w\| \|p\| \cos\theta}{\|w\|}$

$\cos(\theta < 90)$ +ve
$\cos 90 = 0$
$\cos(90+) \rightarrow -ve$

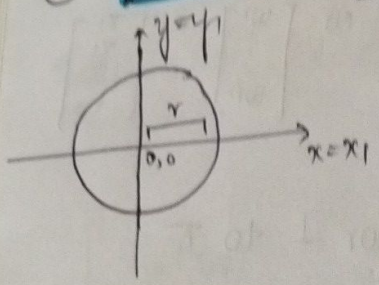
$\theta_1 = <90 \; [90 - \theta_{P_1}] \rightarrow \cos\theta \gtrless 0$ (+ve)

$\theta_2 = >90 \; [90 + \theta_{P_2}] \rightarrow \cos\theta$ (-ve)

$\therefore \boxed{d_1 = +ve \mid d_2 = -ve}$

∴ pnts in same dir as $\vec{w}$ then (d = +ve) else (d = -ve)

⑧ **circle** → consider point $P(x_1, x_2)$



if $x_1^2 + x_2^2 < r^2$ → lies inside circle
if $x_1^2 + x_2^2 > r^2$ → outside circle
if $x_1^2 + x_2^2 = r^2$ → on the circle

$x^2 + y^2 = r^2$ origin $(0,0)$

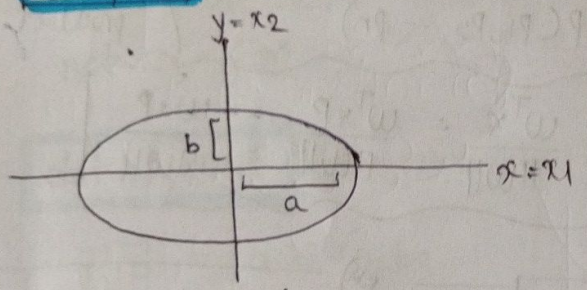
sphere - 3d - $x_1^2 + x_2^2 + x_3^2 = r^2$

hyper sphere - nd - $\sum_{i=1}^{n} x_i^2 = r^2$

if origin is (h,k)

$(x-h)^2 + (y-k)^2 = r^2$

Ellipse



a - major axis
b - minor axis

$$\left(\frac{x^2}{a^2} + \frac{y^2}{b^2}\right) = 1$$

< 1 → inside
> 1 → outside
$= 1$ → on elipse

ellipsoid ndimension

$$\sum_{i=1}^{n} \left(\frac{x_i^2}{a_i^2}\right) = 1$$

⑦ $\frac{d}{dx}(\sin x) = \cos x$

⑧ $\frac{d}{dx}\cos x = -\sin x$

⑨ $\frac{d}{dx}\tan x = \sec^2 x$

⑩ $\sec x = \sec x \cdot \tan x$

⑪ $\operatorname{cosec} x = -\operatorname{cosec} x \cdot \cot x$

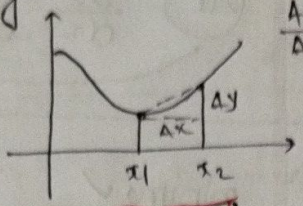
⑫ $\cot x = -\operatorname{cosec}^2 x$

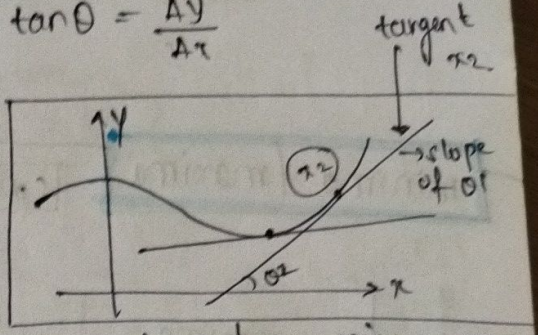
Differentiation

$\frac{dy}{dx}$ rate of change of y, as x changes

$\frac{\Delta y}{\Delta x} = \frac{y_2 - y_1}{x_2 - x_1}$; $\tan\theta = \frac{\Delta y}{\Delta x}$ ← slope

$\rightarrow \frac{dy}{dx} = \lim_{\Delta x \to 0} \frac{\Delta y}{\Delta x}$





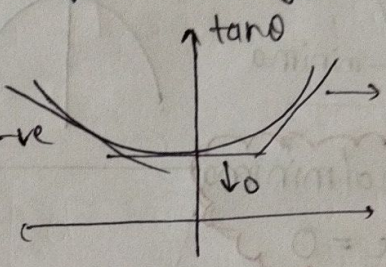
$\rightarrow \frac{dy}{dx}$ = slope of tangent to $f(x)$ → $\tan\theta$ → angle b/w tangent and x-axis

slope = $\tan\theta$

① $\theta = 0$, $\tan\theta = 0$

② $0 < \theta < 90 \rightarrow \tan\theta > 0$

③ $180 < \theta < 90 \rightarrow \tan\theta < 0$



$\tan 90 = \text{und}$

Basic Formulae

① $\frac{d}{dx}(x^n) = nx^{n-1}$

② $\frac{d}{dx}(c) = 0$

③ $\frac{d}{dx}(cx^n) = c \cdot nx^{n-1}$

④ $\frac{d}{dx}(\log x) = \frac{1}{x}$

⑤ $\frac{d}{dx}(e^x) = e^x$

⑥ $\frac{d}{dx}(f(x) + g(x)) = \frac{d}{dx}f(x) + \frac{d}{dx}g(x)$

chain rule

$\frac{d}{dx} f(g(x)) = \frac{df}{dg} \cdot \frac{dg}{dx}$

Ex:- $g(x) = (a - bx)$; $f(x) = x^2$; $f(g(x)) = (a-bx)^2$

$\rightarrow \frac{d}{dx} f(g(x)) = \frac{df}{dg} \cdot \frac{dg}{dx}$

$\rightarrow \frac{dg}{dx} = \frac{d(a-bx)}{dx} = -b \cdot \frac{dx}{dx} = -b$; $\frac{dg}{dx} = -b$

$\rightarrow \frac{df}{dg} = \frac{d}{dg}$ {let $(a-bx) = z$}

$\rightarrow \frac{df}{dg} = \frac{dz^2}{dz} = 2z \rightarrow (a-bx)$

$\frac{df}{dg} = \frac{d(a-bx)^2}{d(a-bx)} \cdot 2(a-bx)$

$\therefore \frac{dy}{dx} = (-b)(2(a-bx))$

$\frac{dy}{dx} = -2b(a-bx)$

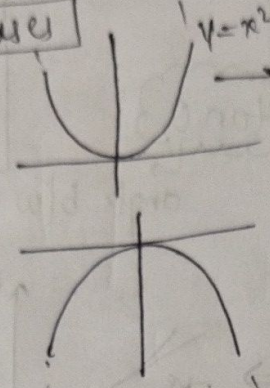
(10) * $\frac{d}{dx}(\sin x)^2 = 2 \cdot \sin x \cdot \cos x$

steps let $\sin x = u$, $\frac{dy}{dx} = \frac{d}{du}u^2 \times \frac{du}{dx} = \frac{d(u)}{du} \cdot \frac{du}{dx} = 2 \cdot \cos x \cdot \sin x$

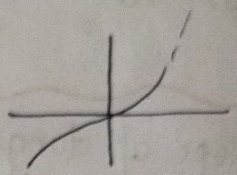
$\frac{du^2}{du} = 2u$ $\quad \frac{du}{dx} = \frac{d \sin x}{dx} = \cos x$

