ML Fundamentals
-by Muthu

1. Algebra & Geometry
2. Calculus
3. PCA (more than dimensionality Reduction)

*one step to break above myth

Linear Algebra & Co-ordinate Geometry.

applications

matrix factorization → Recommadation sys.

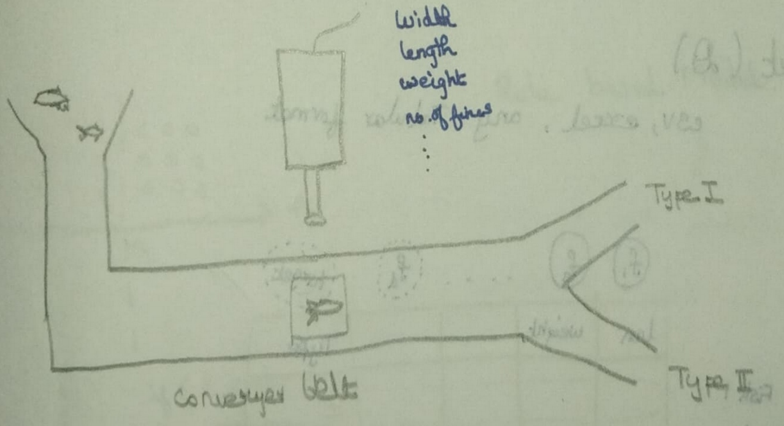
convolution → Computer vision, Deep Learning

Line
plane
vector
matrices
dot-product
circle, sphere, eclipse.

} cover perspective of ML.

Irodov book for JEE Advance.

eg. pattern classification
– Richard Duda hart



Scanner gives
↳ len, width, weight, ...

features — ML community
Axis — Math
independent variable. — statistics

Two o/p : Type 1 / Type 2.
(categorical)

predict age of fish using feature — regression

takes $\mathbb{R}^+$

→ classification : There is no natural ordering among classes.

eg.

we can't say Type 1 is better than type 2.

marks of student $\in \{1, 2, \ldots\ldots 100\}$

↳ multi-classification (Yes but) Regression is approp.

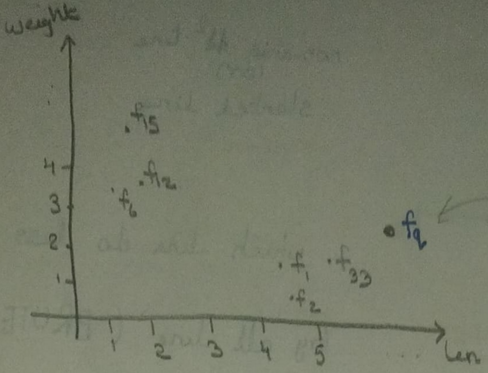
DataSet ($\mathcal{D}$)

csv, excel, any Tabular format

	f_1	f_2	....	f_d	target
	len	weight			Type.
Fish 1					II
Fish 2					II
⋮					
					II → Fish 15
Fish n					

$(n \times d)$

In graph, (scatter plot) hue = target



probably ∈ Type I

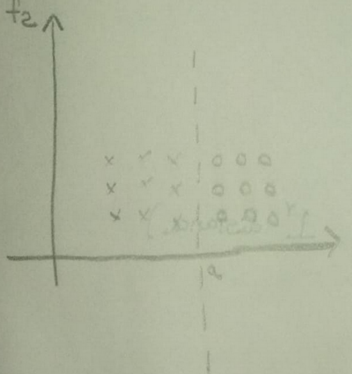
Train / Build a model?

separate point

↳ axis parallel line, slant line/tilted.

↳ curve

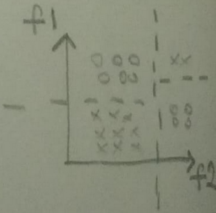
↳ shape (circle, eclipse)



Rule based Model.

if (f_1 < a):
  return I
else
  return II.

Set of axis $\|^{el}$ line

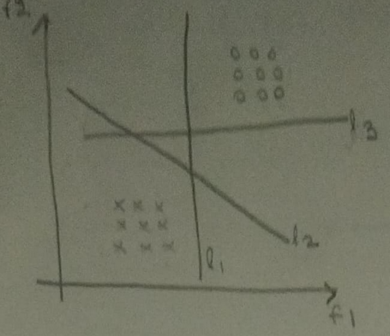


$\|^{ly}$,

fancier model,

DT
GBDT
RF

computationally Discover.

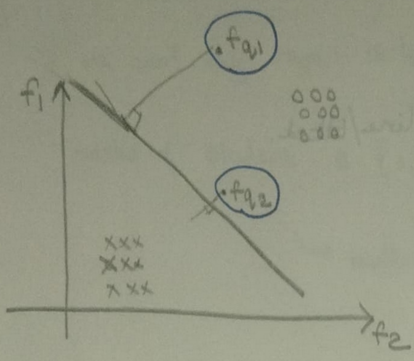


non-axis $\parallel^l$ line
(or)
slanted line

which line do less misclassification

try all line (BRUTE FORCE)

smart way to find
"Maxima & Minima" Calculus



Inference.

f_{q_1} → be Type II (99% confidence)

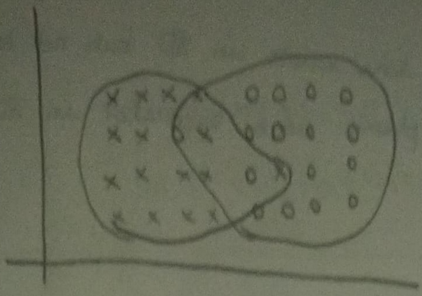
f_{q_2} → be Type II (25% confidence)

Model :

eqn. of line
distance b/w a point & line ($\perp^r$ distance)
which side it belongs to.

how to define err?

Pick all lines	# of misclassification
L_1	5
L_2	9
L_3 — best line —	0
L_4	2
L_5	10



Shapes

circle

Parabola

eclipse

(quadratic/conic)

close to canter more confident

boundary are less confident

Most of Research papers

set of axis-$\|^{l}$ line

slanted line

curved line

conic

- circle
- eclipse
- parabola
- ...

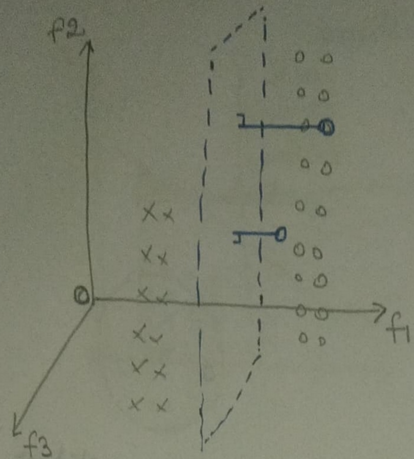
} overfit will occur

ML Models:

DF , RF , GBDT —— axis $\|^{l}$ lines

SVM, perceptron, logistic reg —— slanted line

KNN , DL , Naive Bayes (probabilistic models) ⟶ NOT finding hyperplane.



line exists in 3D but not separate

plane is a seperator in 3D

2D → line, circle, eclipse

3D → plane, sphere, ellipsoid

d D - hyperplane, hypersphere - - - -

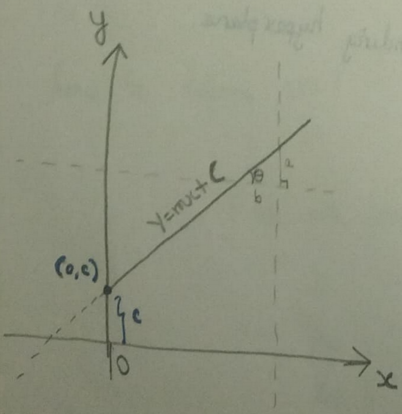
What if my independent variables are feature categorical?

soln:

we convert them into R^d (feature transforming)

one-hot encoding
target encoding
Label encoding

Lines



c → intercept

if x is 0 then y is at C

$Y = mx + C$

General form of line,

$ax + by + C = 0$

$m = \frac{-a}{b}$ $\qquad$ $m = \tan\theta = \frac{y}{x}$ $\qquad$ a $\parallel^{l}$ Y-axis

b $\parallel^{l}$ X-axis

$ax + by + c = 0$

$w_1x_1 + w_2x_2 + w_0 = 0$

In d-dimension, $\quad W_1x_1 + W_2x_2 + W_3x_3 + \ldots\ldots + W_dx_d + W_0 = 0$

Vector notation, $\quad \sum_{i=1}^{d} W_i x_i + w_0 = 0$

In general,

Vector Represent Column Matrix

$$A = \begin{bmatrix} a_1 \\ a_2 \\ a_3 \end{bmatrix} \qquad A^T = \begin{bmatrix} a_1 & a_2 & a_3 \end{bmatrix}$$

dot product

$$\begin{bmatrix} w_1 & w_2 & \cdots & w_d \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix} + w_0 = 0$$

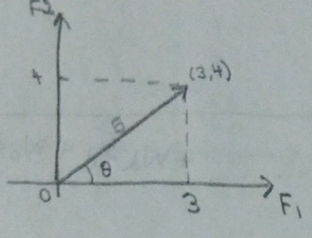
Matrix notation, $\quad \boxed{W^T X + w_0 = 0}$

$W \in \mathbb{R}^d$ d-dim Real value no.

$X \in \mathbb{R}^d$

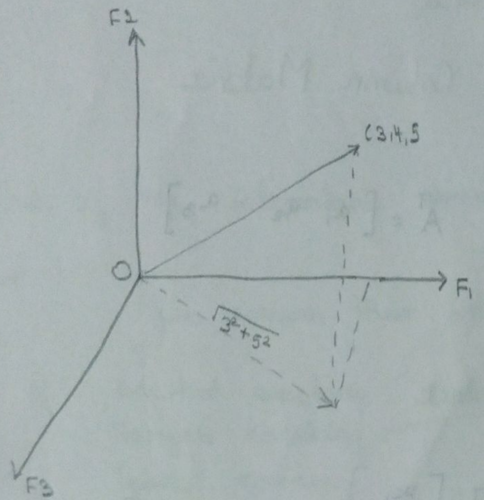
$w_0 \in \mathbb{R}$

pythagores theorem,



magnitude/distance from origin

$$= \sqrt{(y_2 - \underset{0}{y_1})^2 + (x_2 - \underset{0}{x_1})^2}$$



$$d = \sqrt{(y_2-y_1)^2 + (x_2-x_1)^2 + (z_2-z_1)^2}$$

In ML context,

magnitude of vector = $\| f_q \|$

Norm of vector.

DataSet Notation

	F_1	F_2	......	F_d	Type
n_1					y_1
n_2					y_2
.					
n_i^{th}	←	x_i	→		y_i
n					y_n

$x_i \in \mathbb{R}^d$

$y_i \in \{1, 2\}$

DataSet Representation

$$\mathcal{D} = \left\{ (x_i, y_i)_{i=1}^{n} \; ; \; x_i \in \mathbb{R}^d \; ; \; y_i \in \{1,2\} \right\}$$

eqn. of hyperplane, $W^T X + w_0 = 0$ $\quad W \in \mathbb{R}^d \; ; \; X \in \mathbb{R}^d \; ; \; w_0 \in \mathbb{R}$

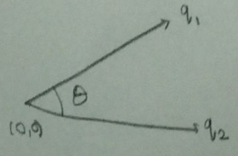
Agenda.

