Contingency Tables and Introduction to Stats

24/09/2022

1-3 B

$$\text{avg. age of India} = \frac{a_1 + a_2 + a_3 \dots + a_n}{n}$$

1 – 12L
2 – 13L
3 – 12L
4 – 15L
:
20 – 17L

Population: all possible subjects

Sample: collected subset for our experiments.

Summary metrics

Summarize data

— Mean (avg.) avg. CTC. was 10L

—

[2, 3, 4, 2, 16] $CTC = \frac{27}{5} = 5.4L$

Median

22 ← [2, 2, 3, 4, 16] Median. CTC → 3L

Mode: [20, 0, 0, 0, 30, 10, 5]

Mode = 0

[2, 2, 3, 4, 16]

range of pkg that is ~~offered~~ in the class: (2 to 16L)

S.d. / variance: wide ~~high~~ range → high (SD. / σ / variance)

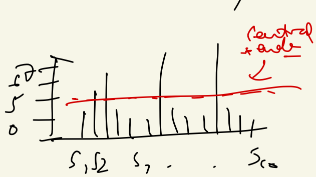
mean / median ← (central tendency)

S.d / variance ← variability

1, 3

-4, 1, 4

60%



Population $(\bar{x})$

measure:

Parameter (μ)

(any prop. of underlying population) (σ)

Sample $(\hat{x})$

Statistic $(\hat{x})$

(estimating the property from sample → statistic)

avg. wt of all Indians → parameter (humanly Impractical)

sample of 1M Indians → statistic (Practical)

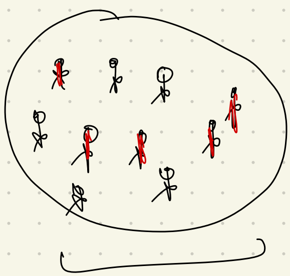
method to find out statistic from sample is called Estimation

answer will be estimate

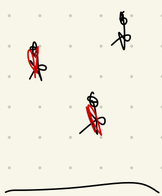
sample of 1M Indian to find avg wt

→ Statistic

Val. of avg → estimate



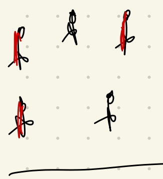
Population



prop. of ~~ppl~~

ppl w/ diabetes $= \left(\frac{2}{5}\right)$

$(\hat{p}) = .40$



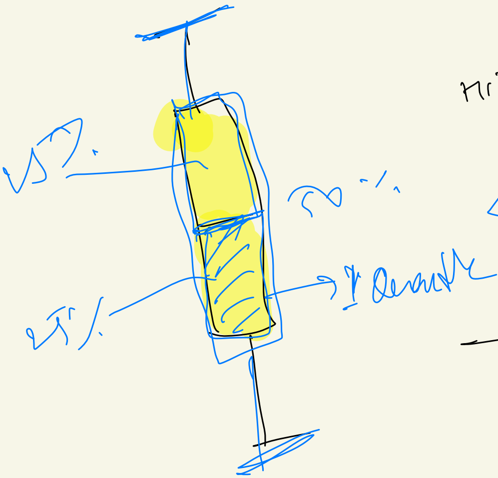
$\hat{p} = \left(\frac{3}{5}\right) = 60\%$

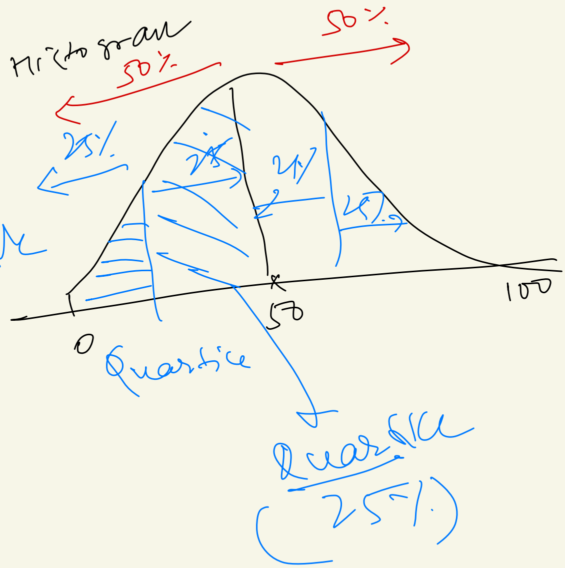
diabetic

	M	F
Men	.	.
Woman	.	.

prop. of ppl who are diabetic are

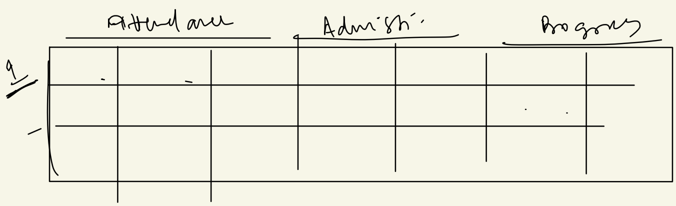
$(30\% - 40\%)$ (95% Confidence)

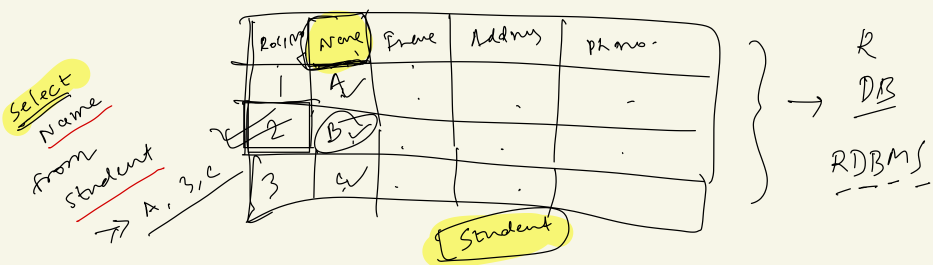

25%
50%
I Quartile
25%


Histogram
50%
50%
25%
25%
25%
25%
0
50
100
Quartile
Quartile
(25%)

Structured Query Language

Relational DB
Filtering rows & columns
Querying using SQL



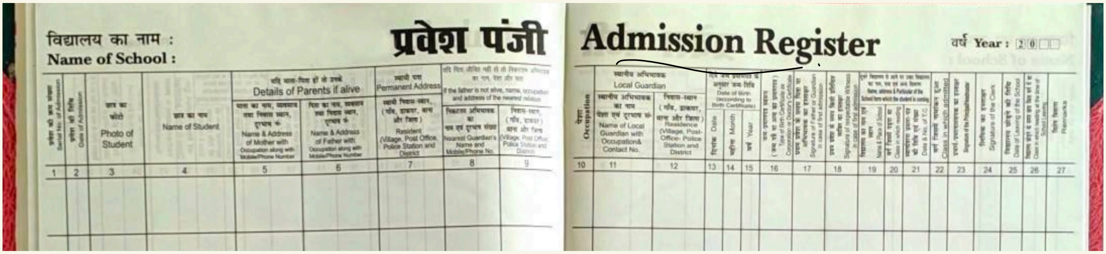


SQL

Information is stored in fixed tabular structure.