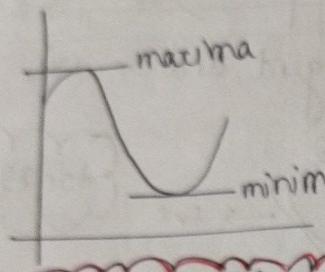
minima | maxima

case 1

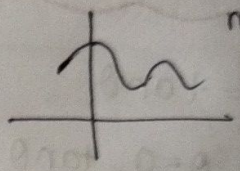


$y = x^2$ → minima ✓, no maxima

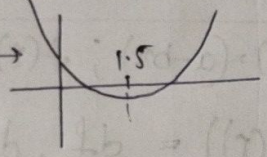




→ maxima, no minima



NB at maxima/minima slope = 0

Ex: $f(x) = x^2 - 3x + 2$ → 

for max/min slope = 0

$\frac{d}{dx}(x^2 - 3x + 2) = 0$

$2x - 3 = 0$

$x = \frac{3}{2} = 1.5$

min $f(x) = -\max -f(x)$

max $f(x) = -\min -f(x)$

at $x = 1.5 \rightarrow$ slope = 0 < maxima / minima ✓

$\rightarrow f(1.5) = -0.25$

$f(1) = 0$

as $f(1) \geq f(1.5)$ (some pt is higher)

∴ $f(1.5)$ is minima

minima at 1.5

(2) $f(x) = \log(1 + e^{ax}) = y$

$\frac{df(x)}{dx} = 0$

$\frac{dy}{dx} = \frac{1}{(1 + e^{ax})} \cdot e^{ax} \cdot a = 0$

$\frac{a \cdot e^{ax}}{1 + e^{ax}} = 0 \rightarrow a \cdot e^{ax} = 0$

→ function is complex and not as easy as previous.

↓

Gradient Descend can do it easily

Vector differentiation (grad / dell)

→ x : vector $y = a^T x$; $x = \langle x_1, x_2 \dots x_d \rangle$

$a = \langle a_1 \dots a_d \rangle$

$f(x) = y = \sum_{i=1}^{n} a_i x_i$

$a^T x = a_1 x_1 + a_2 x_2 + \dots a_n x_n$

$\frac{\partial f}{\partial x_1} = a_1$, $\frac{\partial f}{\partial x_2} = a_2$

***

$\frac{df}{dx} = \nabla_x f$ → grad / dell

$\frac{df}{dx}$ → vector

grad / dell

$\frac{df}{dx} = \begin{bmatrix} \frac{df}{dx_1} \\ \frac{df}{dx_2} \\ \vdots \\ \frac{df}{dx_d} \end{bmatrix}$ → vector $\in \mathbb{R}^d$

$\frac{df}{dx_1}$ → partial diff

$= \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_n \end{bmatrix} = \vec{a}$

$\therefore \frac{df}{d\vec{x}} = \frac{d(a^T \cdot \vec{x})}{d\vec{x}}$

$= a$

Example 2

* logistic loss(L) = $\sum \log(1 + \exp(-y_i w^T x_i)) + (\lambda w^T w)$

$\frac{\lambda w^2}{dw} = 2w\lambda$

x_i, y_i → constants

$\frac{dL}{d\vec{w}} = \nabla_w L = \frac{\exp(-y_i w^T x_i) \cdot (-y_i \cdot x_i)}{1 + \exp(-y_i w^T x_i)} + 2\lambda w = 0$

→ solved by gradient descend

$dw_i = \frac{dL}{dw_i}$

$w = [w_1\ w_2\ w_3]$

$x = [x_1\ x_2 \dots x_3]$

$w_{t+1} = w_{t-1} - \eta (\nabla L)_{w_{t-2}}$

$\begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix} - \eta \begin{bmatrix} dw_1 \\ dw_2 \\ dw_3 \end{bmatrix} = \begin{bmatrix} w_1 - \eta dw_1 \\ w_2 - \eta dw_2 \\ w_3 - \eta dw_3 \end{bmatrix}$

→ run forward-prop

→ run back-prop → you get (dw_1, dw_2, dw_3) w.g = $[dw_1, dw_2, dw_3]$

(12)

GRADIENT DESCENT

Let r = 1

Step 1 pick any random (x_0) as initial point

Step 2

$$x_1 = x_0 - r\left[\frac{df(x)}{dx}\right]_{x_0}$$

UPDATE FUNCTION

$$x_{i+1} = x_i - r\left[\frac{df(x)}{dx}\right]_{x_i}$$

lr / step size — $\frac{df}{dx}$ at x_i

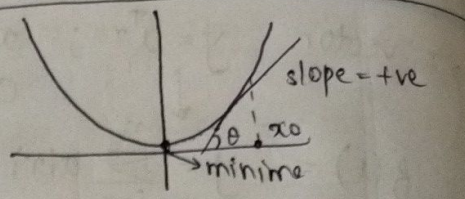
$x_0, x_1 \; - \; - \; x_K$

Step 4

→ if $(x_{K+1} - x_K)$ is very small,

declare $\boxed{x^* = x_K}$

Case 1



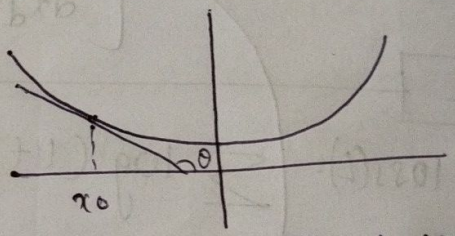
$x_0 \rightarrow$ slope = +ve

$\rightarrow x_1 = x_0 - 1\left(\frac{df}{dx}\right)_{x_0}$ → +ve

$\rightarrow x_1 = x_0 -$ positive value

$\therefore \boxed{x_1 < x_0}$ closer to minima

Case 2



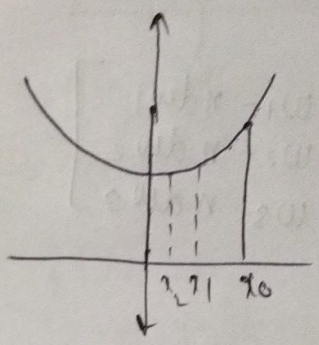
$\rightarrow \left[\frac{df}{dx}\right]_{x_0} < 0 \; [\tan\theta < 0]$

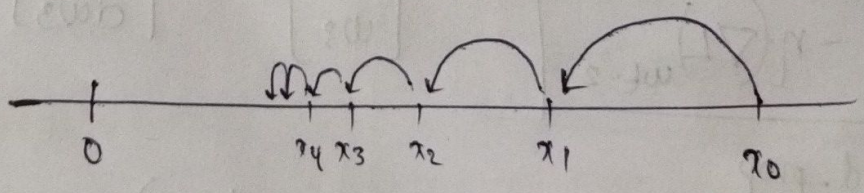
$x_1 = x_0 - 1(-ve) = x_0 + (val)$

$\therefore \boxed{x_1 > x_0}$ closer to minima

Note

$\left[\frac{df}{dx}\right]_{x_0} \geq \left[\frac{df}{dx}\right]_{x_1} \geq \left[\frac{df}{dx}\right]_{x_2}$ → because $\theta \rightarrow 0$ or as $\theta \rightarrow 90°$, $\left[\frac{df}{dx}\right] \uparrow$





→ step distances decreases after each iteration

problem : Oscillation → updation may lead to same 2 value
ex → 0.5, -0.5, 0.5, -0.5

→ sol^n :- (i) change [r] with each iteration
→ reduce (r) with each iteration

Stochastic gradient descent for Linear regresion

→ In GD,

$$x_{i+1} = x_i - r\left[\frac{df}{dx}\right]_{x_i}$$

Incase of Linear regresion → $\frac{df}{dx} = \frac{df}{dw} = \sum_{i=1}^{n} -2x_i (y_i - w^T x_i)$