- eqn. of line/plane
- half spaces
- how far are we from Π_d / confidence.

angle b/w two vector. → dot product

~~cross product~~ not much in ML

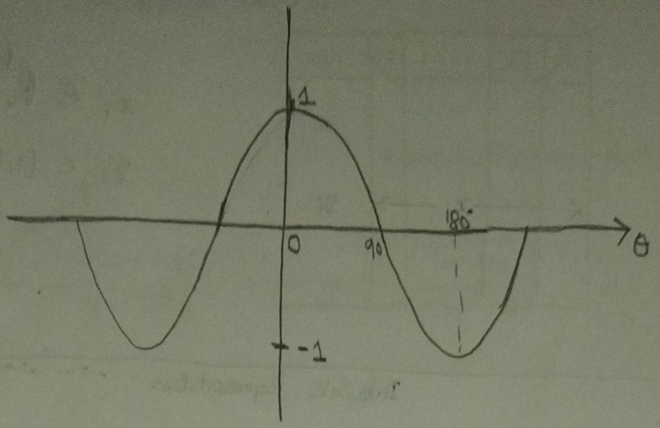
w_0 in geometry.



angle b/w two vector q_1 & q_2 = $q_1^T \cdot q$

$= |q_1| |q_2| \cos\theta$

$$\theta = \cos^{-1}\left(\frac{q_1^T \cdot q_2}{\|q_1\| \|q_2\|}\right)$$



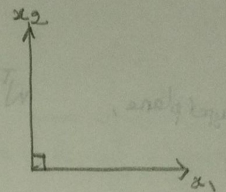
If $x_1 \cdot x_2 = 0$! $\boxed{\theta = ?}$

$\|x_1\| \neq 0$

$\|x_2\| \neq 0$

$\cos\theta = 0$

$\theta = 90^\circ$



orthogonal vectors

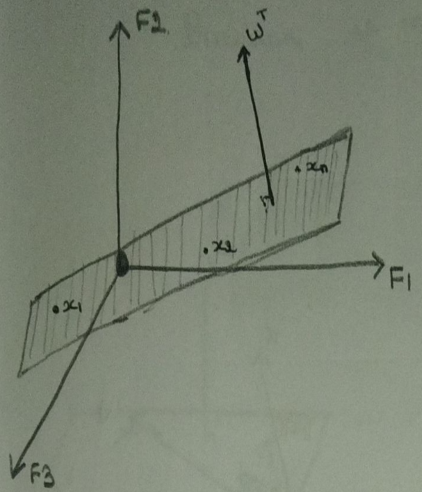
① plane passes through origin

plane not passes through origin

② half space

③ distance b/w query points & plane.

plane passes through origin



$$W^T x + w_0 = 0$$

$$\boxed{\begin{aligned} W^T x &= 0 \qquad (w_0 = 0) \\ \|W^T\| \|x\| \cos\theta &= 0 \\ W^T &\perp^r x. \end{aligned}}$$

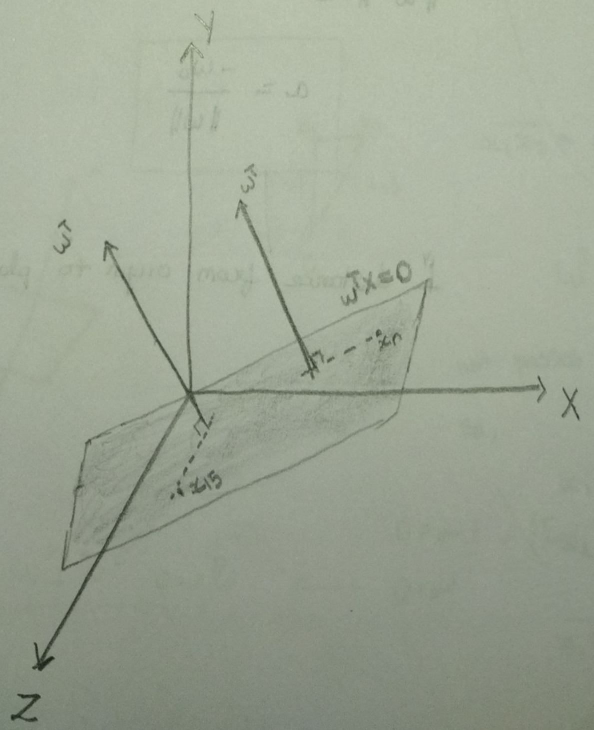
where,
x represents any points on plane.

$$\boxed{\begin{gathered} \text{assume/ put } x = 0 \text{ in } W^T x + w_0 = 0 \\ \cancel{[w_1, w_2 \ldots w_n] \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}} + w_0 = 0 \\ w_0 = 0 \end{gathered}}$$

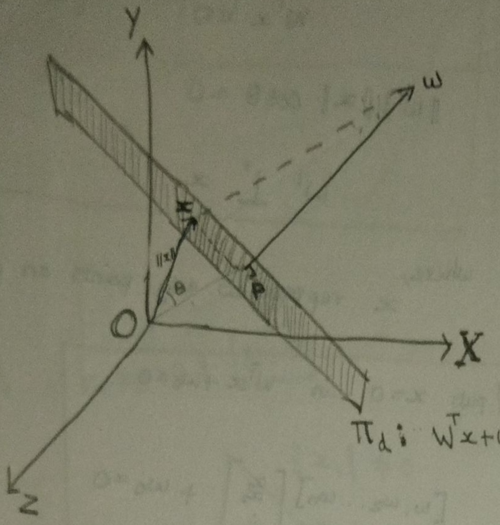
Inference.

So, if plane pass through origin $W^T X = 0$; $W^T \perp^r X$

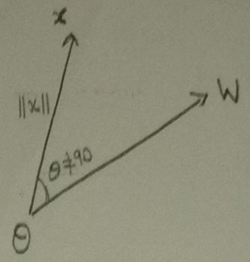
X be random point which is on plane.

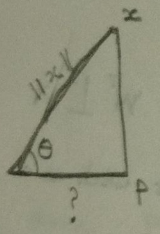


plane passes not through origin



$\pi_d : w^T x + w_0 = 0$





$\cos\theta = \frac{?}{||x||}$

$? = ||x|| \cos\theta$

$w^T x + w_0 = 0$

$||w^T|| \, ||x|| \cos\theta + w_0 = 0$

$||w^T|| \, a + w_0 = 0$

$$\boxed{a = \frac{-w_0}{||w||}}$$

⊥ distance from origin to plane. (distance not -ve)

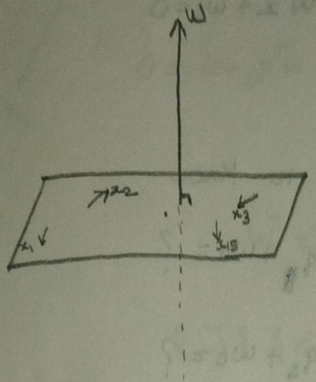
if $a = 0$

$\cancel{||w^T|| a} + w_0 = 0$

$w_0 = 0 \longrightarrow w^T x = 0 \quad w^T \perp x$ (passes through origin)

By convention we assume $\|w\| = 1$ $\hat{w}$ as unit vector

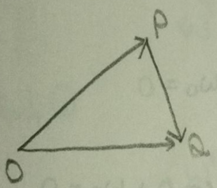
Direction is more important.



To represent plane we need Normal vector

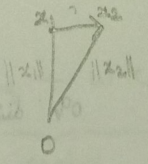
bcz, so many $\parallel^{l}$ vector we can draw (leads to ambiguity)

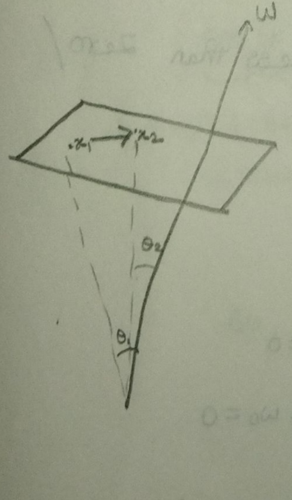
Vector addition



$\vec{OP} + \vec{PQ} = \vec{OQ}$

$\vec{PQ} = \vec{OQ} - \vec{OP}$





$\vec{x_1 x_2} = \vec{O x_2} - \vec{O x_1}$

$w^T \perp^r \pi_d$

not passes origin,

so,

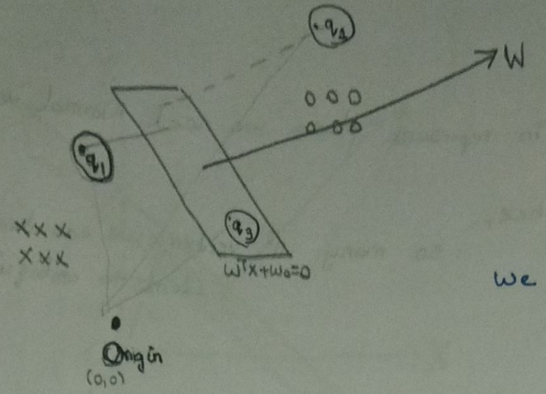
$x_1 \perp^r w$ NO.

$x_2 \perp^r w$ NO.

$\vec{x_1 x_2} \perp^r w$ NO.

distance from origin to π_d : $\frac{w_0}{\|w\|}$

Half Space.



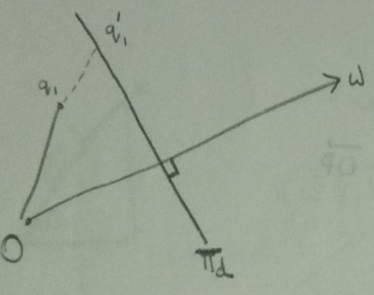
$w^T x + w_0 = 0$

so $w^T q_3 + w_0 = 0$

we want to know

$w^T q_1 + w_0 = ?$

$w^T q_2 + w_0 = ?$



$w^T x'_{q_1} + w_0 = 0$

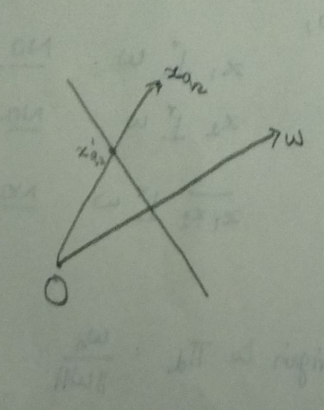
$\|w^T\| \|x'_{q_1}\| \cos\theta + w_0 = 0$

$\|w\| \|\vec{oq_1} + \vec{q_1q'_1}\| \cos\theta + w_0 = 0$

$|\vec{oq_1}| < \|oq'_1\|$

(0)

oq_1 distance less than zero/



$w^T x'_{q_2} + w_0 = 0$

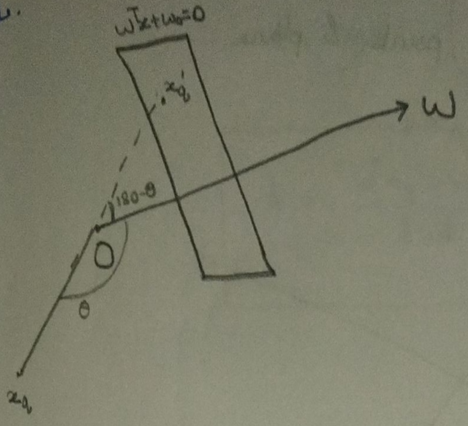
$\|w^T\| \|x'_{q_2}\| \cos\theta + w_0 = 0$

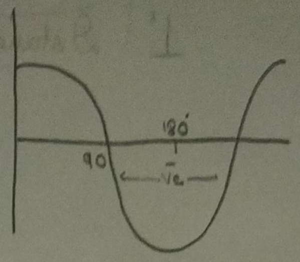
$\|ox'_{q_2}\| < \|o x_{q_2}\|$

(0)

ox_{q_2} distance greater than zero/!