select Name from Student where
Roll No = 2. >> B

wait Time - 7.10 pm A
7.12 PM ATT
7.15 PM PR



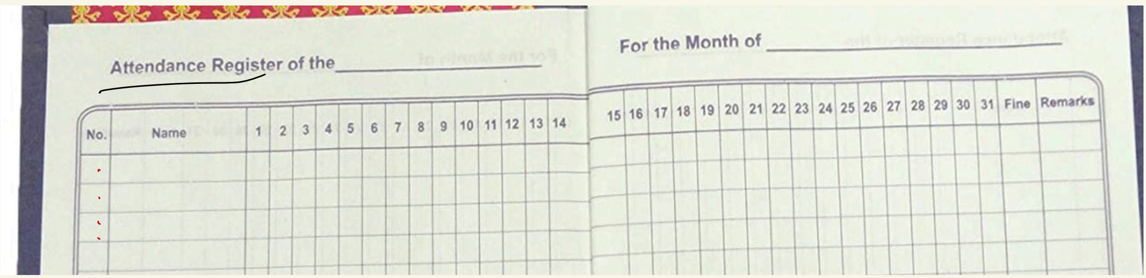
विद्यालय का नाम :
Name of School :

प्रवेश पंजी

Admission Register

वर्ष Year : 20

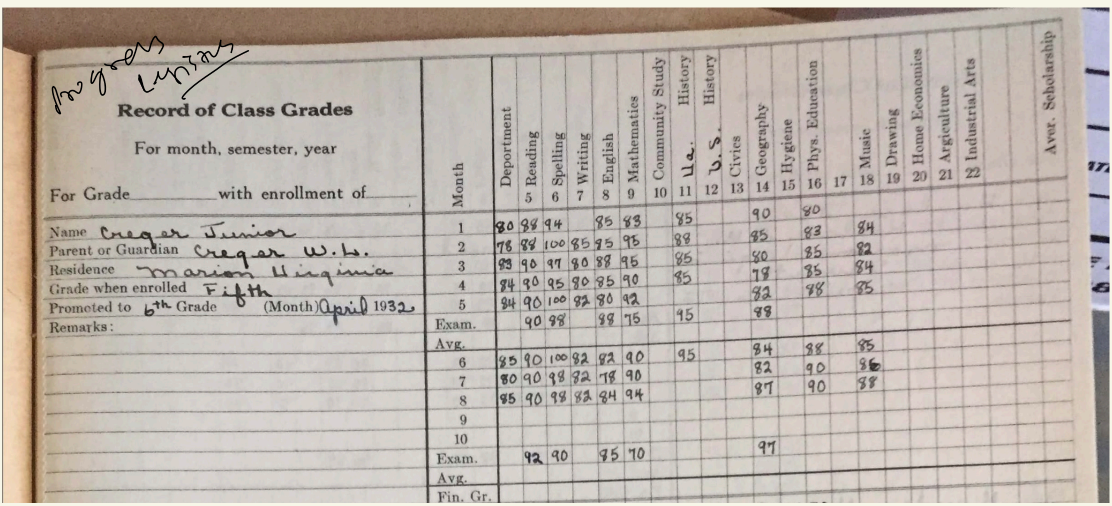
Serial No. of Admission	Date of Admission	Photo of Student	Name of Student	Details of Parents if alive: Name & Address of Mother with Occupation along with Mobile/Phone Number	Name & Address of Father with Occupation along with Mobile/Phone Number	Permanent Address: Resident (Village, Post Office, Police Station and District)	If the father is not alive, name, occupation and address of the nearest relation: Nearest Guardian's Name and Mobile/Phone No.	Village, Post Office, Police Station and District	Occupation	Local Guardian: Name of Local Guardian with Occupation & Contact No.	Residence (Village, Post-Office, Police Station and District)	Date of Birth: Date	Month	Year	16	17	18	19	20	21	22	23	24	25	26	Remarks
1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27



Attendance Register of the ____________

For the Month of ____________

No.	Name	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	Fine	Remarks



Progress Reports

Record of Class Grades

For month, semester, year

For Grade ________ with enrollment of ________

Name Creger Junior
Parent or Guardian Creger W. L.
Residence Marion Virginia
Grade when enrolled Fifth
Promoted to 6th Grade (Month) April 1932
Remarks:

Month	Deportment	5 Reading	6 Spelling	7 Writing	8 English	9 Mathematics	10 Community Study	11 Va. History	12 U.S. History	13 Civics	14 Geography	15 Hygiene	16 Phys. Education	17	18 Music	19 Drawing	20 Home Economics	21 Argiculture	22 Industrial Arts			Aver. Scholarship
1	80	88	94		85	83		85			90		80		84							
2	78	88	100	85	95	95		88			85		83		82							
3	83	90	97	80	88	95		85			80		85		84							
4	84	90	95	80	85	90		85			78		85		85							
5	84	90	100	82	80	92					82		88									
Exam.		90	88		88	75		95			88											
Avg.																						
6	85	90	100	82	82	90		95			84		88		85							
7	80	90	98	82	78	90					82		90		86							
8	85	90	98	82	84	94					87		90		88							
9																						
10																						
Exam.		92	90		85	70					97											
Avg.																						
Fin. Gr.																						

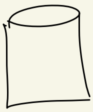
"Normalization"

Primary key

Name	Class (Year)	Dept	HoD.	% Marks
ABC	BADS	DS		70%
CDE	BADS	DS		95%

HOD

DS	PK
MG	BR



Server

< cust_id, / a/c > < ac/ bal / $ >

A C I D

"transaction in its entirety"

CBS

a/c no	Balance	
1	~~100~~ 200	−10
2	200	+10
3	400	

+30

−30

800 → 800

S −S R

+S

t20 → S: 100 R: 150

~~t = 10 → S: 90 R: 160~~

A	B	C
+10	+10	+10
−10		

Elementary Matrix Operations

Module 2

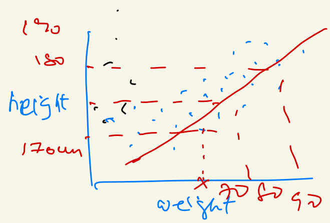
(numpy) → python library for numerical processing

Goal: build some "system" which will predict height of a person given its weight.

ht (cm)	wt (kg)
170	70
150	65
180	(80)
:	:
200	?

~~Alt~~ Altman Z score.

Wang → Deposit → Interp → FS → Bank → 0/1, % of default



height ∝ weight

Car Tyres

Durability of Car Tyres?

diameter
Thickness
friction
weight
materials

→ probability of ~~how~~ car tyre ~~going~~ "failing"

Data → Model → Results (Predictions)

{Current cars' m/s}

	D	T	F	W	M	F
1						
2						
3						
4						

Data will have features - muscular body, high wt, body hair, voice (deep). etc.

5 cols, 3 rows

Mus.	wt	B.H	V	Lbl
1	58	1	1	M
0				
0	—	—	—	F

1 - True (present)
0 - False (absent)

String of data: (Matrix | Matrices)
singular | plural.

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} \quad (3\times3)$$

"Matrix"
(3 cols, 3 rows)

$$\begin{bmatrix} 2 & 3 & 6 \\ 2 & 1 & 6 \end{bmatrix}$$

COLUMN

row

rows 2
cols : 3

(2 × 3)

[1] → 1 row, 1 col.