→ you cont calulate (i=1 → n) fr all the points
it is computationally expensive

(so) pick any (K) points instead of (n) where (K << n)

$$x_{i+1} = x_i - r \sum_{i=1}^{n} (\cdots) \Longrightarrow \boxed{x_{i+1} = x_i - r \sum_{i=1}^{K} (\cdots)}$$

if (K == 1) → stochastic gradient descent
if (K >>> 1 and K << n)
→ Batch - stochastic gradient descent (BGD)

Constrained Optimization →

maximize f(x) → objective function
such that
g(x) = c
h(x) ≤ 0 → constraint function

Ex :- max $\frac{1}{n} \sum_{i=1}^{n} (u^T x_i)^2$ given $u^T \cdot u = 1$

(14)

Lagrangian multipliers

Pblm

$x^* = \max f(x)$ given $g(x) = c$, $h(x) = d$

$$L(x, \lambda, \mu) = f(x) - \lambda\{g(x) - c\} - \mu\{(h(x) - d)\}$$

→ we can proove that,

diff L wrt x, λ, μ

$$\frac{\partial L}{\partial x} = 0 \; ; \; \frac{\partial L}{\partial \lambda} = 0 \; ; \; \frac{\partial L}{\partial \mu} = 0$$

we get (3) equations → solving → $\tilde{x}, \tilde{\mu}, \tilde{\lambda}$

$\therefore \tilde{x} \equiv x^*$

Norms

L-p norm $= \left(\sum |x_i|^p\right)^{1/p}$

L1 → $\sum |x_i|$

L2 → $\sqrt{\sum x_i^2}$

Uniform distribution → Equiprobable

- → discrete ud
- → continous ud

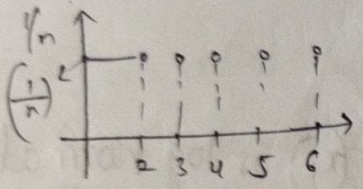
① Discrete uniform d:-

Ex:- coin toss / dice roll

U(2,6) → 2, 3, 4, 5, 6

mean = $\frac{a+b}{2}$ = median

prob-mass function ≡ pdf



Bernullis distrib:-

↳ 2 outcomes only.

Ex:- coin toss

p - prob of outcome 1

q = prob of outcome 2

$p + q = 1$

Binomial Distribut:-

X — Bernouli (p=0.5)

↓

y — Binomial (n, p)

↳ no of trials

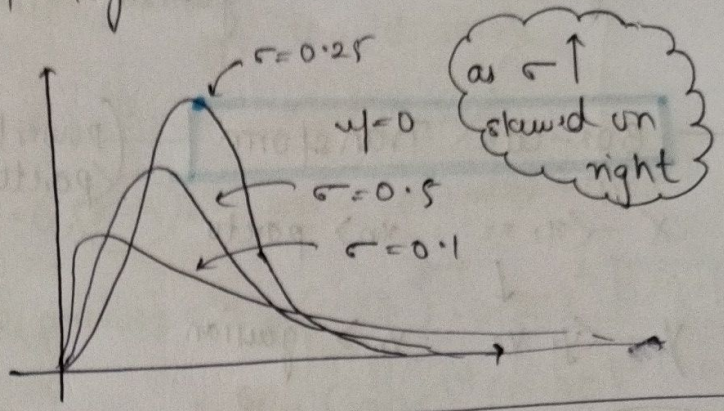
$pmf = \binom{n}{k} p^k (1-p)^{n-k}$

k = success / (q=K)

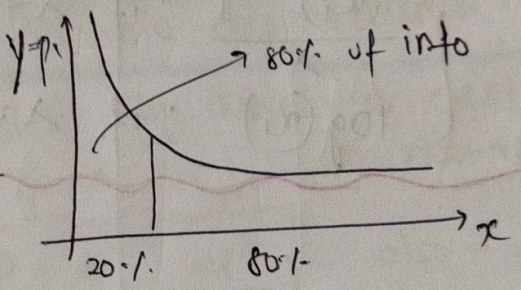
log-Normal

$X = \log \text{Normal}(\mu, \sigma)$

if $\log(x) \in$ Normal distribut?



Power law distrib:-

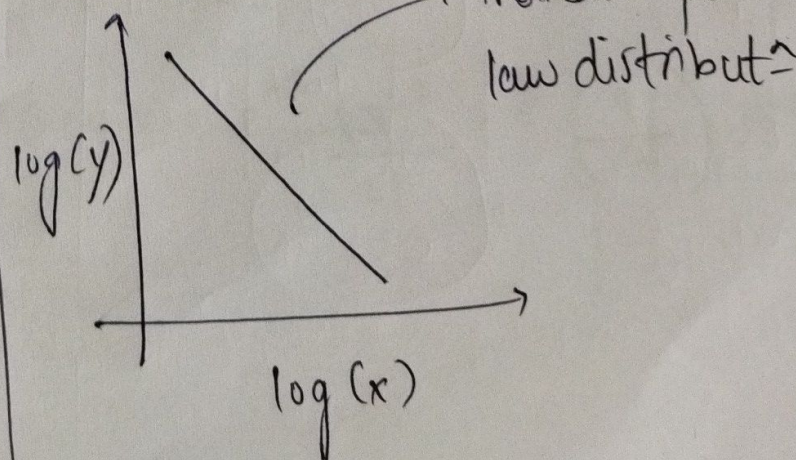


Check power law distribution

x → x

y → f(x) / p(x)

log-log graph :-



(16) **lognormal → Gaussian**

$X \xrightarrow{\log(x)} Y$

X: log normal

Y: gaussian

Box-Cox Transform (powerlaw / pareto → gaussian)

$X \rightarrow \langle x_1\ x_2\ \ldots x_n \rangle$ pareto

$\downarrow$

$y \rightarrow \langle y_1\ y_2\ \ldots y_n \rangle$ gaussian

step 1: Box-cox(X) $\rightarrow \lambda$

step 2:

$$y_i = \begin{cases} \dfrac{(x_i)^{\lambda} - 1}{\lambda} & ; \text{if } \lambda \neq 0 \\ \log(x_i) & ; \text{if } \lambda = 0 \end{cases}$$

→ means x_i is log normal

Probability & distributions

Binomial Distribⁿ

2 outcomes { yes / no

$$P(K/N) = \binom{N}{K} p^K q^{N-K}$$

N → no of trails
p → prob of outcome (K)
q = 1 − p
K → no of succes / fail

Ex 2 die rolled 3 times
(i) P(5 doesn't apear)
(ii) P(~~atleast~~ 1,5)

→ N = 3

(i) $P_S = \frac{1}{6}$ (success)

K = 0 ; no 5

$$P(\text{NO 5}) = {}^3C_0 \left(\frac{1}{6}\right)^0 \left(\frac{5}{6}\right)^3$$

(ii) $p(n=1) = {}^3C_1 \left(\frac{1}{6}\right)^1 \left(\frac{5}{6}\right)^2$

$p(n \geq 1) = p(n=1) + p(n=2) + p(n=3)$
atleast 1 time

Ex → in 100 ppl — 75 die
what is prob that out of 6 ppl 4 will survie

→ P(die) = 75% P(sur) = 25%

→ N = 6, K = 4 ← survie p(success) = 0.25

→ $P(\text{survive}) = {}^6C_4 (0.75)^{6-4} (0.25)^4$

$= {}^6C_4 (0.25)^4 (0.75)^2$

$= 0.0329$

Poisons - distribution

used in time based or event period (INTERVALS)

$$P = \underset{P(x)}{e^{-\mu}} \cdot \frac{\mu^x}{x!}$$

also − $\mu = np$ ← pop^n size

Ex - person sells 3 cars / week.
what is probability that he sells 2 cars this week?