Q.





$(\theta > 90°)$ & $(\theta < 270°) \rightarrow$ -ve value

$\cos(180-\theta) = -\cos\theta$

$$W^T x_q' + w_0 = 0$$

$$\|\vec{w}\| \, \|x_q'\| \cos(180-\theta) + w_0 = 0$$

$$-w \, x_q' \cos\theta + w_0 = 0$$

$$w_0 = w \cdot x_q' \cos\theta \quad \text{--- (1)}$$

$$W^T x_q + w_0$$

Sub (1)

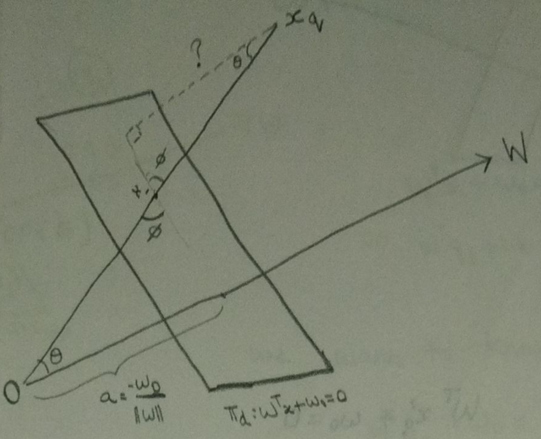
$$w^T x_q + w \cdot x_q' \cos\theta$$

$$\|w\| \, \|x_q\| \cos\theta + w \cdot x_q' \cos\theta$$

$$\underbrace{\cos\theta}_{\substack{\text{if } (\theta > 90°) \,\&\, (\theta < 270°) \\ \text{-ve value}}} \; (\underbrace{w x_q + w x_q'}_{\text{scaler}})$$

so, $W^T x_q + w_0 < 0$, if $\theta < 270°$ & $\theta > 90°$

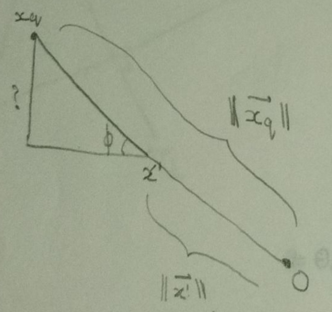
$\perp^r$ Distance b/w query point & plane

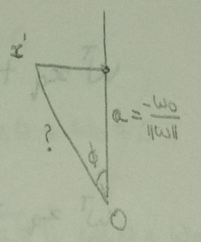


given,

$\perp^r$ distance from origin to plane : $\frac{-w_0}{\|w\|}$

plane not passes through origin, $\cancel{\vec{x_q} \perp^r \vec{w}^T}$ NOT





$$\cos\theta = \frac{a}{\|\overrightarrow{ox'}\|}$$

$$\|\overrightarrow{ox'}\| = a / \cos\theta$$

$$\|\overrightarrow{ox'}\| = \frac{-w_0}{\|w\| \cos\theta}$$

$$\|\overrightarrow{x' x_q}\| = \|\overrightarrow{ox_q}\| - \|\overrightarrow{ox'}\|$$

$$\|\overrightarrow{x' x_q}\| = x_q + \frac{w_0}{\|w\| \cos\theta}$$

$$\cos\theta = \frac{d}{\|\overrightarrow{x' x_q}\|}$$

$$d = \cos\theta \cdot \left(x_q + \frac{w_0}{\|w\| \cos\theta} \right)$$

$$w^T x_q$$

$$d = \cancel{\cos\theta}\left(\frac{\|w\|\, x_q \cos\theta + w_0}{\|w\| \cancel{\cos\theta}}\right)$$

$$\boxed{d = \frac{w^T x_q + w_0}{\|w\|}}$$

Recap.

eqn of plane. $w^T x + w_0 = 0$

pass through origin $\quad w^T x = 0 \longrightarrow w^T \perp^r x \;-\; w_0 = 0$

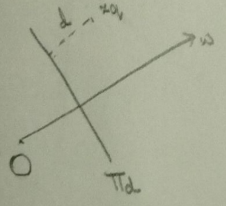
not pass through origin $\quad a = \frac{-w_0}{\|w\|}$

point lie on the plane. $\quad w^T x_i + w_0 = 0$

either side, $\quad w^T x_i + w_0 < 0 \;/\; w^T x_i + w_0 > 0$

orthogonal distance from pnt to separator plan

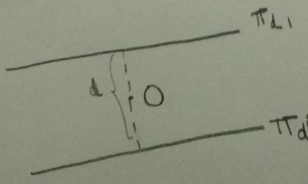
$$d = \frac{w^T x_q + w_0}{\|w\|}$$



distance b/w 2 plane: $\frac{w_1 - (-w_2)}{\|w\|}$

$$= \frac{w_1 + w_2}{\|w\|}$$

Q
SVM



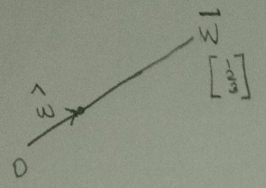
what if 3-types of fish

Multi-classification

→ 2 hyperplane

→ ONE Vs rest

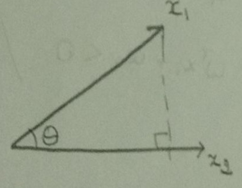
Unit Vector



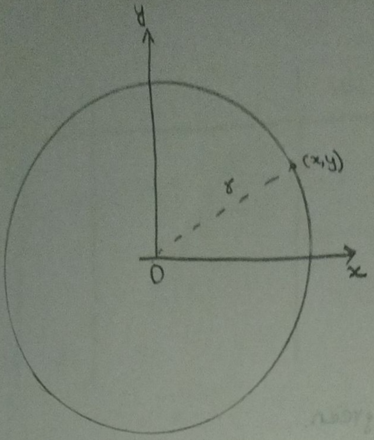
$$\|w^T\| = \sqrt{1^2+2^2+3^2} = \sqrt{14}$$

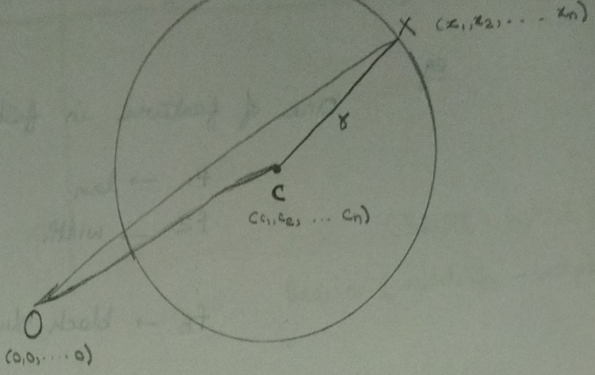
$$\hat{w} = \begin{bmatrix} 1/\sqrt{14} \\ 2/\sqrt{14} \\ 3/\sqrt{14} \end{bmatrix}$$

projection



$Proj_{x_2}(x_1)$





To define hypersphere,

→ distance O to C

→ radius

$$\| \vec{cx} \| = r$$

$$\| \vec{ox} \| - \| \vec{oc} \| = r$$

$$\left\| \begin{matrix} x_1 - c_1 \\ x_2 - c_2 \\ \vdots \\ x_n - c_n \end{matrix} \right\| = r$$

$$\sqrt{(x_1 - c_1)^2 + \dots + (x_n - c_n)^2} = r$$

$$\boxed{\sum_{i=1}^{n} (x_i - c_i)^2 = r^2}$$

$$\sum_{i=1}^{n} (x_i - c_i)^2 - r^2 < 0$$

pnt inside sphere

$$\sum_{i=1}^{n} (x_i - c_i)^2 - r^2 = 0$$

pnt on hypersphere

$$\sum_{i=1}^{n} (x_i - c_i)^2 - r^2 > 0$$

pnt outside hypersphere

ONE HOT ENCODING

for categorical variable,

eg.

One of feature in fish.

$f_1 \rightarrow$ len

$f_2 \rightarrow$ width

:

$f_5 \rightarrow$ black, blue, green.

we can't assign,

black — 0

blue → 1

green — 2

X

if low, medium, high (ordinal

low → 0 10

medium → 1 20

high → 10 100

(Domain Expert)

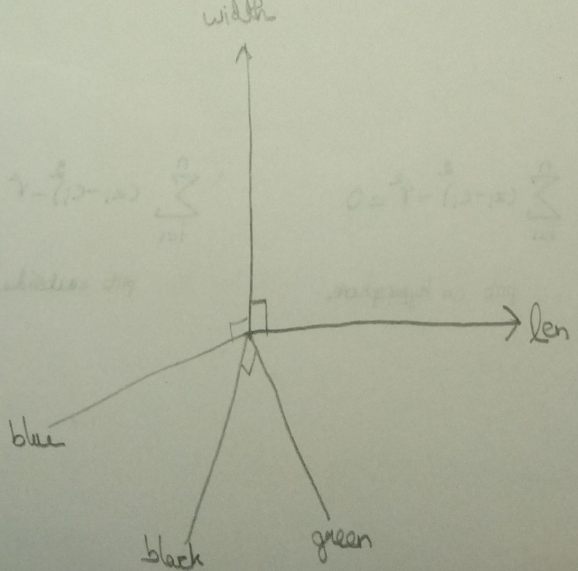
(Rank base-encoding)

fish1 → 12, 5, blue.

fish2 → 9, 2, green

len	width	black	blue	green
12	5	0	1	0
9	2	0	0	1

(Sparse matrix)



3D → 5D

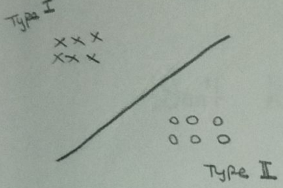
Find Best Hyperplane

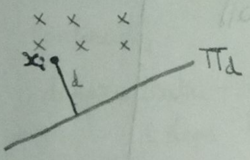
w_1	w_2	. . . .	w_d	w_0	# of misclassification
$\vdots$	$\vdots$		$\vdots$	$\vdots$	$\vdots$
∞	∞			∞	

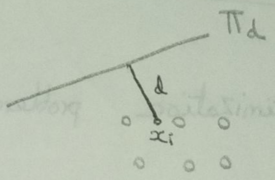
(BRUTE - FORCE)

less no. of misclassify → best plane

$$\mathcal{D} = (x_i, y_i)_{i=1}^{n} \; ; \; x_i \in \mathbb{R}^d \; ; \; y_i \in \{-1, 1\}$$







distance from x_i to π_d = $\dfrac{w^T x_i + w_0}{\|w\|}$

$d = \dfrac{w^T x_i + w_0}{\|w\|}$

$y_i = -1$

Label $y_i = +1$

we want both far as possible.