[1, 2 3] → (1 row, 3 col)
(1×3)

Size of matrix:
(row X cols)

50 car type features

30,000 car types

(30,000 X 50)

chp.
(12)
LR.

Vector: (direction + magnitude)
Quant

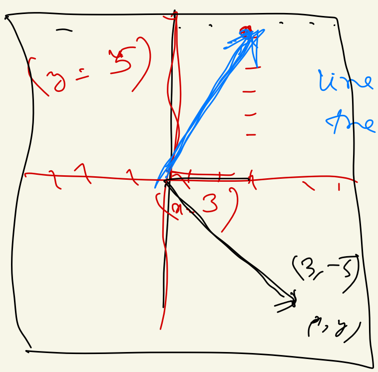
speed velocity ✓
acc. ✓

Col: vector
(n x 1)

row vector
1 row, n col

$n \times 1$ → col vector
$1 \times n$ → row vector

x ← 3
y ← -5



2x + 3y = 1

line connecting the pt (3,5) with origin (0,0) is vector.

Matrix (operation)

Addition $\underline{A} + \underline{B}$

(2x2) (2x2)

$$\begin{bmatrix} 2 & 2 \\ 1 & 0 \end{bmatrix} + \begin{bmatrix} 3 & 0 \\ 1 & 2 \end{bmatrix}$$

$$\boxed{\cancel{\begin{bmatrix} 3 & 2 & 1 \\ 0 & 1 & 2 \end{bmatrix} + \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}}} = \begin{bmatrix} 5 & 0 \\ 2 & 2 \end{bmatrix}$$

$(2\times 3) \longleftrightarrow (2\times 2)$

Subtraction $\underline{A} - \underline{B}$

$$\begin{bmatrix} -1 & 0 \\ 0 & -2 \end{bmatrix}$$

Multiplication:

$$\begin{bmatrix} 2 & 3 \\ 4 & 5 \end{bmatrix} \times \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

$$\begin{bmatrix} 2\cdot 1 + 3\cdot 0 & 2\cdot 2 + 3\cdot 1 \\ 4\cdot 1 + 5\cdot 0 & 4\cdot 2 + 5\cdot 1 \end{bmatrix} = \begin{bmatrix} 2 & 7 \\ 4 & 13 \end{bmatrix}$$

Matrix multiplication rule:

A $\qquad$ $\underline{B}$

$\underline{m} \times \underline{n} \qquad n \times p$

no. of cols of 1st matrix = no. of rows of 2nd matrix

$$\begin{bmatrix} 2 & 3 & 2 \\ 4 & 1 & 6 \end{bmatrix} \begin{bmatrix} 2 & 3 & 2 \\ 4 & 1 & 6 \end{bmatrix}$$

$2\times 3 \quad \cancel{\times} \quad 2\times 3$

$$3 \overset{A}{\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}} \qquad \underline{3A} = \begin{bmatrix} 3 & 6 \\ 9 & 12 \end{bmatrix}$$

~~Transp~~: Transpose:

$$A = \begin{bmatrix} 2 & 3 \\ 2 & 1 \end{bmatrix} \cancel{2}$$

$$\cancel{A^T} \quad A^T = \begin{bmatrix} 2 & 2 \\ 3 & 1 \end{bmatrix}$$

$$A = \begin{bmatrix} 2 & 3 & 2 \\ 1 & 0 & 1 \end{bmatrix} \quad (2 \times 3)$$

$$A^T = \begin{bmatrix} 2 & 1 \\ 3 & 0 \\ 2 & 1 \end{bmatrix} \quad (3 \times 2)$$

$$\underset{A}{\begin{bmatrix} 2 & 3 \\ 1 & 2 \end{bmatrix}} \times \underset{C}{\begin{bmatrix} 3 \\ 0 \end{bmatrix}}$$

$$= \begin{bmatrix} \quad \end{bmatrix}$$

wt size a

$x \% 2$

modulo operator

$$2 \overline{)21} \; (10$$
$$20$$
$$\overline{1} \rightarrow \text{remainder (answer)}$$

$21 \% 2 = 1$

$7 \% 3 = 1$

$19 \% 4 = \underline{3}$

python numpy array

matrices (p)

matrix

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$

(2×2)

$$\begin{bmatrix} 1 & 2 & 3 & 1 \\ 4 & 5 & 6 & 2 \\ 9 & 8 & 0 & 0 \end{bmatrix}$$

$(3, 4)$

(3×4)

no. of rows $= 3$

no. of cols $= 4 \rightarrow \boxed{2}$

$$\begin{bmatrix} 1 & 2 & 3 & 4 & 5 \\ 6 & 7 & 8 & 9 & 1 \\ 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \end{bmatrix}$$

$\boxed{(4 \times 5)}$ $\boxed{\text{dim} = 2}$

$(1 \times n)$

$$\begin{bmatrix} 1 & 2 & 3 \end{bmatrix}$$

row vector

$$\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}$$

col. vector

$(n \times 1)$

$$A = \begin{bmatrix} 2 & 3 & 4 \\ 5 & 6 & 7 \\ 2 & 1 & 2 \end{bmatrix}$$

Transpose: interchanges row & cols

$$\boxed{Y = (X^T X)^{-1} Y}$$

Transpose of A: $(A^T) = \begin{bmatrix} 2 & 5 & 2 \\ 3 & 6 & 1 \\ 4 & 7 & 2 \end{bmatrix}$

~~Trans~~

Trace of Matrix:

$tr(A) = 2 + 6 + 2$
$= 10$

$tr(A) = tr(A^T)$

Prime Number

divisible by 1 or itself.

21

identify whether a number is prime or not?

$2 \div 21$ ×
$3 \div$

$21 \div 2$ ×
$\cancel{21 \div 3}$ ✓

$19 \div 2$
$19 \div 3$
$19 \div 4$
⋮
$19 -$

11

~~2~~	~~3~~	~~4~~	~~5~~	~~6~~	~~7~~	~~8~~	~~9~~	~~10~~

take the number
– divide by all number starting from 2 through (no. − 1)

$\boxed{2 \dots 10}$

17 → 2 3 4 5 4 $\sqrt{17}$ 5

if not divisible then prime.

(n)
Testing primality
Test from '2' to $\sqrt{n} + 1$

113

$\sqrt{113} = 11$

2, ... 19, ~~29~~ . . .

2, 3, 4, 5 . . . , 9, 10, 11

193

11 157

< 2

Sieve of Eratosthenes

Prime number till 30.