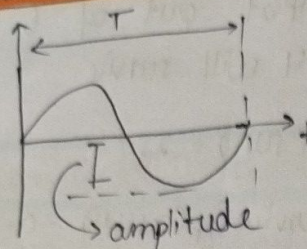
$\mu = 3$, $x = 2$

$$P(x=2) = \frac{e^{-3} \cdot (3)^2}{2!} = e^{-3}\left(\frac{9}{2}\right)$$

(18)

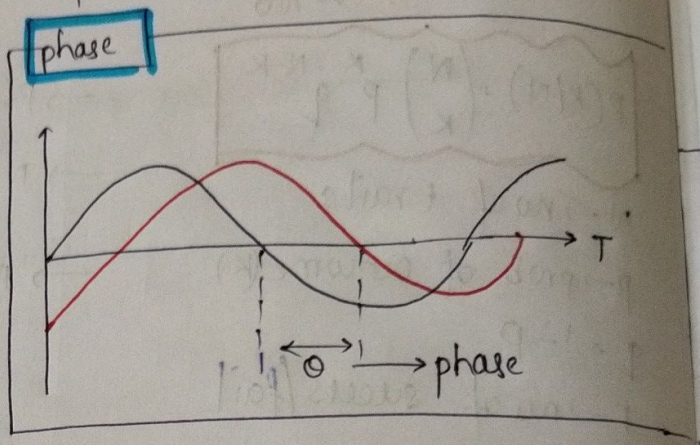
Fourier Transforms

Basic terms



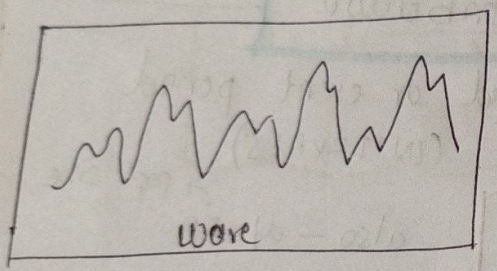
* amplitude : height of wave
* T : time to complete one oscilation
* f : $\frac{1}{T}$

phase

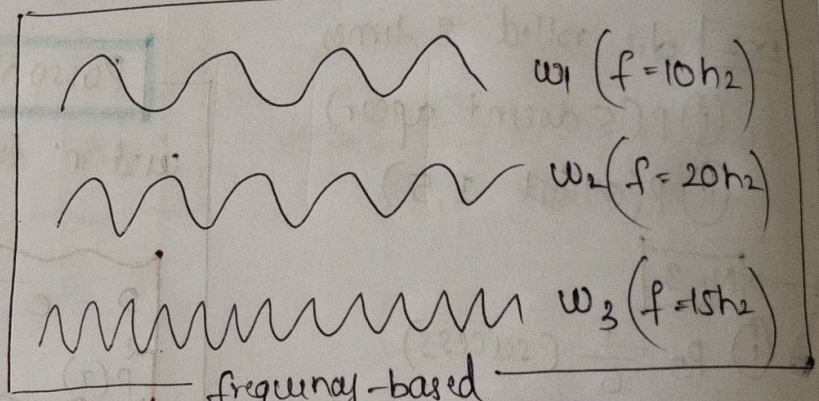


logic

* if pattern is repeating
→ wave can be decomposed into multiple waves

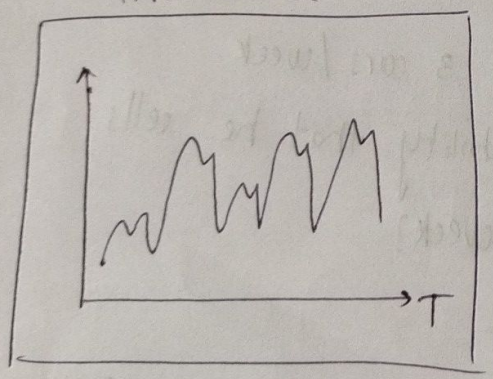


FT →

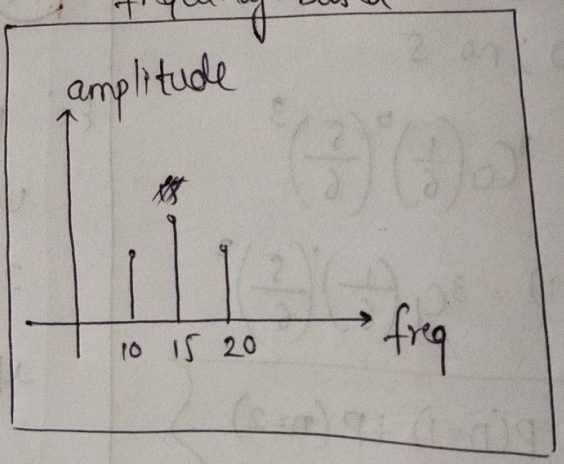


frequency-based

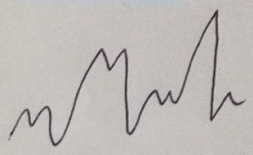
time-based



Fourier transform →



Steps



→ (f_1, a_1)
(f_2, a_2)
(f_3, a_3) →

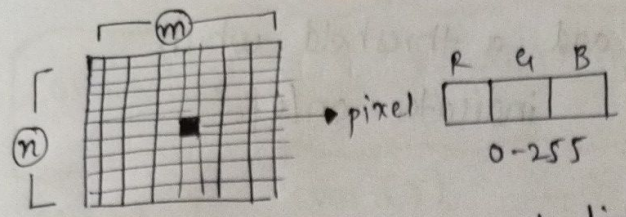
feature

f_1	f_2	f_3	a_1	a_2	a_3

⇓

x_i

Image Data



→ cant distinguish shapes | only colors

Color Histograms

get Red values for each pixel (n×m) → plot hist → convert to vector

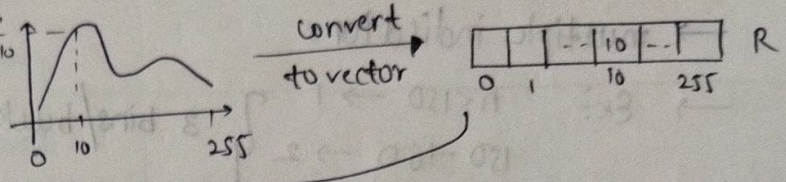
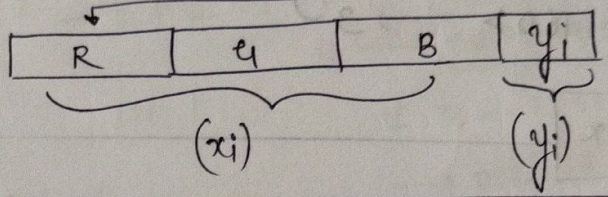


image →



Edge Histograms

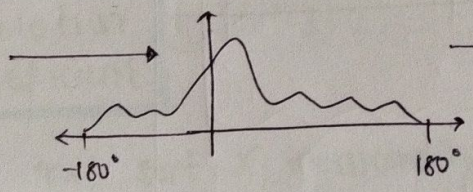
Edge → drastic change in color → value = location

↳ angle of edge → / = 45°, — = 0°

divide image to regions

for each region

calculate edge-value or edge-angle →



→ convert to vector of features

used in face detection ⇒ Haar features → object detection

SIFT (scale invariant feature transforms)

↳ popular to detect objects in image

Key idea : stores important features and then compares

↳ SIFT features

main advantage :- scale independent } to some extent
rotatation independent }

(20) **Feature indicator**

Ex: feature f_1 ⇒ height ⟶ add a threshold which indicates on/off.