$\dfrac{w^T x_i + w_0}{\|w\|} \cdot y_i$

(distance MAX)

$\dfrac{w^T x_i + w_0}{\|w\|} \cdot y_i$

distance is in -ve

Label -1

$-1 \times -1 = (+1)$

(distance MAX)

for each point x_i,

$$z_i = \frac{w^T x_i + w_0}{\|w\|} \cdot y_i$$

$$\frac{w^T x_i + w_0}{\|w\|} \cdot y_i$$

w_1	w_2	...	w_d	w_0	y_i	z_i

we want z_i will max

math trick,

$$\max_{\substack{w, w_0 \\ d+1 \text{ dim}}} (z_1 + z_2 + \cdots + z_n) = \max_{w_0, w} \sum_{i=1}^{n} z_i$$

$$= \max_{w_0, w} \sum_{i=1}^{n} y_i \frac{w^T x_i + w_0}{\|w\|}$$

why sum? not product?

if any one zero $\longrightarrow$ collapsed.
(any pnt lie on plane)

optimization problem. $\max_{w, w_0} \sum_{i=1}^{n} y_i \frac{w^T x_i + w_0}{\|w\|}$

$\max_{w, w_0} f(w, w_0)$

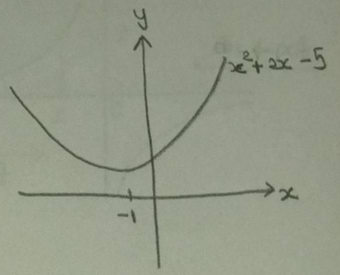
CALCULUS

11th & 12th math

$y = f(x) = x^2 + 2x - 5$

find min,

$y' = f'(x) = 2x + 2$



$y' = 0$

$2x + 2 = 0$

$\boxed{x = -1}$

it's minima

$y'' = f''(x) = 2 > 0$

Don't know why we are studying :(

Detour

functions & limits

Continuity

differentiate - one variable
(slope & geometry)

Maxima & minima

partial derivative

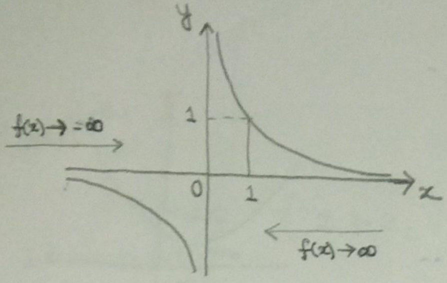
Gradient Descent

SGD

Constraints Optimization

Functions & Limits

plot (1/x)



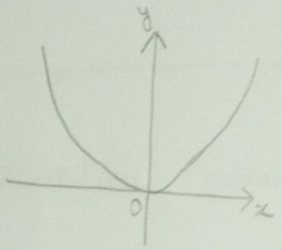
one-side limit

as move $x^{+ve} \longrightarrow 0 \Rightarrow f(x) = \infty$

as move $x^{-ve} \longrightarrow \quad \Rightarrow f(x) = -\infty$

$$\boxed{\begin{array}{l} \lim_{x \to 0^+} \frac{1}{x} = \infty \\ \lim_{x \to 0^-} \frac{1}{x} = -\infty \end{array}}$$

plot (x^2)



$\lim_{x \to 0^+} x^2 = 0$; $\lim_{x \to 0^-} x^2 = 0$

if both side value are same "Two side limit"

$$\lim_{x \to 0} x^2 = 0$$

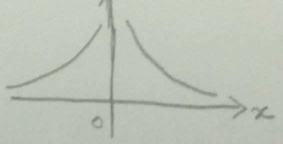
Q.

$\lim_{x \to 0} \frac{1}{x}$ = undefined (2-side limit not exists)

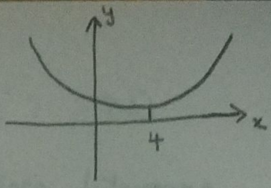
(exists at point x@1)

$\lim_{x \to 1} \frac{1}{x} = 1$

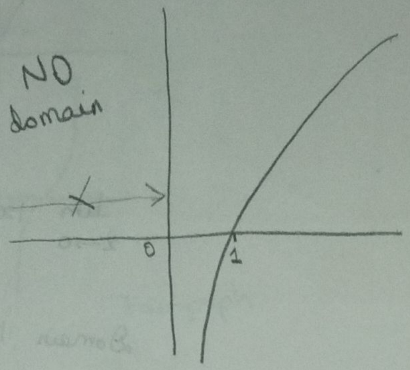
$\lim_{x \to 0} \left| \frac{1}{x} \right| = \infty$



$\lim_{x \to 1} x^2 - 2x + 5 = 4$



2-side Limit
No-discontinuity



$\lim_{x \to 0} \log(x)$

$\lim_{x \to 0^+} \log(x) = \infty$

$\log(-10)$ = undefined

$\lim_{x \to 0^-} \log(x)$ = undefined

Domain : All values X can take

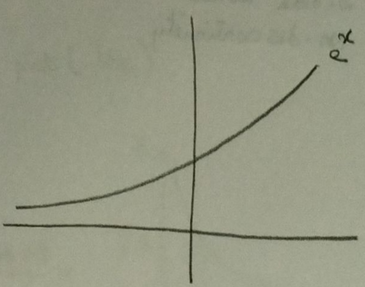
Range : All values Y can take

So,

Domain of $\log(x) = R^+$ (∵ exclude 0)

0 neither +ve nor -ve.

exponential graph



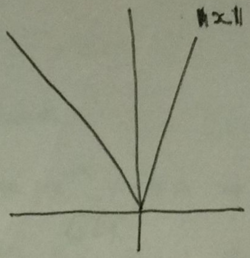
$\lim_{x \to \infty} e^x = \infty$

$\lim_{x \to -\infty} e^x = 0$

Domain : $\mathbb{R}$

Range : $\mathbb{R}^+$

absolute graph

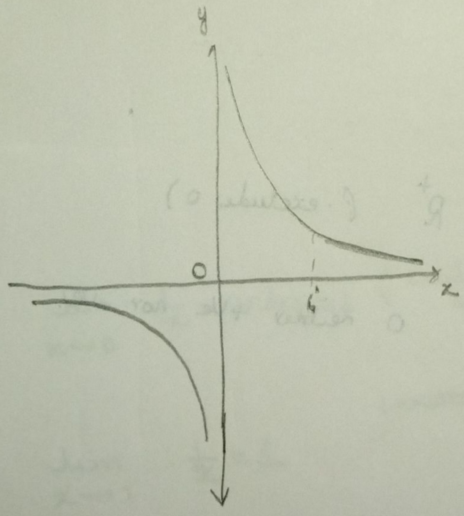


$\lim_{x \to 0} |x| = 0$

Domain $|x| = \mathbb{R}$

Range $|x| = \mathbb{R}^+_{+0}$

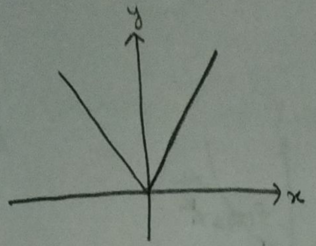
$[0, \infty)$

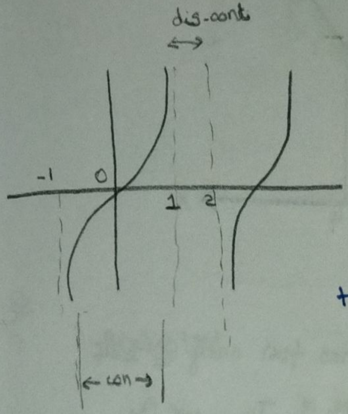


$\frac{1}{x}$ discontinuity @ Zero

$\frac{1}{x}$ non-discontinuity @ $x=6$

is |x| discontinous ? NO





tan-graph.

definition

f(x) is continous @ $z=a$

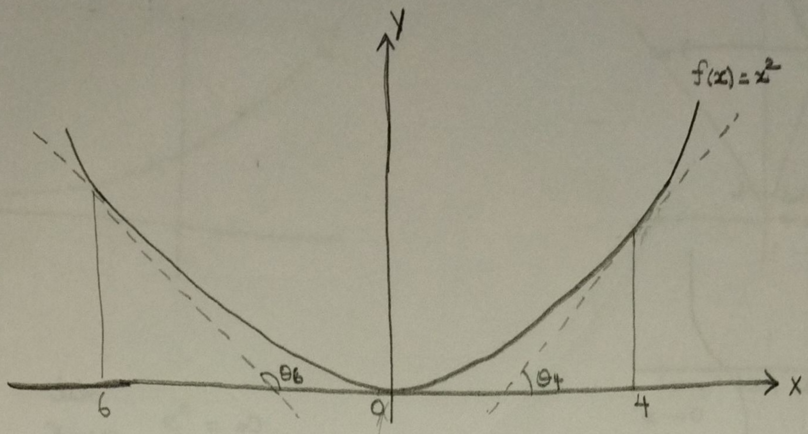
if $\lim_{x \to a^-} f(x) = \lim_{x \to a^+} f(x) = f(a)$

why continuity ?

Concept we cover Maxima / Minima

if graph breaks $\longrightarrow$ maxima/minima breaks.