2, 3, 5, 7, 11, 13, 17, 19, 23, 29

Matrix Multiplication:

① matrix is square

$$\begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \end{bmatrix} \times \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

(2×3) A

(3×3) B

$A \times B$

$m \times m$ $\quad$ $m \times m$

$(a \times b)$ $\quad$ $(b \times \_)$

Size of result:

$a \times b \times b \times c$

$= a \times c$

$$= \begin{bmatrix} 1 \cdot 1 + 2 \cdot 0 + 3 \cdot 0 & 2 & 3 \\ 2 & 1 & 0 \end{bmatrix}$$

2×3

$a \times b \quad b \times K$

$a \times K$

$2 \times \boxed{3 \times 3} \times 3$

2×3

$2 \times 3 \quad 3 \times 1$

(2×1)

$$\boxed{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 1 & 0 \end{bmatrix}$$

Trace of matrice:

usually defined

$$\begin{bmatrix} 2 & 3 & 2 \\ 1 & 0 & 1 \\ 1 & 1 & 1 \end{bmatrix} = 2 + 0 + 1 = \boxed{3}$$

Transpose:

$$A^T = \begin{bmatrix} 1 & 2 \\ 2 & 1 \\ 3 & 0 \end{bmatrix}$$

Matrix Vector multiplication

$$\begin{bmatrix} 2 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 0 \\ 1 \end{bmatrix}$$

3×3 $\quad$ 3×1

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

3×3 (Identity matrix)

It has all 1's in its diagonal and 0's else where (Square matrix)

$$= \begin{pmatrix} 2 + 0 + 1 \\ 1 + 0 + 0 \\ 0 + 0 + 1 \end{pmatrix} = 3 \times 1$$

$$= \begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix}$$

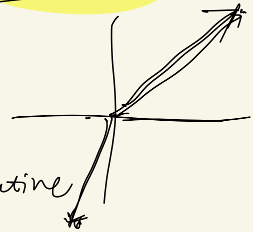
2x2 identity matrix : $\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$

5x5 $\begin{bmatrix}1 & & & & \\ 0 & 1 & & & \\ & & 1 & & \\ & & & 1 & \\ & & & & 1\end{bmatrix}$

$$\begin{bmatrix}-1 & 9\\ 11 & 6\end{bmatrix}\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} = \begin{bmatrix}-1 & 9\\ 11 & 6\end{bmatrix}$$

2x2 $\quad I_2$

$$\boxed{A \times I = A}$$



$A \times B \neq B \times A$ (commutative

(2x2) (2x2)

5x3 = 15
3x5 = 15

A = 3x5
B = 5x1

$\boxed{A \times B = 3 \times 1}$

B x A = 5x1 x 3x5

no. of axis = no. of dimension

{0,1}

{0,1,2}

$$\begin{bmatrix}1 & 2 & 3\\ - & - & -\\ - & - & -\end{bmatrix}$$

axis = 0

row axis = 1
col = 0

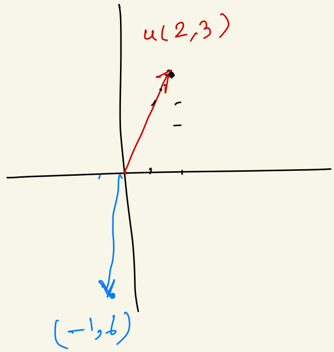
Inner Product

$u = \begin{bmatrix} 2 \\ 3 \end{bmatrix}$ $v = \begin{bmatrix} -1 \\ 6 \end{bmatrix}$

u:

$$u^T v = \begin{bmatrix} 2 & 3 \end{bmatrix} \begin{bmatrix} -1 \\ 6 \end{bmatrix}$$

1×2 2×1



=

1×2 2×1

$= 2(-1) + 3(6)$

$= -2 + 18$

$$\boxed{u^T v = \boxed{16}}$$

$(-\infty\ \infty)$

Outer Product: uv^T

Inverse of matrix: $\frac{1}{\det(A)} (\operatorname{adj} A)$

$\begin{bmatrix} a & b \\ c & d \end{bmatrix}$

$\det(A) = ab - cd$

det $\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}$

$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix}$

$((6)(-3)) = 3$

$\frac{1}{\det A} \begin{bmatrix} 6 & -3 \\ 2 & 1 \end{bmatrix}$

$\begin{bmatrix} 1 & 2 \\ 3 & 6 \end{bmatrix} = 6 - 6$

$= 0$

$u = [\,2 \quad 3 \quad 6\,]$

$$\|u\| = \sqrt{2^2+3^2+6^2}$$

norm

$$= \sqrt{4+9+36}$$

$$= \sqrt{49} = 7$$

$$\|u\|_2 = \sqrt{x_1^2 + x_2^2 + x_3^2}$$

$$\|u\|_3 = \sqrt[3]{x_1^3 + x_2^3 + x_3^3}$$

$$\|u\|_1 = x_1 + x_2 + x_3.$$

$$\begin{array}{c} \begin{array}{ccc} 0 & 1 & \end{array} \\ \begin{bmatrix} 0 & 3 & 6 \\ 1 & 4 & 7 \\ 2 & 5 & 8 \end{bmatrix} \end{array}$$

Column major form

Linearizing a matrix:

0, 1, 2, 3, 4, 5, 6, 7, 8

row #	column #	element #
0	0	0
1	0	1
2	0	2
0	1	3
1	1	4
2	1	5
	⋮	6

no. of rows = 3

$\boxed{\text{element} = \text{row}}$

element = row + (column × no. of rows)

$= 1 + (1 \times 3)$

$= 4$

$= 2 + 1 \times 3 = 5$ . . .

10 → 100 sec
11 → 121 sec
12 → 144 sec

	T_1	T_2	T_3	T_4	..	T_m
U_1	✓	x				
U_2			x			
U_3				✓		
⋮						
U_n	x					✓

no. of entries in matrix < 5%

95% → zero

" sparse "

20,000 books

w_1, w_2 w_3 w_4 .. unique words w_n

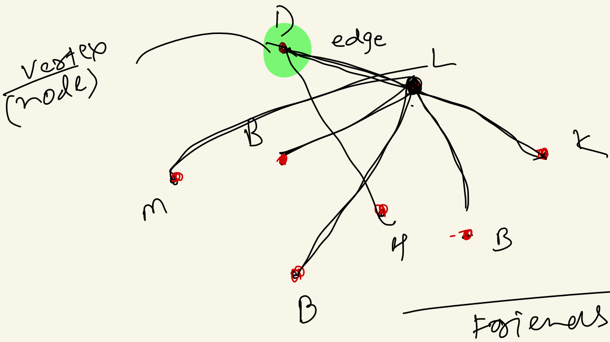
13 - - - . -

b_1
b_2
row b_3
b_m

TD Matrix

Graphs

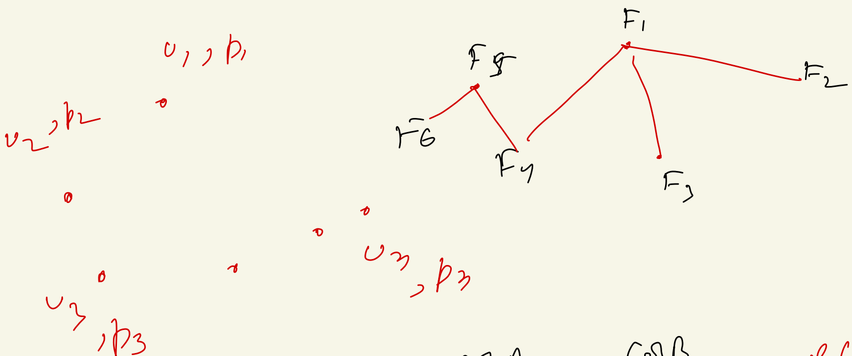
collection of nodes (vertex) and edges



Connection → edges

nodes / vertices

Friends network



df :

col A	col B
1	A
2	B
3	C
~~D~~	~~d~~
d	D
A	1
B	2
C	3
~~d~~	~~D~~
~~D~~	~~d~~

df

A	B
B	A

drop dupl

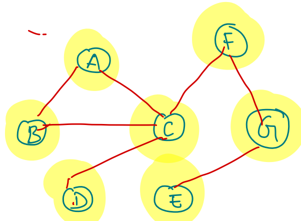
Note Book 10 (Numpy) : Part 3 (Sparse matrices).