if (h<150) ret 0

else

ret 1

feature bin

⟶ multiple indicators

⟶ Ex: h<150 → 1
150-160 → 2
160> → 3
} 3 bins/buckets

TN FP
FN TP

conf_matrix

actual

pred	0	1
0	TN	FN
1	FP	TP

sklearn

pred

actual	0	1
0	TN	FP
1	FN	TP

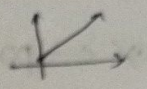
Covariance

u cant say how much corelated

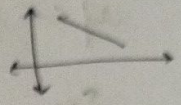
$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$

$$cov(X,X) = var(X) = \frac{1}{n}\sum(x_i - \mu_x)^2$$

cov = +ve



cov = -ve



drawbacks : changes if metrics for measurement are changed

Pearsons Corelation Co-ef

$$\rho_{x,y} = \frac{cov(X,Y)}{\sigma_x \cdot \sigma_y}$$

$-1 \le \rho \le +1$

if $\rho = 0$,
→ no corelation

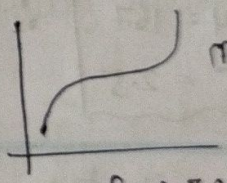
if $\rho = -1, +1$,
→ perfect linear relat⁻
→ slope of line doesnt matter

prob-m : biased towards perfect linear relationship
→ cant capture non-linear rel⁻ (sin, cos ...)

Spearman rank corelation coeficient

x	y	r_x	r_y
160	50	3	2
150	70	2	4
140	40	1	1
190	60	4	3

r_x : sort x in ascending order
→ smallest - rank 1



monotonically increasing

$\rho = 0.79$
spearman = 1

$$\rho = \rho_{r_x, r_y} = \frac{cov(r_x, r_y)}{\sigma_{r_x} \cdot \sigma_{r_y}}$$

* it cares only if $x_1 > x_2, y_1 > y_2$ and doesnt care abt linear

* Better when outliers are present

Causal models

Note | Corelation ≠ causation | x corelated to y ≠ x causes y

(22) **Confidence Interval**

x ← population (heights of ppl)

μ ← population mean

$\{x_1, \ldots x_{10}\}$ sample of $n=10$, drawn from x.

$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$ (sample mean)

* point estimate = $\mu \approx \bar{x}$ (as $n\uparrow$, $(\mu = \bar{x})\uparrow$)

* CI $\mu \in [x, y]$ with (95%) probability.

[x, y] → interval

(95%) → confidence

all these are to find μ

if x is gausian

$X \leftarrow N(\mu, \sigma)$ → wkt 95% points in (N) lie b/w $\mu+2\sigma$, $\mu-2\sigma$

let $\mu = 167$, $\sigma = 2.5$ ∴ $\mu = [167-5, 167+5]$ with 95% prob

if distribution is unknown

→ $X \sim F$ (any distribution)

u know the pop-std $= \sigma$

mean of k-samples each of size n → [(1--n), (1-n) (1--n)]

then,

Central limit theorem gives: $\bar{x}$ = sample mean $= \sum_{i=1}^{(n)} x_i$ → sample size

then, $\boxed{\bar{x} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)}$

$\mu \rightarrow$ pop-mean

$\sigma \rightarrow$ pop-std (given)

the distribution of means of sample-pop of size (n) is a gausion distribution

then we can say,

$\mu \in \left[\bar{x} - \frac{\sigma}{\sqrt{n}} \cdot 2,\ \bar{x} + \frac{2\sigma}{\sqrt{n}}\right]$ with prob of 95%.

case 2 you dont know σ of pop^n

→ t-distribution

$\bar{x} \sim t(n-1)$ → degree of freedom → as df↑ peakness↑

How to find confidence intervals for σ, median, percentile? → **Bootstrap CI**

① find 95% of CI for median

X → distribution unknown

Sample with size 'n' : $\{x_1, \ldots x_n\}$ find CI of median of X ?

(only one S)

$S = \{x_1, x_2 \ldots x_n\}$ → sampling with replacement (using uniform r.v)

→ $S_1 = \{x_1', x_2' \ldots x_m'\}$: random sample of size (m) picked

$S_2 = \{$ $\}$ from (S) where $m < n$

$S_k = \{x_1^k, x_2^k \ldots x_m^k\}$ → called Bootstrap samples

lets take $K = 1000$,

→ find medians for each bootstrap sample → **CI for median**

$m_1, m_2 \ldots m_{1000}$

↓ sort in increasing order

$m_1', m_2', \ldots m_{1000}'$

25 val $< m_{25}$ | $m_{25}' \ldots m_{975}'$ | 25 val $> m_{975}$

95% CI of median of X = $[m_{25}', m_{975}']$

(24) **CI for variance**

calculate bs variances

$\sigma_1^2, \sigma_2^2 \; - - \; \sigma_{1000}^2$

↓ sort

$\sigma_1'^2, \sigma_2'^2, \; - - \; \sigma_{1000}'^2$

95% = P/C

lower = $\frac{1-C}{2}$

higher = $\frac{1+C}{2}$

lower = 2.5 %

higher = 97.5 %

2.5 ////////// 97.5

CI

95% CI for variance

$= [\sigma_{25}'^2, \sigma_{975}'^2]$

Hypothesis test

given two class student heights. Each class has 50 students

Question : Is there difference in height of C_1 compared to C_2

① choose test measure

lets takes mean as measure, μ_1, μ_2 → mean of heights of C_1 and C_2

② Null hypothesis (H_0)

H_0 : no difference in μ_1 and μ_2

(alternate hyp) H_1 : there is difference

③ P-value

Ex Given a coin determine if it is biased

biased $= p(H|T) \neq 0.5$

→ Expt : flip coin 5 times

Expt obs :- Head occured 5 times (X = # of heads = 5)

↓ after performing experiment

$X \leftarrow$ test statistic

obs $= (X = 5)$ after doing exp

→ H_0 : unbiased, H_1 : biased

→ $p(\text{obs} | H_0) = \left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(\frac{1}{2}\right)$

$= \frac{1}{2^5} = \frac{1}{32}$

$p(\text{obs} | H_0) = 3\%$

as $p < 5\%$, reject H_0, accept H_1

→ try this for different experiments and take average result

$p(\text{obs} | \text{asumption}) = x\%$

if ($x < 5\%$) :

→ maybe asumption is wrong

→ reject asumpt :

Calculate percentile

① sort →

12	14	35	43	45	47	48	78	80	98
1	2	3	4	5	6	7	8	9	10

rank

$$\%P = \left(\frac{P}{100}\right)(n+1)$$

→ 25% → 0.25×11 → 2.75

if % is decimal

$$x.y = a[x] + (a[x+1] - a[x])\,(y)$$

$\therefore$ 2.75

$a[2]$ + 75% of $d(a[2], a[3])$ → difference

14 + (.75) of (14 ↔ 35)

→ (14) + (.75)(21)

= 29.75

35 − 14 diff

Ex ② 75% → (.75)(11) = 8.25 ans = 78 + (.25 * (80 − 78))

Errors

① $MSE = \frac{1}{n}\sum(y-\hat{y})$

② MAPE (mean abs percentile error)

$$MAPE = \frac{1}{n}\sum\left|\frac{y-\hat{y}}{y}\right|$$

→ if $y=0$ // exception

$$\bar{y} = \frac{\sum y_i}{n}, \quad n\bar{y} = \sum|y_i|$$

③ $R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$ → $\sum(y_i - \hat{y}_i)$

modified-MAPE

$$M\text{-}MAPE = \frac{1}{n}\sum\left(\frac{y-\hat{y}}{\bar{y}}\right)$$

$\bar{y}$ → mean(y)

$$MM = \frac{1}{n}\,\frac{|e_1| + |e_2| + \dots + |e_n|}{\bar{y}}$$

$$MM = \frac{|e_1| + |e_2| + \dots + |e_n|}{\sum|y_i|}$$

model agnostic → **Forward Feature Selection** → time complexity is high

(98) consider dataset with 10 features $f_1 - f_{10}$

it 1 Train with only single feature and save accuracy obtained

$f_1 \to a_1$, $f_2 \to a_2$ --

→ pick feature with highest a

Ex f_3

it 2 f_3 is fixed, → which feature along with f_3 would give us most value?