Derivative



tangent

slope = 0

touches f(x) only @ x=4

slope of $T_4 = \tan\theta_4$

$$\left.\frac{df(x)}{dx}\right|_{x=\theta_6} = \tan\theta_6$$

$90^\circ < \theta_6 < 180^\circ$

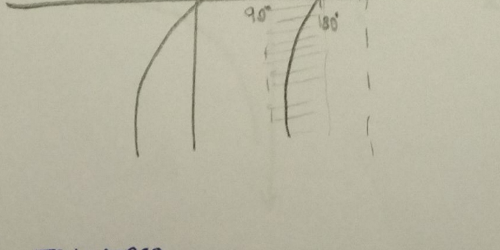
(−ve value)

$$\left.\frac{d\,f(x)}{dx}\right|_{x=4} = \tan\theta_4$$

$0^\circ < \theta_4 < 90^\circ$

(+ve value)

$\tan\theta_0 = 0$



$\tan\theta$

$0 < \theta < 90 \rightarrow$ inc curve

$90 \rightarrow$ zero

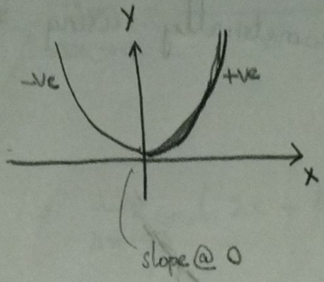
$90 < \theta < 180^\circ \rightarrow$ dec curve

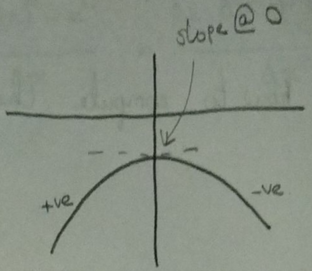
Inference

At T_4 curve is increasing $\rightarrow$ slope (+ve)

T_0 curve is flat $\rightarrow 0$

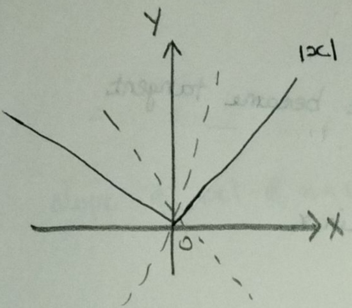
T_6 curve is decreasing $\rightarrow$ slope (−ve) eg





Q.

$f(x) = |x|$



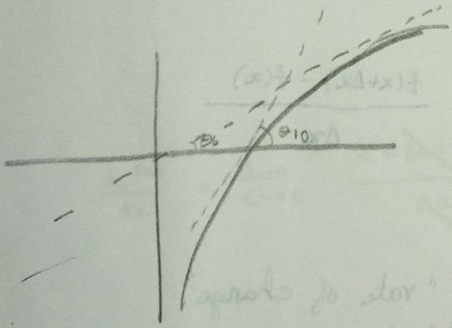
" tangent has to be unique "

slope @ $x=-1$ $\longrightarrow$ -ve

slope @ $x=1$ $\longrightarrow$ +ve

slope @ $x=0$ $\searrow$ multiple slope (or) not defined

log(x)

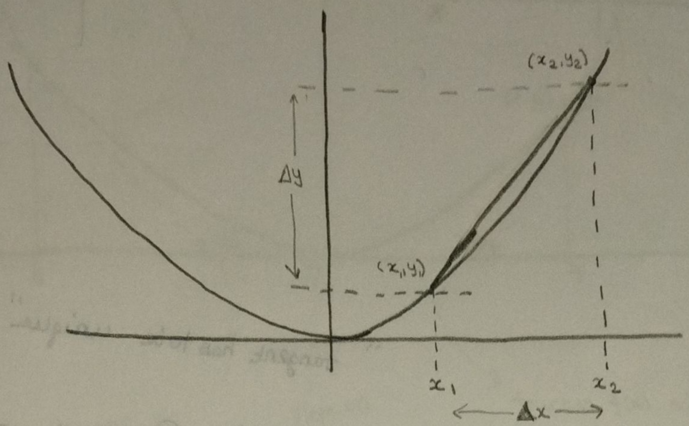


slope @ $x=6$ $<$ slope @ $x=10$

magnitude tells how fast grow/dec.

Agenda

how to compute the slope geometrically using secant?



x_2 closer to x_1 $\longrightarrow$ $\Delta x \to 0$ $\longrightarrow$ secant became tangent

dx

$\hookrightarrow$ small change in x.

$$\text{slope of tangent} = \frac{\Delta y}{\Delta x}$$

$$= \frac{f(x+\Delta x) - f(x)}{(x+\Delta x) - x}$$

$$= \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x}$$

Secant measures $\rightarrow$ "rate of change" eg.

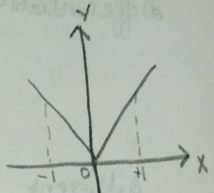
$$\frac{dx^2}{dx} = \lim_{\Delta x \to 0} \frac{f(x+\Delta x) - f(x)}{\Delta x} = \frac{(x+\Delta x)^2 - x^2}{\Delta x} = \frac{2x \cdot \cancel{\Delta x} + \Delta x^{\cancel{2}}}{\cancel{\Delta x}}$$

$$= \lim_{\Delta x \to 0} \left(2x + \underbrace{\Delta x}_{0}\right)$$

$$\frac{dx^2}{dx} = 2x$$

slope of x^2 @ $x=4$ $\quad \left.\frac{dx^2}{dx}\right|_{x=4} = 2(x=4) = 8$

slope of $|x|$ @ $x=1$ $\quad \left.\frac{d|x|}{dx}\right|_{x=1} = 1$



$$\left.\frac{d|x|}{dx}\right|_{x=-1} = -1$$

tan45° = 1

$$|x| = \begin{cases} +x & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ -x & \text{if } x > 0 \end{cases}$$

$$\frac{d|x|}{dx}' = \lim_{x \to 0} \frac{|x + \cancel{\Delta x}| - |\cancel{x}|}{\cancel{\Delta x}} = \lim_{x \to 0} 1 = 1.$$

$$\frac{d|x|}{dx} = \lim_{x \to 0} \frac{-|x+\Delta x| - (-|x|)}{\Delta x} = \frac{-\cancel{|x|} - |\cancel{\Delta x}| + \cancel{|x|}}{\cancel{\Delta x}} = -1.$$

common derivatives

$$\frac{d}{dx} x^n = nx^{n-1} \quad n \neq 0 \quad n \in R$$

$$\frac{d}{dx} \sqrt{x} = \frac{d}{dx} x^{1/2} = \frac{1}{2} x^{-1/2}$$

$$\frac{d}{dx} \log x = \frac{1}{x}$$

$$\frac{d}{dx} e^x = e^x$$

Differentiability

if left-side limit and right-side limit are different at point a then f(x) not-differentiability @ $x = a$

geometrically,

contin & smoothing

Jeff Dean

→ distribute System

code:

```
def func(x):
    return x^3 + x^2 - 5


def compute_differentiate(a)

    Δx = 0.00001

    res = (func(a + Δx) - func(a)) / Δx

    return res
```

Notation

$$\frac{d}{dx} f(x) = \frac{df}{dx} = f' = y' = \frac{dy}{dx} = \dot{y}$$

rules

Sum rule: $\frac{d}{dx}\{ f(x) + g(x) \} = f' + g'$

$$\frac{dc}{dx} = 0$$

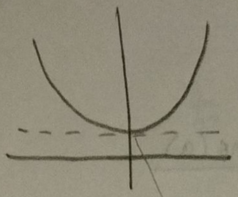
$$\frac{d}{dx}\{ f(x) \cdot g(x) \} = f g' + g f'$$

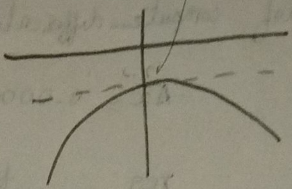
$$\frac{d}{dx} f(g(x)) = f'(g(x)) \cdot g'(x)$$

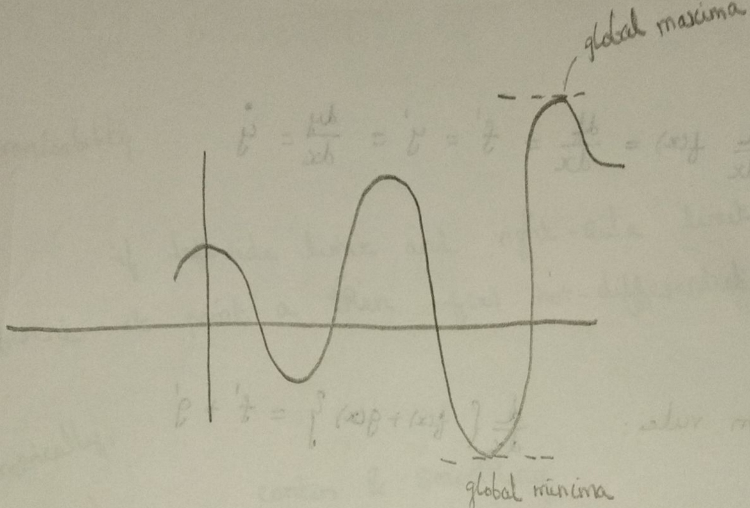
$$\frac{d}{dx}\left\{ \frac{f(x)}{g(x)} \right\} = \frac{f' g - g' f}{(g')^2}$$

Maxima & Minima.

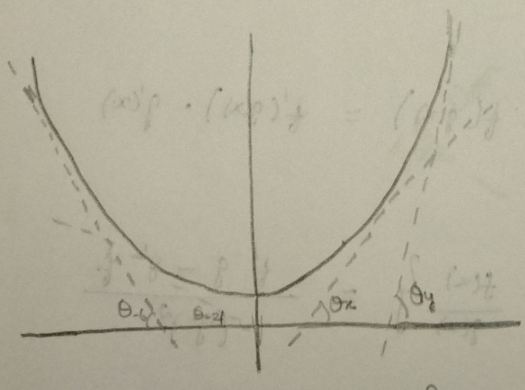
First non-zero higher order derivative $\begin{cases} \nearrow +ve \\ \searrow -ve \end{cases}$







how to get minima



if 2^{nd} derivative (+ve)

we get local minima.

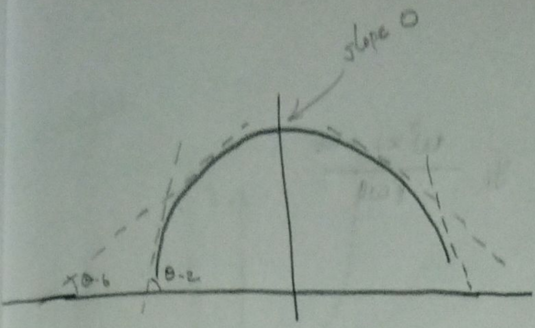
$\theta_x < \theta_y$

$\theta_{-6} > \theta_2$

big $\longrightarrow$ small   small $\longrightarrow$ big

-ve   0   +ve

$\longrightarrow$



small ——→ big

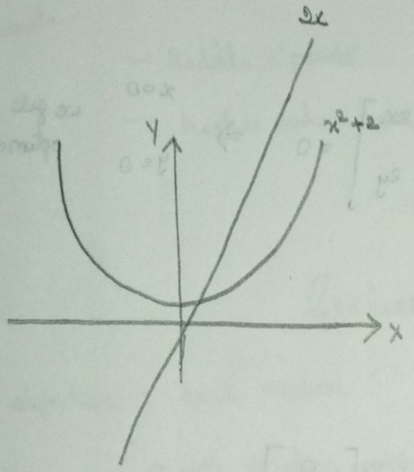
$\theta_{-6} < \theta_{-2}$

big ——→ small

$\theta_4 > \theta_{10}$

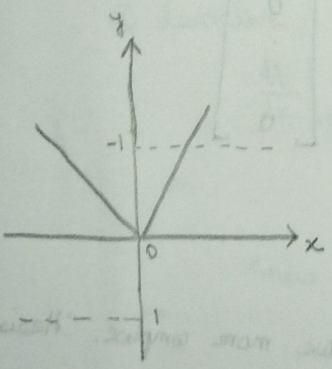
+ve ——→ −ve

if 2nd derivative (−ve) we get local maxima.