DS/ML → Data (matrix)

H/W

we want to

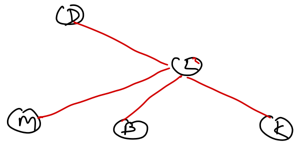
Friendship network in computer → Graphs are used.



" Friend Network "

	M	B	J	E
M		1		
B	1			
J				1
E			1	

Collection of nodes and edges



Storing Graphs in Computers :

(i) Adjacency matrix

	A	B	C	D	E	F	G
A		1	1				
B	1		1				
C	1	1		1		1	
D			1				
E							1
F			1				1
G						1	1

COO ~~Most~~ format :

Row :	A	A	B	B			C	G
col :	B	C	A	C	-	-	F ..	F
val :	1	1	1	1			1	1

Src	Target
1	a
2	b
4	d
d	4
a	1
b	2
d	4
4	d
	a

$$
\begin{array}{c|ccc}
 & 0 & 1 & 2 \\
\hline
0 & 0 & -2.5 & 1.2 \\
1 & 0.1 & 1.0 & 0 \\
2 & 6.0 & -1.0 & 0
\end{array}
\quad
\begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix}
=
\begin{bmatrix} 0+(-5)+3.6 \\ \\ \\ \end{bmatrix}
$$

$3\times3 \qquad 3\times1 \qquad 3\times1$

matrix $\times$ vector = vector

$(m\times n) \quad (n\times 1) = (m\times 1)$

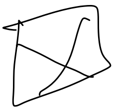
default dictionary

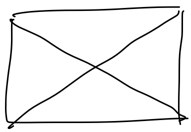
(dim) key val

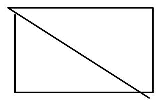
{ 0 → 1: -2.5, 2: 1.2 }

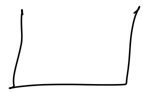
{ 1 → 0: 0.1, 1: 1.0 }

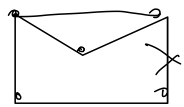
{ 2 → 0: 6.0, 1: -1.0 }

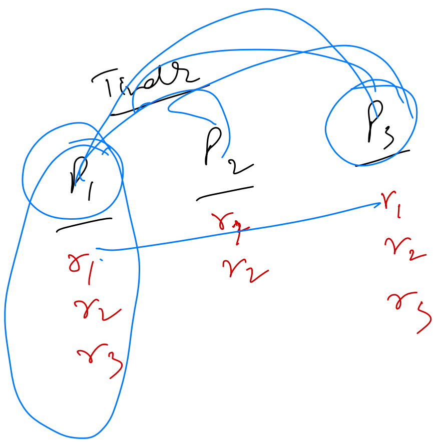












JEE

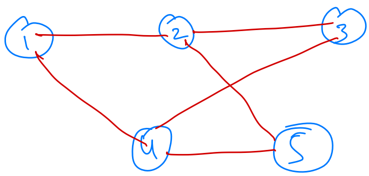
Roll no	Rank
1	5
2	1
3	2

A CS,

B - ~

$$\left\{ \frac{100K}{5\ pr} \right\}$$

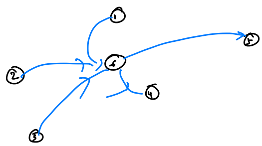
100!



Page Rank: Ranking algorithm of webpages.

(Slow) → Execution of lookup/ Query Retrival (Fast)

Story:



website
- Text / Image/ MM
- Links

1 - 30%
2 - ~~20%~~
3 - 40%
4 - 20%
5 - 5%
6 - 5%

$x \leftarrow xA$

$1\times n$ $1\times n$ norm

matrix X vector = vector

$\boxed{xA = x}$

A X eigen vector = e.v

val. of ev : eigen value

$$\begin{array}{c|ccc} & 1 & 2 & 3 \\ \hline 1 & 0 & 0 & 0 \\ 2 & 0 & 1 & 1 \\ 3 & 0 & 0 & 0 \end{array}$$

9x9

e

Row →	2	2
Col →	2	3
val →	1	1

$\left[\; - \quad 1 \quad 1 \;\right]$

Markov Chains

memoryless

60%	~~75%~~	80%	"85%"
6	7	8	9

"1 step back"

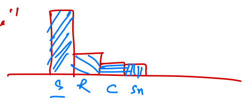
(8) 90%

2 step

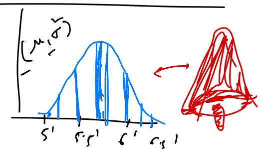
(7,8)

[60% 20% 10% 10%]

Sunny Rainy Cold Snow



Height of student in Class:

① 

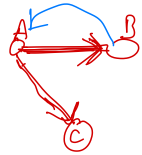
Symmetric: Normal distribution (Bell curve) Gaussian Distr.

② Poisson Distribution:

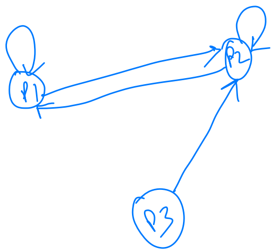
#. white color vehicle.

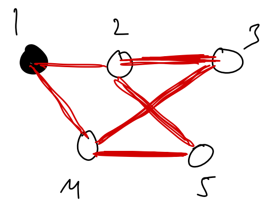
{Time → # " }

	A	B	C	D	E	F
A		1	1	1		
B	1		1		1	
C						
D						
E						
F						



	A	B	C
A		1	1
B	1		
C			



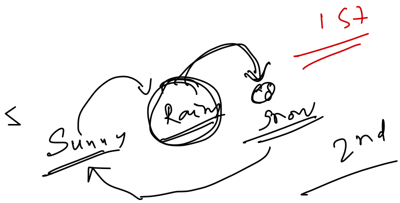


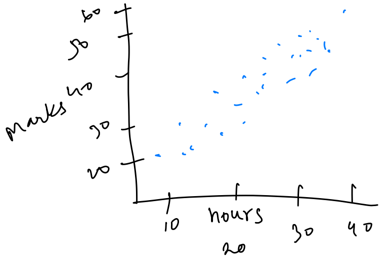
Markovian:

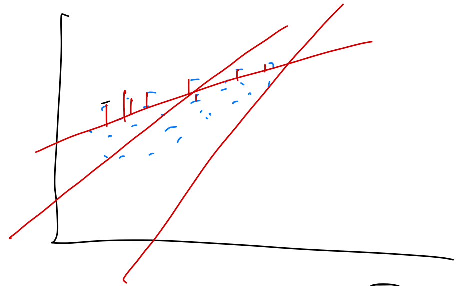
Future depends on past

1 steps

2 steps.