Now train each feature in $F - \{f_3\}$ and calculate accuracy

→ $f_1, f_3 \to a_1$
$f_2, f_3 \to a_2$
$f_4, f_3 \to a_4$
} best obt for f_5, f_3

it 3 f_5, f_3 is fixed, now pick from $f - \{f_3, f_5\}$ and continue.

problems → can be applied to any algorithm but computationally expensive

Backward feature selection

step 1 → pick all weights and calculate accuracy

it 1 In each it, drop one weight and check which feature droped cause least decrease in accuracy comp to acc with all feature

it n remove a feature at each iteration which causes lowest decrease in accuracy

Time complexity

it 1 → train d models

it 2 → train (d-1) models

it 3 → train (d-1) models

⋮

seaborn as sns

IRIS Dataset

→ iris = pd.read_csv()

→ iris["species"].value_counts()

→ **2d plot**

* iris.plot(kind = 'scatter', x = '', y = '')
* plt.show() ↳ field

→ **3D plot (color)**

* sns.set_style("whitegrid")
* sns.FacetGrid(iris, hue = "species", size = 4) (height)
  .map(plt.scatter, 'sepal_length', sepal_w)
  .add_legend();
* plt.show()

PAIR PLOT → no of plots = ${}^{n}C_2$ (n → no of features)
→ useful upto 6 features

sns.pairplot(iris, hue = 'species', size = 3)

UNIVARIATE ANALYSIS
↳ single variable analysis (pick 1 best feature)

* sns.FacetGrid(iris, hue = 'species')
  .map(sns.distplot, 'petal_length')
  ↓
  histogram on petal length

EXPLORATORY DATA ANALYSES

MEAN → μ → $\frac{\sum x_i}{len(x_i)} = \frac{\sum x_i}{n}$
↳ outlier cause deviation

csv
Header, ---
data1, data2, -

MEDIAN ① sort in ascending order ② pick middle value
same problem

variance → spread → how far from mean.

$var = \frac{1}{n}\sum_{i=0}^{n}(x_i - \mu)^2$ → to remove -ve

-ve ——— μ ——— +ve

$\frac{\sum x_i}{n}$ → avg dist of pts from mean

$std = \sqrt{var} = \sqrt{\frac{1}{n}\sum_{i=0}^{n}(x_i - \mu)^2}$
↳ average spread from mean

np.mean()
np.std()
np.median()

MEDIAN BASED

percentiles → where do you lie in sorted list

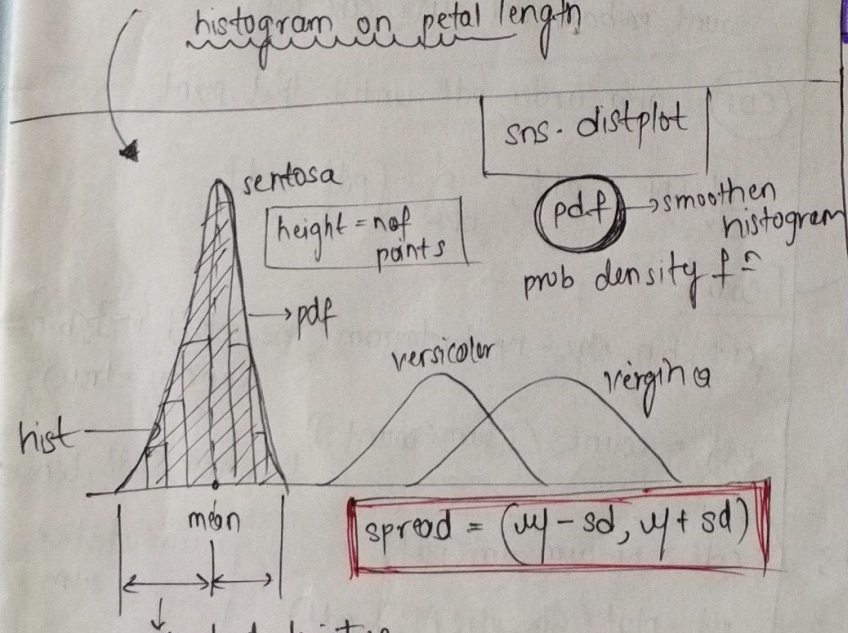
50th percentile = arr.sort()] len = 100
arr[50]
↳ meaning → 49 values are smaller than it

50th percentile = median ***

Quantile

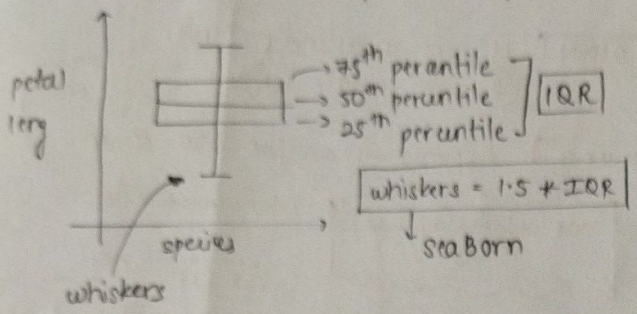
25th percentile ——→ 1st Quantile
50th " ←——→ MEDIAN ——→ 2nd Quantile
75th " ——→ 3rd Quantile
100th " ——→ 4rd Quantile (max value).

* np.percentile( )
  ↳ arr, percentile



BOX PLOTS AND WHISKERS

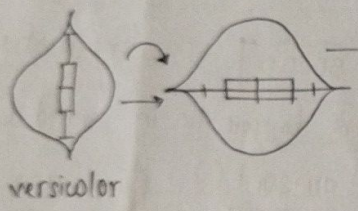
* sns.boxplot(x=, y=, data="iris")
  dataset
  ex → (x = species, y = petal_length)



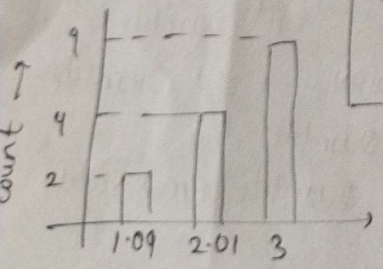
VIOLIN PLOT

↳ Boxplot + histogram

* sns.violinplot(x = species, y = len, data = iris)



→ np.histogram(arr, bins = 10)
returns
(count, bin_edge)



([2,4,9], [0.1.09, 2.01, 3])

[1:] → bin_edge

pdf → count / sum(counts)

contour

sns.jointplot(x = " ", y = " ", data = ds, kind = "kde")

median absolute deviation
~~mean~~ absolute deviation
↳ how far away from (median)
sd → mean