$y = x^2+2$

$y' = 2x$



if $\left.\frac{df}{dx}\right|_{x=x_1} = 0$ then we get optima / saddle point

$$f(w, w_0) = \max_{w, w_0} \sum_{i=1}^{n} y_i \frac{w^T x_i + w_0}{\|w\|}$$

(d+1) function

$$= \min_{w, w_0} \sum_{i=1}^{n} -y_i \frac{w^T x_i + w_0}{\|w\|}$$

Partial derivative

eg.

$$f(x) = x^2 + y^2$$

each variable find optima independently

$$\nabla f = \begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \end{bmatrix} = \begin{bmatrix} 2x \\ 2y \end{bmatrix} = 0 \Rightarrow \begin{matrix} x = 0 \\ y = 0 \end{matrix}$$

we get optima.

11ly

$$\nabla f_{(w, w_0)} = \begin{bmatrix} \partial f/\partial w_0 \\ \partial f/\partial w_1 \\ \vdots \\ \partial f/\partial w_d \end{bmatrix}_{d+1} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

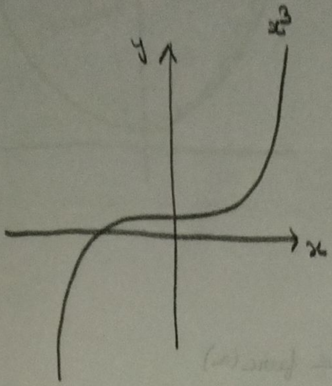
we won't go for multi / 2nd order derivative more complex. "Hesian"

for one variable it's easy.

2nd order.

$$\begin{bmatrix} \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right) & \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right) \\ \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial y}\right) & \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial y}\right) \end{bmatrix}$$

exception



$f(x) = x^3$

$f' = 3x^2 = 0 \Rightarrow x = 0$ (optima)

$f'' = 6x = 0 \Rightarrow$ neither +ve / -ve

(saddle point)

$f = (x^2 - y^2)$

handle

- saddle points
- higher order derivation

} → computation hacks/ algo.

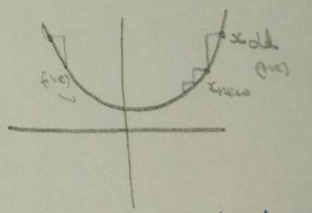
Gradient Descent

objective : find minima

① pick random point

② derivative

$\left.\frac{df}{dx}\right|_{x=x_{old}} = +ve$ [inc func]



update x in -ve direction

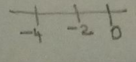
$$x_{new} = x_{old} + \eta \left(\frac{-df}{dx}\right)_{x=x_{old}}.$$

if derivative -ve.

$\left.\frac{df}{dx}\right|_{x_{old}} = -ve$ (dec func)

$$x_{new} = x_{old} + \eta\left(-\frac{df}{dx}\right)$$

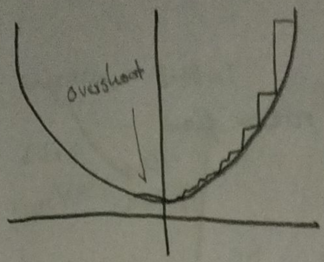
$- \times -1 = +ve$



-4
$+2$
-2

code

```
def func(x):
        return x^3+x^2 - 5


def compute_differentiate(a)
        delX = 0.00001
        res = (func(a+delx) - func(a)) / delx

        return res


def gradient_descent():
        x_old = random.random()

        i = 0
        eta = 0.01
        x_i = x_old

        df_dx = compute_defferentiate(x_i)

        while ( df_dx - 0.0001 > τ ):      threshold
                print(i)
                print(x_i)
                print(df_dx)
                print('****-----****')

                df_dx = compute_defferiate(x_i)

                x_i = x_i - (eta * df_dx)

                i = i+1

        return x_i
```



'e' ↑ iteration ↓

Multivarient GD

$z = f(x, y)$

→ init x_0, y_0 randomly.

→ while $\frac{df}{dx}\Big| = 0.0001$ and $\frac{df}{dy}\Big| = 0.0001$

update

$$x = x_0 - \eta \left.\frac{df}{dx}\right|_{x_0}$$

$$y = y_0 - \eta \left.\frac{df}{dy}\right|_{y_0}$$

→ we get x^*, y^*

Q. how to check its maxima/minima.

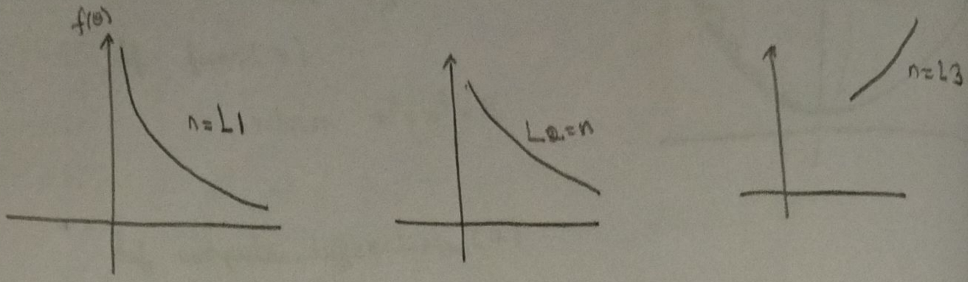
if $(x^*, y^*) = 10$

$f(x^* \pm 0.1, y^* \pm 0.1)$ < 9 ⇒ x^*, y^* can't be minima; 12 ⇒ x^*, y^* can't be maxima

function should be,

differentiable

continuous



$$\boxed{L2 < L1 < L3}$$

$$J_{\omega, w_0} = \min_{\omega, w_0} \sum_{i=1}^{n} -y_i \frac{\omega^T x + w_0}{\|\omega\|} \quad \mathcal{L}_i$$

$$= \min_{\omega, w_0} \sum_{i=1}^{n} \mathcal{L}_i$$

$$\nabla_{\omega, w_0} \mathcal{L} = \begin{bmatrix} \partial L / \partial w_0 \\ \partial L / \partial w_1 \\ \vdots \\ \partial L / \partial w_d \end{bmatrix}$$

each will update independently

$$\mathcal{L} = -\sum_{i=1}^{n} y_i \frac{\omega^T x + w_0}{\|\omega\|}$$

--- unconstrained problem.

Constrained problem.

enforce, $\|w\| = 1$

unit vector

direction NOT change.

$$\mathcal{L} = \min_{w, w_0} \sum_{i=1}^{n} y_i \frac{w^T x_i + w_0}{1} \qquad \text{s.t } \|w\| = 1$$

equality constraint

unconstrained $\longrightarrow$ constrained.

(Introduce Lagrange multiplier)

$$\min_{x,y} f(x,y) \quad \text{s.t } g(x,y) = C \qquad \approx \qquad \min_{x,y} f(x,y) + \lambda \{ g(x,y) - C \}$$

x^*, y^* same x^*, y^*

$$\min_{w, w_0} \sum_{i=1}^{n} -y_i (w^T x_i + w_0) \quad \text{s.t } \|w\|^2 = 1 \qquad \approx \qquad \min_{w, w_0} \sum_{i=1}^{n} -y_i (w^T x_i + w_0) + \lambda (\|w^2\| - 1)$$

solve

using GD find best w_i^*, w_0^*

misclassification

(using w_0, w)

hyper parameter search

$\lambda = 0 \longrightarrow w_0^*, w^* \longrightarrow \pi_1 \longrightarrow$

$\lambda = 1 \longrightarrow w_0^*, w^* \longrightarrow \pi_2 \longrightarrow$

$\lambda = 10 \longrightarrow$

.

.

.

solving

$x^2 + y^2 - c$

$x^2 + y^2$

(same)

$$\min_{w, w_0} \sum_{i=1}^{n} -y_i \left(w^T x_i + w_0\right) + \lambda \underbrace{\|w^2\| - 1}_{\text{constant}}$$

$$\min_{w, w_0} \sum_{i=1}^{n} -y_i \left(w_0 + w_1 x_1 + \dots + w_j x_j + \dots + w_d x_d\right) +$$

$$\underbrace{\lambda}_{\text{assume fixed}} \left(w_1^2 + w_2^2 + \dots + w_j^2 + \dots + w_d^2\right)$$

$$\min_{w_0} \frac{\partial L}{\partial w_0} = \sum_{i=1}^{n} -y_i \cdot 1 + \cancel{\lambda 0} \Rightarrow \frac{\partial L}{\partial w_0} = \sum_{i=1}^{n} -y_i$$

$$\min_{w_1 \dots w_d} \frac{\partial L}{\partial w_j} = \sum_{i=1}^{n} -y_j x_j + \lambda (2 w_j) \Rightarrow \frac{\partial L}{\partial w_j} = \sum_{j=1}^{n} -y_i x_{ij} + 2\lambda w_j$$

$L(w, w_0, \lambda)$

$\frac{\partial L}{\partial \lambda} = ?$ ⟶ we need some control on bias-variance

$0 + g(x, y) - c = 0 \Rightarrow g(x, y) = c$

$\boxed{\text{constant}}$

here,

for update eqn, we compute $\frac{\partial L}{\partial w_j}$ $\forall_j$ $1 \to n$ all points.

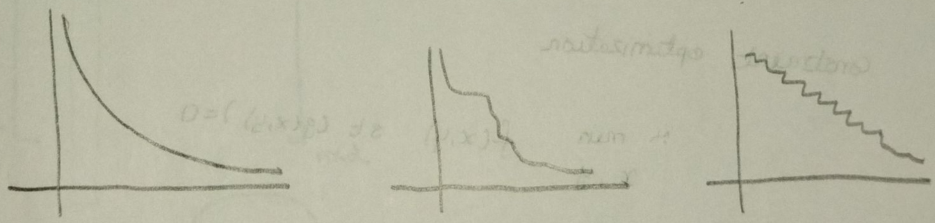
compute - batch wide: $\frac{\tilde{\partial L}}{\partial w_j} = \sum_{i=1}^{k} (-y_i x_{ij}) + 2\lambda w_j$ (estimate)

closer

batch	mini-batch	SGD
	$k=100$	$k=1$
$\frac{\partial L}{\partial w_j}$	$\frac{\partial L}{\partial w_j}$	$\frac{\partial L}{\partial w_j}$
iteration I_1	iteration I_2	iteration I_3

less accuracy

$$I_1 < I_2 < I_3$$

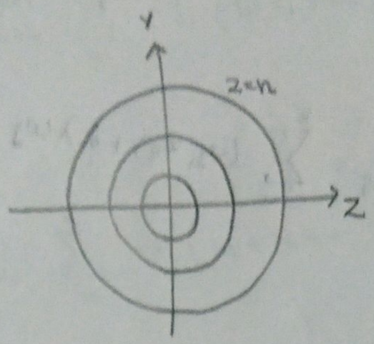


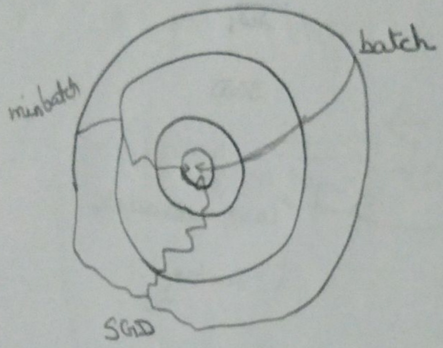
$$\min_{w, w_0} \sum_{i=1}^{n} L_i + \lambda \|w^2\|$$