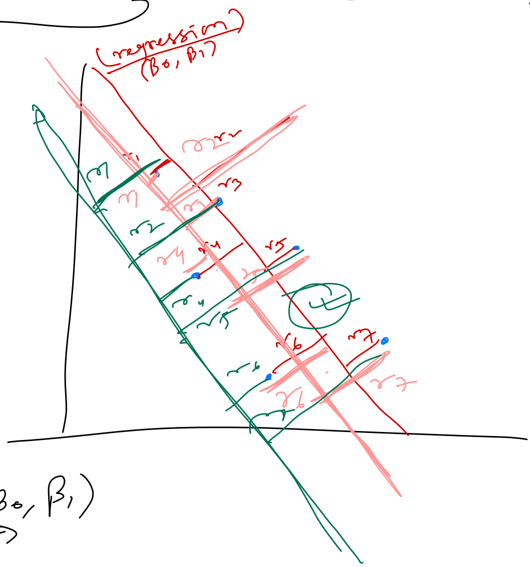






$$\sum_{i=1}^{n} x = \underbrace{x + x + x + \dots + x}_{n} = \sum x = \boxed{nx}$$

minimize
the residuals
$r_1 + r_2 + \dots r_7$
should be as
smallest as
possible



Data

(hrs) (x), (marks) (y) $\longrightarrow$ Model $\longrightarrow$ (β_0, β_1)

min. $2x^2 - 3x + 5 \rightarrow 2 \cdot \frac{9}{16} - 3 \cdot \frac{3}{4} + 5$

$(5-2)^2 = 9$

$(2-5)^2 = 9$

$\frac{d}{dx}(2x^2 - 3x + 5) = 0 \rightarrow \frac{9}{8} - \frac{2 \cdot 9}{2 \cdot 4} + 5$

$4x - 3 = 0$

$\boxed{x = \frac{3}{4}}$

$= \frac{9}{8} - \frac{18}{8} + 5$

Σ

$= 5 - \frac{9}{8}$

$= \cancel{3}$

$\hat{y}_i = \beta_0 + \beta_1 x_i$ (estimated from data)

$(y - \hat{y})$: residual.

SSR (sum of square residuals)

$$= \sum_i (y_i - \hat{y}_i)^2 \text{ (to cancel the effect of opp. signs)}$$

$$= \sum_i (y_i - (\beta_0 + \beta_1 x_i))^2 \qquad [(a-b)^2 = a^2 - 2ab + b^2]$$

$$= \sum_i \left[y_i^2 - 2y_i(\beta_0 + \beta_1 x_i) + (\beta_0 + \beta_1 x_i)^2 \right] \text{ --- } 1$$

$$= \sum_i \left[y_i^2 - 2y_i \beta_0 - 2y_i \beta_1 x_i + \beta_0^2 + \beta_1^2 x_i^2 + 2\beta_0 \beta_1 x_i \right]$$

$$\left[\frac{\partial\, SSR}{\partial \beta_0} = \sum_i \left[0 - 2y_i + 0 + 2\beta_0 + 2\beta_1 x_i \right] = 0 \right.$$

$$= 2\left[\sum \beta_0 + \beta_1 x_i - y_i \right] = 0$$

$$\Rightarrow \sum (\beta_0 + \beta_1 x_i - y_i) = 0$$

$$\Rightarrow \sum_{i=1}^{n} \beta_0 = \sum_{i=1}^{n} (y_i - \beta_1 x_i)$$

$$\Rightarrow n\beta_0 = (y_1 + y_2 + y_3 + \cdots y_n) - \beta_1 (x_1 + x_2 + \cdots x_n)$$

$$\beta_0 = \frac{(y_1 + y_2 + \cdots + y_n)}{n} - \beta_1 \cdot \frac{(x_1 + x_2 + \cdots x_n)}{n}$$

$$\boxed{\beta_0 = \bar{y} - \beta_1 \bar{x}}$$

$$SSR = \sum_i \left[y_i^2 - 2y_i\beta_0 - 2y_i\beta_1 x_i + \beta_0^2 + \beta_1^2 x_i^2 + 2\beta_0\beta_1 x_i \right] \quad \text{— (1)}$$

$$\frac{\partial SSR}{\partial \beta_1} = \sum_i \left[2\beta_1 x_i^2 + 2\beta_0 x_i - 2x_i y_i \right] = 0$$

$$= \sum_i -2x_i \left[y_i - (\beta_0 + \beta_1 x_i) \right] = 0 \quad \text{— (2)}$$

Substituting β_0 in (2)

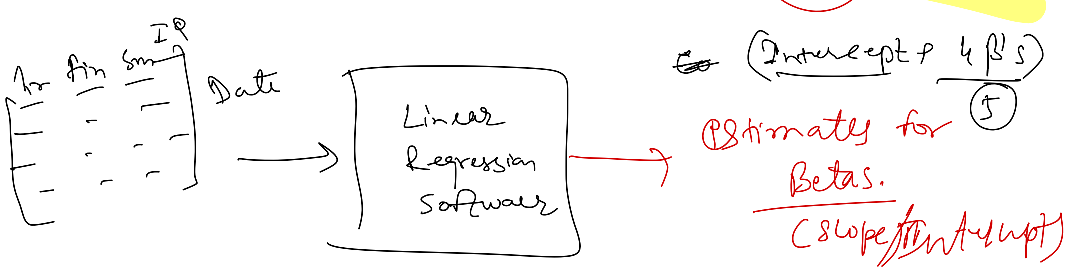
$$= \sum x_i \left[y_i - \left((\bar{y} - \beta_1 \bar{x}) + \beta_1 x_i \right) \right]$$

$$= \sum x_i \left[y_i - \bar{y} - \beta_1 (x_i - \bar{x}) \right] = 0$$

$$= \sum x_i (y_i - \bar{y}) - \sum \beta_1 x_i (x_i - \bar{x}) = 0$$

$$\Rightarrow \sum x_i (y_i - \bar{y}) = \beta_1 \sum x_i (x_i - \bar{x})$$

$$\boxed{\beta_1 = \frac{\sum x_i (y_i - \bar{y})}{\sum x_i (x_i - \bar{x})} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2}}$$



Multiple Linear Regression:

Predictor → variables
hours studied,
IQ, financial status,
Smoking-since,

Marks scored = $\beta_0 + \beta_1 \times$ hours studied $+$
β_2 IQ $+$ β_3 financial status $+$
β_4 Smoking since (Yrs)

$2 \times$ IQ $+$ β_3 fin $+$ β_4 Smoking

$$\begin{bmatrix} \text{hrs studied}, & \text{IQ}, & \text{finance}, & \text{Smoking}, & \text{Marks Scored} \\ - & - & - & - & - \\ - & - & - & - & \end{bmatrix}$$

MLR : has more than 1 predictor

response/prediction

$$\boxed{Y = \beta X + \epsilon}$$

"noise"

co-efficient

predictor (independent vars)

$$= \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \epsilon$$

$$\boxed{y = \beta_0 + \beta_1 x}$$

$y_0 = \beta_0 \cdot 1 + \beta_1 x_0$

$y_1 = \beta_0 \cdot 1 + \beta_1 \cdot x_1$

$y_2 = \beta_0 1 + \beta_1 x_2$

$$\begin{bmatrix} y_0 \\ y_1 \\ y_2 \\ y_3 \\ y_4 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$(n \times 1) = (n) \times 2 \times (2 \times 1) + (n \times 1)$$

$$\begin{bmatrix} 1 & x_0 & y_0 \\ 1 & x_1 & y_1 \\ \vdots & \vdots & \vdots \\ 1 & x_n & n \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{bmatrix}$$

$\beta_0 \beta_1 x_0$
$\beta_2 y_0$

$$\boxed{Y = X \beta + \epsilon}$$

This is matrix

design matrix

weights matrix

noise/disturbance matrix.