{ no outlier problems

→ from statsmodels import robust

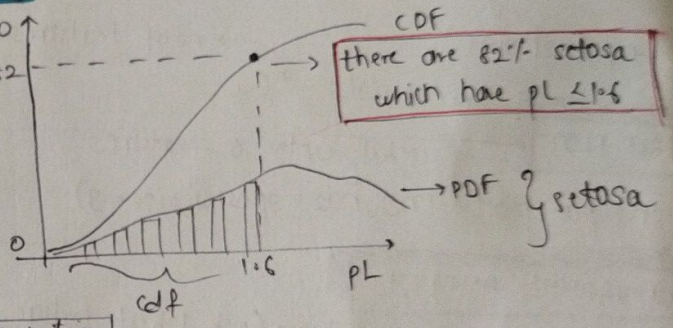
* robust.mad( )

$$MAD = \sum \sqrt{|x_i - median|} / n$$

Inter-Quartile range | IQR

75th percentile − 25th percentile = IQR

CDF cumulative distribution function



Calculation

→ x = val
→ y = $\frac{\text{how many pts have val} \le x}{\text{total pts}} = \frac{41}{50} = 0.82$

or

count probabilities upto our point

(CDF) = area under pdf untill that point

$\frac{d}{dx} CDF = pdf$, $\int pdf = cdf$

Code

cnt, bin_edge = np.histogram(iris_setos['pl'], bins=, density = true)

pdf = counts / (sum(counts))
↳ print(pdf, bin_edge)

cdf = np.cumsum(pdf) → cumulative sum
plt.plot(bin_edge[1:], pdf)
(, cdf))

Mathematics for ML

population

sample

Bell curve

Gaussian / Normal distribution

peak at μ

μ → height, reduced width

σ^2 → width, reduced height

X is Normal distr → $X \sim N(\mu, \sigma^2)$

$$P(x) = \frac{1}{\sqrt{2\pi}\cdot\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

$e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$

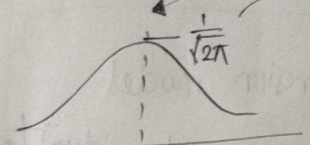
let $\mu = 0$, $\sigma^2 = 1$

$P(x) = \frac{1}{\sqrt{2\pi}} e^{\left(-x^2 \cdot \frac{1}{2}\right)} = e^{-x^2}$

$-(-2)^2 = -4$

$-(2)^2 = -4$

$e^0 = 1$



① $x \uparrow -x^2 \downarrow e^{-x^2} \downarrow$

② $x \downarrow -x^2 \updownarrow e^{-x^2} \downarrow$

③ symetric

④ y reduces (exponential)^2

$P(x) = e^{-x^2} = \frac{1}{e^{x^2}} = y$

Covariance

How elements r related?

split ds

hb-y = hb[hb['rs'] == yes]

$$cov(X,Y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$$ 1/2

$cov(x,x) = var(x)$

Chebyshevs Inequality

if Y ≉ Gaussian distribution

$P(\mu - K\sigma \leftrightarrow \mu + K\sigma) > 1 - \frac{1}{K^2}$ %

Central Limit Theorem

$X \not\approx GD(\mu, \sigma^2)$

$n \geq 30$

$S_1 \rightarrow x_1 + x_2 + \cdots x_{30} = \overline{x_1}$

$S_2 \rightarrow x_1 + x_2 + \cdots x_{30} = \overline{x_2}$

$S_{100} \rightarrow x_1 + x_2 + \cdots x_{30} = \overline{x_{100}}$

→ mean = $\overline{X}$

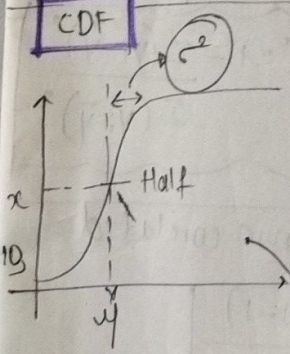
$\overline{X} \approx GD(\mu, \frac{\sigma^2}{n})$

plotting means $\overline{x_1} \cdots \overline{x_n}$

$\overline{x} \approx \mu$

$\sigma \approx \sigma^2/n$

CDF



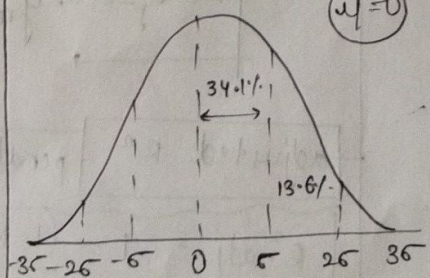
① smaller variance, less apart from mean

probability that of val ≤ y is x

68-95-99.7 rule

SD

$\mu = 0$



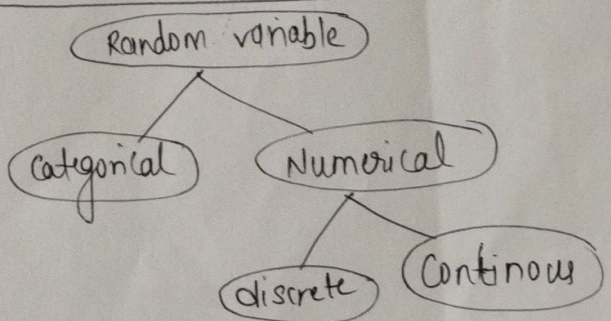
b/w $-\sigma - \sigma$ → 68% pts

$-2\sigma - +2\sigma$ → 95% pts

-3σ to 3σ → 99.7% pts

68% have some quan b/w $\mu-\sigma$ | $\mu+\sigma$

$-\sigma$ to σ



04 CORELATION COEFFICIENT

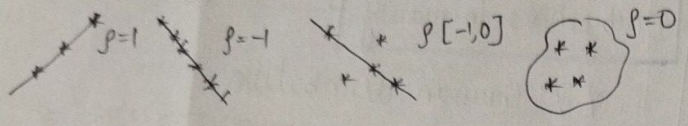
* $cov(x,y) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \mu_x)(y_i - \mu_y)$ → not quantified

$$PCC = \rho(x,y) = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$

↳ op → direction of cor
↳ strength

how much +ve or -ve

$-1 \le \rho \le 1$



- if $\rho(x,y) = 1$, u can drop one

Spearmans (rank) corelation

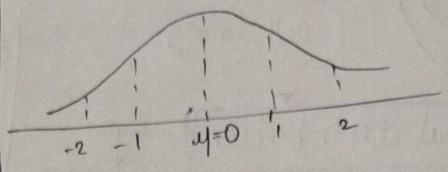
(pearsons only focuses on linear aspects)

$$r_s = 1 - \frac{6 \cdot \sum d_i^2}{n(n^2-1)}$$ ← no of obs

① sort first col, X.
② r_A, r_B → rank acc to ascending order
③ $d_i = r_A - r_b$

Standard Normal distribution

Gausion distribution with $\mu = 0, \sigma = 1$

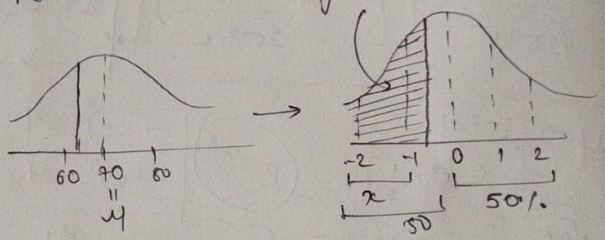


Z-score

Gausian → standard Normal

$$Z_{score} = \frac{x_i - \mu}{\sigma}$$

$P(x > 65) = 50 + (50 - x)$ get from z-table



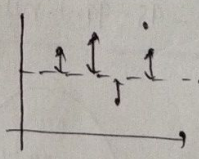
R-Square score

Accuracy of linear regresion model

$$R^2 = 1 - \frac{SS_{Res}}{SS_{tot}}$$

SS_{Res} → sum of residual / error (dif of predicted, original pts)

SS_{tot} → diff b/w points and y_{mean} line



$$R^2 = 1 - \frac{\sum (y_i - \hat{y})^2}{\sum (y_i - \bar{y})^2}$$

adjusted R^2

penalizes non corelated

$$R^2\text{-adj} = 1 - \frac{(1-R^2)(N-1)}{N-p-1}$$

p = no of predictors → (x)
N = sample size

$p \uparrow \cdot R^2 \downarrow$

Encoding ← Categorical variables

no rank

Nominal → 1-hot enc
↓ with many cat var
→ mean encoding

Ordinal → label enc
→ target guided

↑ order maters (education)