loss func

regularization

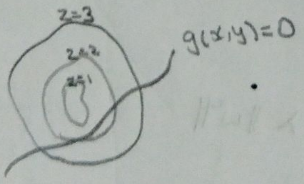
Contours $z = x^2 + y^2$





Constraint optimization

$$\min_{x,y} \quad f(x,y) \quad \text{s.t} \ (g(x,y)) = 0$$



we get z=2 if we solve satisfy constrained pseudocode

mini-batch

$i=0$; w_j^{old} = random-value $\forall j = 0 \ldots d$

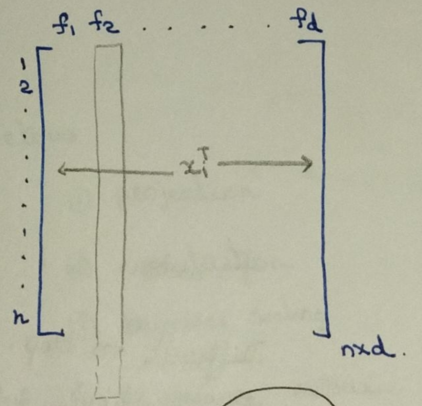
while each of the $\frac{\partial L}{\partial w_j}\Big|_{w_j old}$ is not close to 0.00001

for $j = 0$ to k

$$w_j^{new} = w_j^{old} - \eta \frac{\partial L}{\partial w_j}\Big|_{w_j^{old}}$$

$i = i+1$

Dimensionality Reduction



$x_i \in \mathbb{R}^d$ (no-class label)

in z-distri (or) standize { mean=0, var=1 }

interview Ques

if $d=2$ ⟶ scatter

$d=3$ ⟶ 3-D scatter

$d=4,5\ldots$ ⟶ pair

suppose, $d=100$

$d \rightarrow d'$

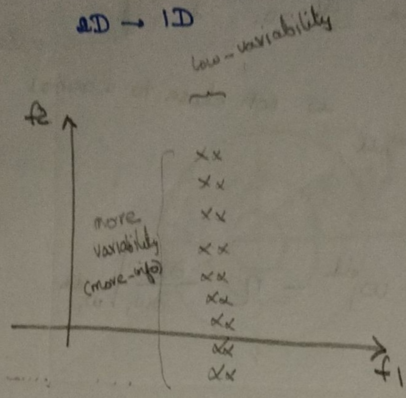
$d \ggg d'$

PCA

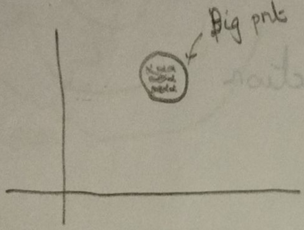
t-SNE

eg.

2D → 1D

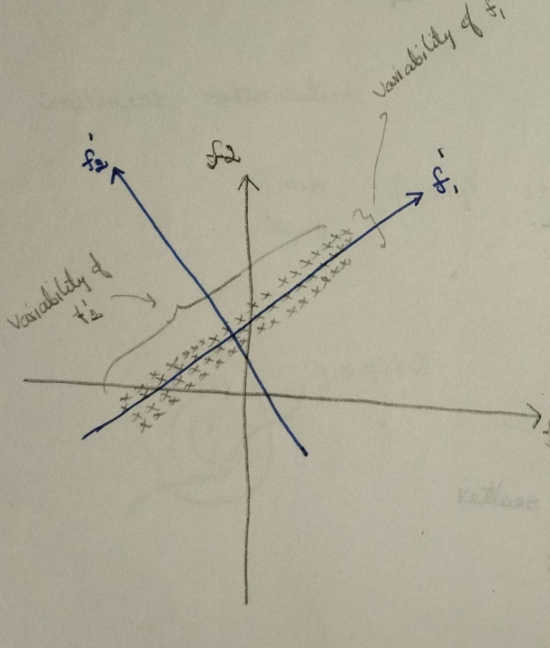


more variability → more info



SOME loss

INEVITABLE



f' - axis

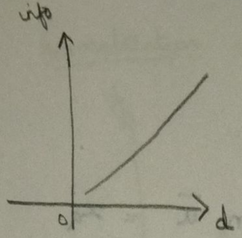
Difficult to drop. So i rotat co-ordinate sys.

standardized data

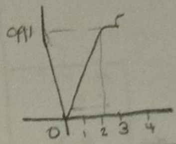
mean-centering

Variance scaling

objective : Find the top d' dimension which preserve as much as info. possible.



but we want low dim high info



2. dim 99% info

3-dim 99.5% info

4-dim 99.9% info

n-dim -100%

Detour

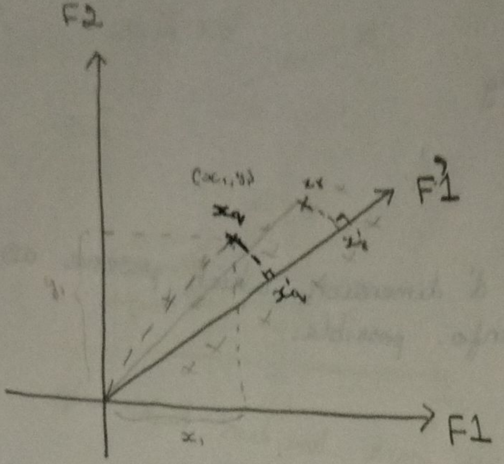
1. projection
2. normalization
3. min-max scaling
4. co-variance matrix

scaling:

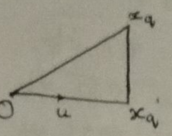
treat all columns by same weight

GD converge the way we expect. If not, certain feature might be very high and lead to very low updates for other features.

(X) remove anomaly before standardized



Projection of x_q on $\bar{u}$ = x_q'



$$\mathrm{Proj}_u x_q = \frac{|\vec{u}| \cdot x_q}{\| \vec{u} \|}$$

$$\cos\theta = \frac{\bar{u} \cdot x_q}{\|u\| \, \|x_q\|}$$

d-dim $\longrightarrow$ $x_q \;\; x_r \; \ldots$

d'-dim $\longrightarrow$ $x_q' \;\; x_r' \; \ldots$

if $\|u\| = 1$, unit vector

$$= \underbrace{\vec{u}}_{1\times d} \cdot \underbrace{x_q}_{d\times 1}$$

= mean of (x_q) scaler

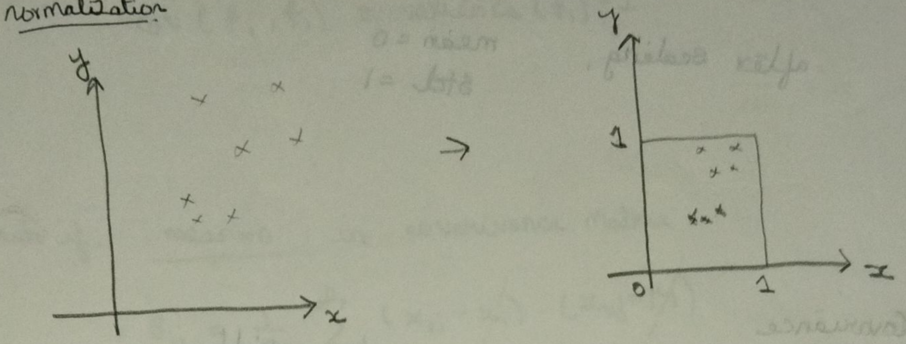
In PCA, why do we std.?

Variance is important

if we don't std. we miss which feature has more information.

Detour.

normalization

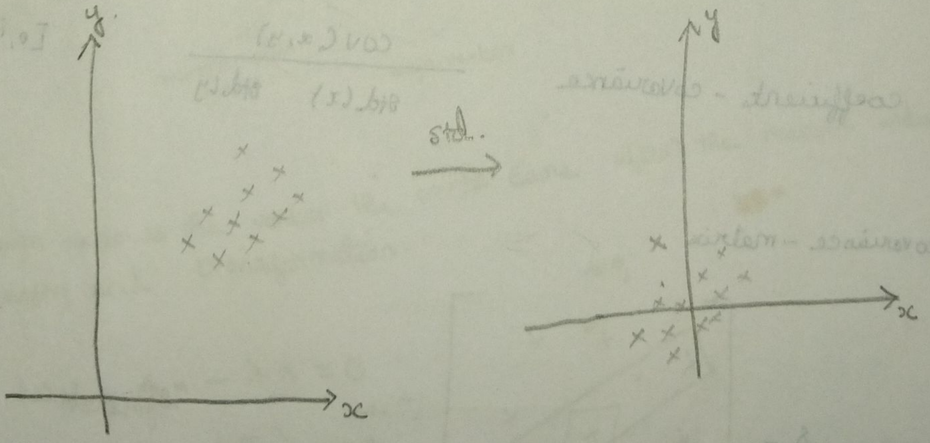


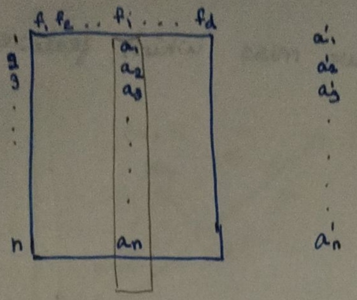
$\forall$ n-dim

Squeeze into [0,1]

Scale got collapsed.

$$a_i' = \frac{a_i - \min(a)}{\max(a) - \min(a)}$$





$$a_m = \frac{a_1 + a_2 + \dots + a_n}{n}$$

$$a_s = \text{std}(a)$$

$$a_i = \frac{a_i - \bar{a}}{a_s}$$

after scaling, mean = 0
std = 1

Covariance

measure of relationship b/w variable.