This is not equivalent to $y = mx + c$ & $c \neq \epsilon$

$\int x^2 + \int e^x$

$\frac{x^3}{3} + c_1 + e^x + c_2$

$y = \beta_0 + \beta_1 \times hr + \beta_2 \times IQ + \beta_3$ fin $+ \beta_4$ Smoking

$i_1 \quad i_2 \quad i_3 \quad i_4$

$\frac{x^3}{3} + e^x + C$

$y = \beta_0 + \beta_1 \times hr + \beta_2 \times IQ + \beta_3\, fin + \beta_4\, smoking$

$y = marks$

$y = (c) + (\beta_1\, hr) + (c_2 + \beta_2 \times IQ) + (c_3 + \beta_3 \times fin)$

$$\underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \end{bmatrix}}_{marks} = \underbrace{\begin{array}{ccccc} & hours & IQ & fin & smoking \\ 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot \\ 1 & \cdot & \cdot & \cdot & \cdot \end{array}} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \end{bmatrix} + \underbrace{\begin{bmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{noise}$$

~~are~~ does not mean intercept is "1".

$$\begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & x_0 \\ 1 & x_1 \\ 1 & \vdots \\ 1 & x_n \end{bmatrix}}_{(n\times 2)} \underbrace{\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}}_{(2\times 1)} + \begin{bmatrix} \epsilon_0 \\ \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_n \end{bmatrix}$$

$$\boxed{Y = X\beta + \epsilon}$$

$\epsilon \neq$ intercept

$y_0 = \beta_0 + \beta_1 x_0 + \epsilon_0$ ← 1st row of dataset

$y_1 = \beta_0 + \beta_1 x_1 + \epsilon_0$ — 2nd row of dataset

$\vdots$

$\epsilon = (Y - X\beta)$

$\frac{d}{dx}(\sin^2 x + \cos^2 x)$

(a) inner product: ⇒ $(x^T x)$

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix} = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}$$

to minimize ϵ, lets find out inner product

$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} (2\times 1)$

$X = (n \times 2)$

$\cancel{\beta^T X^T y}$

$1\times \boxed{2\times 2} \times \boxed{n \quad n} \times 1$

1

$$\epsilon^T \epsilon = (y - X\beta)^T (y - X\beta)$$

$$= (y^T - (X\beta)^T)(y - X\beta)$$

$$= (y^T - \beta^T X^T)(y - X\beta)$$

$$= y^T y - \underline{y^T X\beta} - \underline{\beta^T X^T y} + \beta^T X^T X \beta$$

$$(y^T X\beta)^T = \beta^T X^T (y^T)^T$$

$$= \beta^T X^T y$$

AB

$(AB)^T$

$= B^T A^T$

$(A - B)^T$

$= A^T - B^T$

$$\epsilon^T \epsilon = \underline{y^T y} - \underline{2\beta^T X^T y} + \beta^T X^T X \beta$$

0

$$\frac{\partial \epsilon^T \epsilon}{\partial \beta} \to 0 = -2X^T y + 2X^T X \beta$$

$\dfrac{\partial a^T b}{\partial b} = \dfrac{\partial b^T a}{\partial b} = a$

$\dfrac{\partial b^T A b}{\partial \beta} = 2Ab = 2b^T A$

$$0 = -2X^T y + 2X^T X \beta$$

$$\boxed{X^T X \beta = X^T y}$$

Ordinary Least Square

Normal form Equation

$$(X^T X)^{-1} (X^T X) \beta = (X^T X)^{-1} X^T y$$

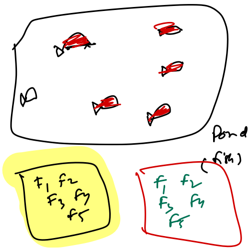
$$\boxed{\beta = (X^T X)^{-1} X^T y}$$

$Y = \beta X$

$\beta = \frac{Y}{X}$ ~~$\beta = X^{-1} Y$~~

(50 x 5)

$= \frac{Y}{(X^T X)} X^T$



~~$Y = (X^T X)^{-1} X^T Y$~~

$$\boxed{Y = (X^T X)^{-1} X^T Y}$$

Inner product

"matrix"

50 examples

$X = (n \times 2)^T$

$= (2 \times n)(n \times 2)$

$= \boxed{2 \times 2}$

$A^T A$

$= \boxed{\beta}$

100M

X Y

Age, Smoking, FH, F.I.

$(X^T X)^{-1} X^T Y$

$\nabla(\ell)$

$\beta \leftarrow \beta_{old} - n \times$

new const.

100 Iterative algorithms

$x^2 - 5x + 6 = 0$

$\frac{-b \pm \sqrt{b^2 - 4ac}}{2a}$

$x^3 - 1 + 2x = 0$

$x^5 - x^4 + 3x^2 = 0$

120

inner prod

$A = m \times n$

$A^T = n \times m$

$A^T A$

$= n \times m \; m \times n$

$\boxed{A = n \times n}$

1.99

2.01

$

cust

risk

High

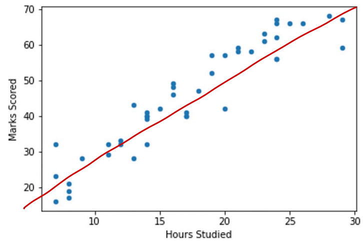
med

low

nil

Logistic Regression

Hours Studied	Marks Scored
12	32
16	46
16	49
13	43
29	59
8	17
24	56
14	40
7	23
15	42
::	::
19	57
24	62
14	40
24	67
19	52



Marks = $\beta_0 + \beta_1 \times$ Hrs.