BIAS & VARIANCE

underfiting → error with training data

overfitting → acc for train ↑ – less for test

BIAS → error of training data

VARIANCE → error for ~~training~~ test data

Generalized model → low variance
low bias

→ Bias - Variance - trade off (manage both)

Probability Density functions

→ Normal
→ students T
→ Binomial

→ Bernoulis
→ uniform
→ poissons

07

Binomial dist

bi → 2 values (True false)
Independent val
→ more than one outcome

$$P(x) = \binom{n}{x} p^x q^{n-x} = \frac{n!}{(n-x)!x!} \cdot p^x (1-p)^{n-x}$$

n → no of trails
x → successes / no of ↑
p → prob of succes
q → (1−p)

Ex:- dice rolled thrice, prob of getting (3) twice
n = 3, p = 1/6, q = 5/6, x = (2)

Poissons Distnbutions

$$p(x) = e^{-\mu} \cdot \frac{\mu^x}{x!}$$

μ → average / mean for interval

→ prob of events in a time period

Normal / Gaussian

$$f(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

→ if $\mu = 0 / \sigma = 1$ → standard normal distrib

→ Normal → standard ND

$$z = \frac{x-\mu}{\sigma}$$ z-score

z-statistics → probability in ranges → z-table

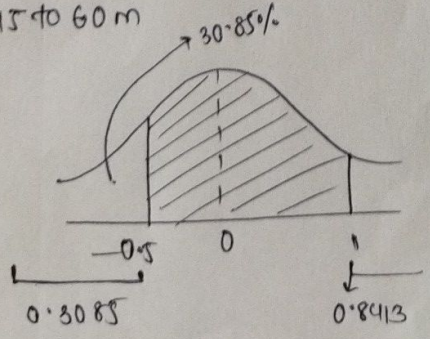
→ σ = 10, μ = 50

→ d.f → ~~prob of~~ work done in x min → [1 .. 4]

→ p(work done in) 45 to 60 m

$\frac{45-50}{10} \leftrightarrow \frac{60-50}{10}$

$-0.5 \leq x \leq 1$



→ 0.8413 − 0.3084 → 0.5329 (53.29%)

z-score estimation

df = profits for 12 year → n = 12, μ = 10, σ = 40

profit 18th year = ?

$$\mu \pm Z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$$

α → 1 − (lvl of confidence)
(lvl of significance)

Ex:- confidence = 90%, α = 1 − 0.9 = 0.1/2 − 0.0

→ $Z_{0.05}$ = 1.621 ← from z table

→ a ≤ x ≤ b

∴ profit lies b/w (a, b) with 90% confi

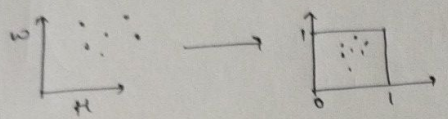
Data Normalization → column normalization

feature

col-a $x_i = \frac{x_i - x_{max}}{}$

$x_i = \frac{x_i - x_{min}}{x_{max} - x_{min}}$ $\quad x_i \to 0 \leftrightarrow 1$

→ gets rid of units
→ restricts data to grid



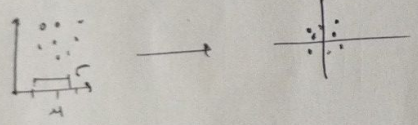
mean vector = $[\bar{f_1}, \bar{f_2}, \bar{f_3} \dots \bar{f_n}]$
↳ central data point

Column standardization

$a_1, a_2 \dots a_n \longrightarrow a_1', a_2', a_3' \dots a_n'$

$\sigma = 1, \mu = 0$

$z\text{-score} = a_i' = \frac{a_i - \mu}{\sigma}$



Covariance

	f_1	f_2	f_n
x_1	x_{11}	x_{12}	x_{1n}
x_2	x_{21}	x_{22}	x_{2n}
⋮			
x_n	x_{n1}	⋯	x_{nn}

$cov(f_1, f_1) = var(f_1)$
$cov(f_1, f_2) = cov(f_2, f_1)$

$cov(f_1, f_2) = \frac{1}{n} \sum (x_{i1} - \mu(f_1))(x_{i2} - \mu(f_2))$

$\mu(f_1)$ $\quad$ $\mu(f_2)$

x_{i1}, x_{i2} → 2 feature of x_i

if data is col standardized

$cov(f_1, f_2) = \frac{1}{n} \sum (x_{i1} \cdot x_{i2})$

$\frac{1}{n} (f_1^T \cdot f_2)$

$\therefore$ $cvm = X^T \cdot X$ (if columns are standardized)
$d \times d \quad d \times n \quad n \times d$

$S \qquad X_{n \times d} \cdot X^T_{d \times n} \longrightarrow cvm_{d \times d}$

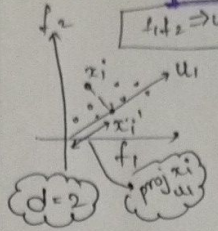
PCA

$a \cdot b = a^T b$

11

$f_1, f_2 \Rightarrow u_1$ $\quad$ u_1 → unit vector

x_i' → projection of x_i on u_1



$x_i' = \text{proj}_{u_1} x_i = \frac{u_1 \cdot x_i}{\|u_1\|} = u_1 \cdot x_i$ $\quad$ $- u_1^T \cdot x_i$

→ find $\vec{u_1}$ st var(proj x_i on u_1) is max

$var(x_i') = \frac{1}{n} \sum_{i=n} (u_1^T x_i - \cancel{u_1^T \cdot \bar{x}_i})^2$ → 0 (col std)

$\therefore$ PCA → maximize $\frac{1}{n} \sum u_1^T \cdot x_i^2$

VARIANCE MAXIMIZING APPROACH

Eigen val and vectors

$S_{d \times d}$ → Eigen values $\lambda_1, \lambda_2 \dots \lambda_d$
vectors $v_1, v_2 \dots v_d$

$v_i \perp v_j$
$v_i \cdot v_j = 0$

↳ $\lambda_1 V_1 = S \cdot v_1$, $\lambda_2 V_2 = S \cdot v_2 \dots$

finally $\vec{u_1} = \vec{v_1}$ → largest Eigen vector ↑ already proved

consider, d = 2

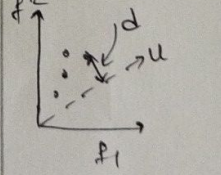
$\lambda_1 = 3$, $\lambda_2 = 0$ → λ_1 give 100%

$\lambda_1 = 3$, $\lambda_2 = 4$ → % = $\frac{3}{7}$

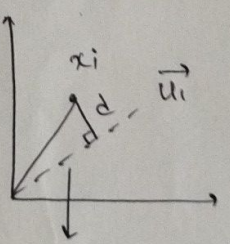
$\therefore$ λ_1 can represent $\frac{3}{7}$% of variance

%var = $\frac{\lambda_1}{\lambda_1 + \lambda_2}$

PCA - distance minimization



minimize $\sum_{i=1}^{n} d_i^2$



$d_i^2 = \|x_i\|^2 - u_1^T \cdot x_i^2$

$d_i^2 = x_i^T \cdot x_i - (u_1^T \cdot x_i)^2$

$\min \sum_{i=1}^{n} (x_i^T x_i - u_1^T x_i)$

$\text{proj}_{u_1} x_i = u_1^T \cdot x_i$

↳ pythagoras → $\|x_i\|^2 = d_i^2 + (u_1^T \cdot x_i)^2$