>0

$=0$

<0

$$S = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu_x) \cdot (y_i - \mu_y)$$

it can be any scale, 5, 50, 0.0005, 5k

coefficient - covariance $\dfrac{\text{cov}(x, y)}{\text{std}(x) \;\; \text{std}(y)}$ $[0, 1]$.

covariance - matrix

j^{th} col

$$S = \begin{bmatrix} x & x & x \\ x & \boxed{x} & x \\ x & x & x \end{bmatrix}$$

i^{th} row

Symmetrical

$$S_{ij} = \frac{1}{n} \sum_{i=1}^{n} (x_{ki} - \mu_i) \cdot (x_{kj} - \mu_j)$$

$$= \text{covariance } (f_i, f_j)$$

property

$$\text{cov}(f_1, f_2) = \text{cov}(f_2, f_1)$$

$f_i \quad f_j$

symmetrical.

$$\text{cov}(f_1, f_1) = \text{variance}(f_1) = 1$$

what if, mean=0 in covariance matrix

$$S_{ij} = \frac{1}{n} \sum_{i=1}^{n} (x_{ki} - \cancel{\mu_i}) \cdot (x_{kj} - \cancel{\mu_j})$$

$$= \frac{1}{n} \sum_{i=1}^{n} x_{ki} \cdot x_{kj}$$

$$= \frac{1}{n} X^T X$$

eign value & eign vector

$$Ax = \lambda x$$

eignvalue

transformation matrix

eign vector

$Ax \parallel x$

eign vector is the vector the points same after the matrix undergoes squeezing and transformation.

$$Ax - \lambda x = 0$$

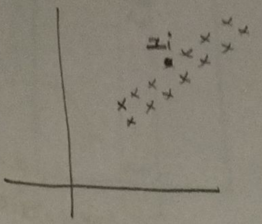
$$(A - \lambda I)x = 0$$

x can't be $\vec{0}$

$(A - \lambda I)$ is singular

If singular matrix, $|A - \lambda I| = 0$

(X) Find the unit-vector s.t max-variance projected x_i is covered.



$$= \max_u \text{ variance } \{u^T x_i\}_{i=1}^{n} \quad \text{st. } \|u\|$$

$$\text{Proj}_{u} x_i = \frac{u^T x_i}{|u|}$$

$$= \max_u \frac{1}{n} \sum_{i=1}^{n} (u^T x_i - \cancel{u^T \bar{x}_i})^2$$

$\cancel{u^T\bar{x}_i}$ (0)

std. mean = 0

in ideal case max-variance = 1

but we need max-variance in projectile x_i

$$= \max_u \frac{1}{n} \sum_{i=1}^{n} (u^T x_i)^2 \quad \text{s.t } \|u\| = 1$$

linear to matrix eqn.

$$= \max_u \frac{1}{n} u^T x_i^T u x_i$$

$$\|u\|^2 = u^T \cdot u$$

$$= \max_u u^T S u \quad \text{s.t } \|u\|^2 = 1$$

$$= \min_u -u^T S u + \lambda (u^T u - 1)$$

$L(u, \lambda)$

$$\frac{\partial L}{\partial \lambda} = 0$$

$$-Su + \lambda u = 0$$

$$\boxed{Su = \lambda u}$$

S-variance matrix (or) transformation matrix

$$Su - \lambda I u = 0$$

$$|S - \lambda I| = 0$$

solving $|S - \lambda I| = 0$

we got d - number of eign value & eign vector.

Important,

PCA is not Interpretable.

$d = 20 \longrightarrow d' = 2$ dim ?

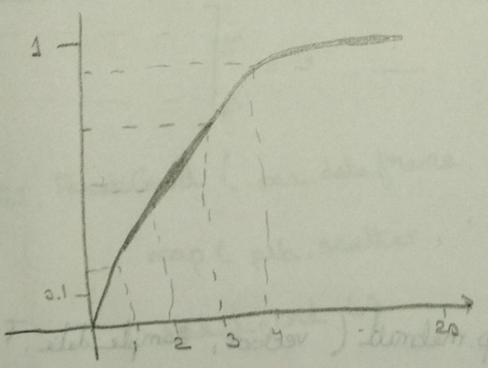
1 dim $\longrightarrow \dfrac{\lambda_1}{\lambda_1 + \lambda_2 + \dots \lambda_{20}}$

2 dim $\longrightarrow \dfrac{\lambda_1 + \lambda_2}{\lambda_1 + \lambda_2 + \dots \lambda_{20}}$

.

.

20 - dim $\longrightarrow \dfrac{\lambda_1 + \lambda_2 + \dots \lambda_{20}}{\lambda_1 + \lambda_2 + \dots + \lambda_{20}} = 1 \rightarrow$ max - variance



$$= \frac{\sum_{i=1}^{d'} \lambda_i}{\sum_{j=1}^{d} \lambda_i}$$

3 eign vector

4 eign vector captures 97 %. variance/

```python
from sklearn.preprocessing import StandardScaler
from scipy.linalg import eigh
```

```python
standarized_data = StandardScaler().fit_transform(data)

covariance_matrix = np.matmul(standarized_data.T, standarized_data)
```

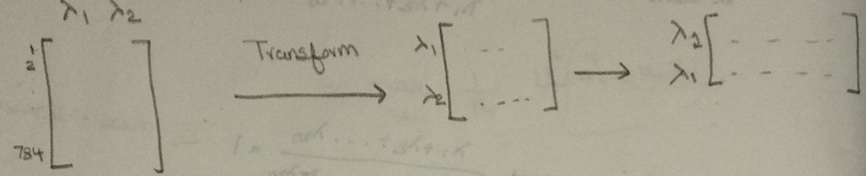
$\lambda_1 \quad \lambda_2$

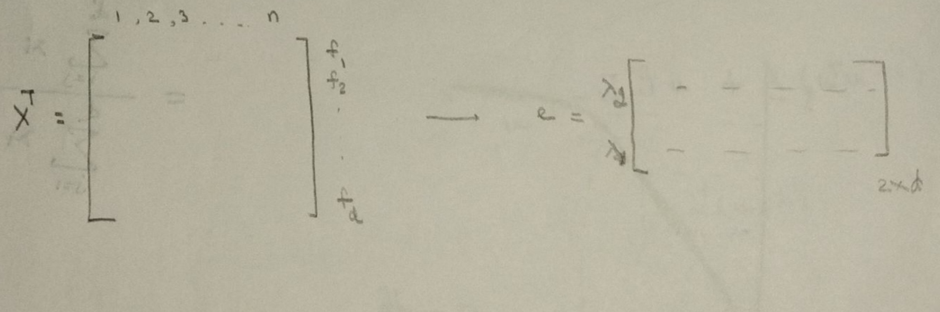
```python
values, vectors = eigh(covariance_matrix, eigvals=(782,783))
```

$\lambda_0, \lambda_1, \ldots \lambda_{784}$

Vectors[[0,1]] = Vector[[1,0]]

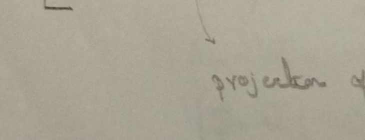
Vectors






```python
pca_data = np.matmul(vector, sample_data.T)
```

$e \cdot X^T$



projection of x_i on V_2 eigh vector

Mnist.

```
df = pd.read_csv('mnist.csv')

data = df.drop('label', axis=1)

label = df['label']


idx = 8
grid_data = data.iloc['idx'].tonumpy().reshape(28,28)
plt.imshow(grid_data)
plt.show()
```

- - - - -
- - - - - } pcA code.
- - - - -

```
pca_data = np.vstack(pca_data, label).T
```

$$\text{pca\_data} \quad \left[\begin{array}{c} \begin{array}{c|ccc} & v_1 & v_2 & v_3 \\ \begin{array}{c} 1 \\ 2 \\ 3 \\ \vdots \\ n \end{array} & & & \end{array}_{n \times 3} \quad \begin{array}{c} \text{label} \\ \left[\quad \right] \end{array} \end{array} \right]_{n \times 4}$$

```
pca_dataframe = pd.DataFrame( pca_data,
                              columns= ["λ1", "λ2", "λ3", 'label'])


sns.FacetGrid( pca_dataframe, hue="label", height=10).
        map( plt.scatter, "λ1", "λ2", "λ3").
        add_legend ()

plt.show()
```

P.T.O

```
from sklearn import decomposition

pca = decomposition.PCA()

pca.n_components = 2

pca_data = pca.fit_transform(sample_data)

pca_data = np.vstack(pca_data.T, labels).T

pca_df = pd.DataFrame(pca_data, columns=['λ1', 'λ2', 'label

sns.FacetGrid(pca_df, hue="label"). map(plt.scatter, "λ1", "λ
                                   add_legend()

plt.show()
```

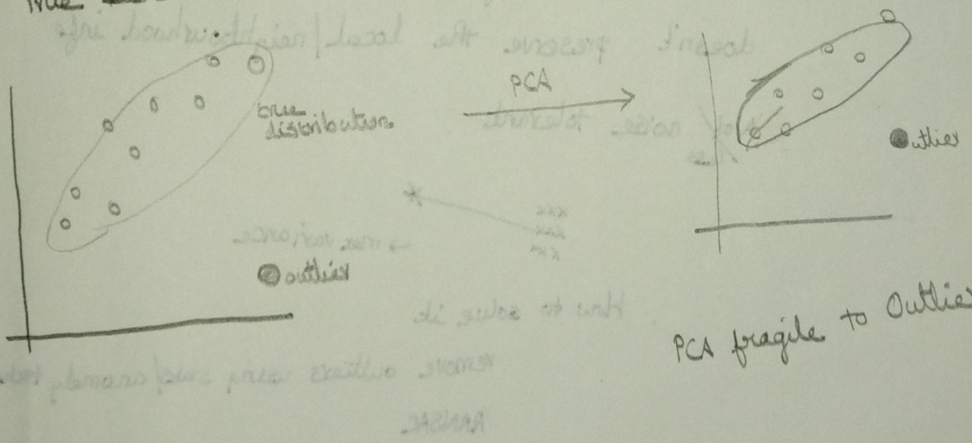
PCA Implementation

```
def (X, n_components):

    Xmean = X - np.mean(X, axis=0)

    cov_mat = np.cov(Xmean)

    value, vectors = np.linalg.eig(cov_mat)

    sorted_index = np.argsort(values)[::-1]
values max ⟶ min
    Sorted_vectors = Vectors[:, sorted_index]
arrange vectors
    eigenvectors_subset = sorted_vectors[:, num_components]

    Xreduced = np.dot(Xmean, eigenvectors_subset)

    return Xreduced
```

```
print ( pca.explained_variance_.shape )

percent_var_explained = pca.explained_variance_/
                        np.sum (pca.explained_variance_)

cum_var_explained = np.cumsum ( percent_var_explained )

plt.figure (1, figsize =(6,4))
plt.plot ( cum_var_explained , linewidth=2)
plt.axis ('tight')
plt.grid ()
plt.xlabel(" n_component")
plt.ylabel(" cumulative_explained_variance")
plt.show.
```

~~tilde~~ , stretch , distorch

True Distribution



PCA fragile to Outlier

RPCA

robust

$$Su = \lambda u$$

characteristic eq, $|S - \lambda I| = 0$

$\lambda = \checkmark$ (eign values)

eign vector,

$$|S - \lambda I|\,\vec{x} = 0$$

$$\begin{bmatrix} a-\lambda & b & c \\ d & e-\lambda & f \\ g & h & i-\lambda \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = 0$$

Sub $\lambda \rightarrow$ values

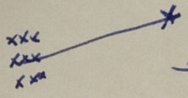
3 vectors $\Rightarrow$ eign vectors

<u>Limitation</u>

find linear mapping (visualization)

doesn't preserve the local/neighbourhood info

~~Not~~ less noise tolerant



→ max variance

How to solve it

remove outliers using stats/anomaly techs

RANSAC

L1 pca - change loss

~~no neighbourhood~~

remove high correlated feature before PCA (best practice)