"(Regression)"

Continuous

$(-\infty, \infty)$

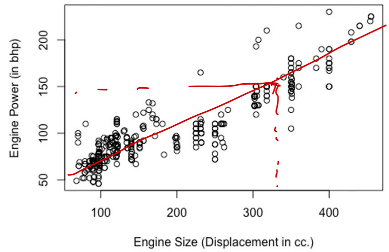
"(Classification)"

Categorical

$(x_1, x_2, x_3, x_4 .. x_n)$ $(\hat{y})$

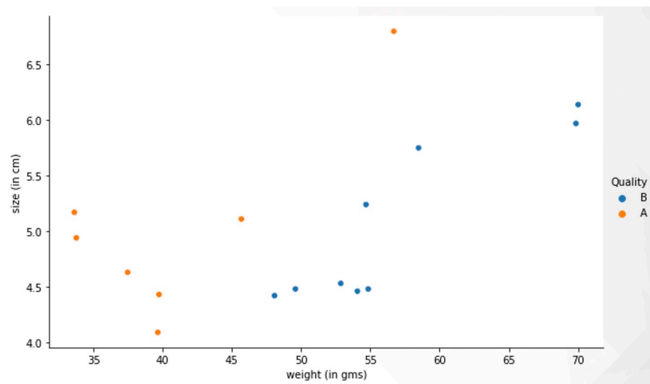
data

Engine Size (cc)	Engine Power (bhp)
307	130
350	165
318	150
304	150
302	140
454	220
440	215
199	97
200	85
97	88



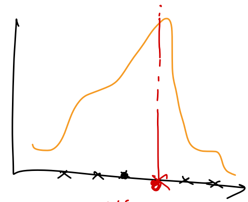
weight (in gms)	size (in cm)	Quality
54.84	4.48	B
48.07	4.42	B
54.68	5.24	B
54.06	4.46	B
70.02	6.14	B
45.67	5.11	A
69.86	5.97	B
37.45	4.63	A
33.60	5.17	A
49.58	4.48	B
39.63	4.09	A
39.72	4.43	A
58.48	5.75	B
33.74	4.94	A
52.86	4.53	B
56.69	6.80	A

Quality A is a better quality than **Quality B**



2 types of predictions — Continuous & Categorical

Given the dataset find most likely parameter of the distribution



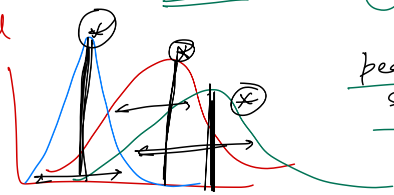
Distribution: a graph showing prob.

μ, σ^2 Normal distribution → distr. of height

Binomial " → p(success) p

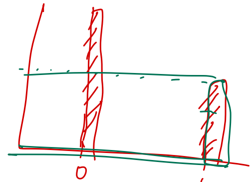
Poisson " λ

Normal

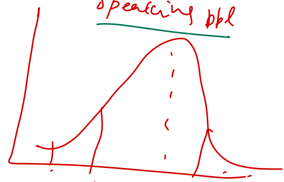


peak position → mean (μ)

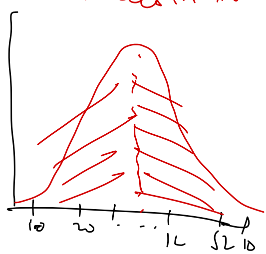
spread → variance (σ^2)



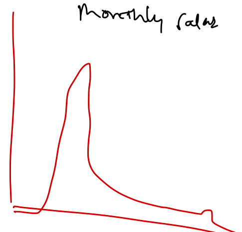
count of english speaking ppl



dist of height of ~~avg~~ 30yr males in IN



Monthly sales



" Log normal "

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

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

$$\frac{d}{dx}\log(x-x^2)$$

$$\underbrace{= \frac{1}{(x-x^2)} \times (1-2x)}_{\text{Chain rule}}$$

$$\boxed{\frac{d(uv)}{dx} = uv' + vu'}$$

$$\cancel{\frac{d}{dp}(y\log p) = y}\ \log$$

$$4\frac{d}{dp}\log(1-p) = \frac{1}{(1-p)}(0-1)$$

$$= \left(\frac{-1}{1-p}\right)$$

Coin $\quad$ H T H T T T H

define: $\quad$ prob. of head $= p(H) = p$

prob. of tail $= p(T) = (1-p)$

MLE:

$$L = p \times (1-p) \times p \times (1-p)^3 \times p$$

$$L = p^3 (1-p)^4$$

$$\log L = \log\left(p^3 \cdot (1-p)^4\right)$$

$$= \log p^3 + \log (1-p)^4$$

$$\boxed{\log L = 3\log p + 4\log(1-p)}$$

$$\frac{\partial \log L}{\partial p} = \frac{3}{p} + \frac{4}{(1-p)}(-1) \quad \text{(from chain rule)}$$

$$\Rightarrow \frac{3}{p} - \frac{4}{(1-p)} = 0$$

$$\Rightarrow \frac{3}{p} = \frac{4}{1-p}$$

$$3 - 3p = 4p$$

$$7p = 3$$

$$\boxed{p = \frac{3}{7}}$$

$$\boxed{\log(ab) = \log a + \log b}$$

logarithm:

$$2^3 = 8$$

$$\log_2 8 = 3$$

$$\log_2 2^3 = 3$$

$$\log_{(base)} number = ans$$

$$\log_2 3 = 6$$

$$3^6$$

$$\cancel{\log_2 3}$$

$$\boxed{\log_r a^b = b\log a}$$

$$\boxed{\log(a \cdot b) = \log a + \log b}$$

$$\frac{d}{dp}(\log p) = \frac{1}{p}$$

$$\log_e\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x$$

ln

$$\boxed{\log_a b = c \Rightarrow b = a^c}$$

$$\underbrace{\text{log odds}}_{(-\infty,\ \infty)}$$

$$\Rightarrow \frac{p}{1-p} = e^{(\beta_0+\beta_1 x)}$$

"Betting"

$$\frac{1}{0} = \infty$$

#	R.F	vs	R.N
	201		180

$$\text{odds} = \frac{\#\text{win}}{\#\text{loss}} =$$

$$\text{odds of R.F winning} = \left(\frac{201}{180}\right) \quad \left(\frac{\text{win}}{\text{loss}}\right) = \frac{1}{0} =$$

$$p = \frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}$$

$$= \frac{t}{1+t} = \frac{t/t}{(1+t)/t} = \frac{1}{\frac{1}{t}+1}$$

$$= \frac{1}{1+t^{-1}}$$

$$= \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

$$f(z) = \frac{1}{1+e^{-z}}$$

sigmoid / logistic

Example (Normal data). *Maximum likelihood estimation can be applied to a vector valued parameter. For a simple random sample of n normal random variables,*

$$\mathbf{L}(\mu,\sigma^2|\mathbf{x}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_1-\mu)^2}{2\sigma^2}\right)\cdots\left(\frac{1}{\sqrt{2\pi\sigma^2}}\exp\frac{-(x_n-\mu)^2}{2\sigma^2}\right) = \frac{1}{\sqrt{(2\pi\sigma^2)^n}}\exp -\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

$$\ln \mathbf{L}(\mu,\sigma^2|\mathbf{x}) = -\frac{n}{2}\ln 2\pi\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2.$$

$$\frac{\partial}{\partial\mu}\ln \mathbf{L}(\mu,\sigma^2|\mathbf{x}) = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu) = .\frac{1}{\sigma^2}n(\bar{x}-\mu)$$

Because the second partial derivative with respect to μ is negative,

$$\hat{\mu}(\mathbf{x}) = \bar{x}$$

is the maximum likelihood estimator.

$$\frac{\partial}{\partial\sigma^2}\ln \mathbf{L}(\mu,\sigma^2|\mathbf{x}) = -\frac{n}{\sigma^2} + \frac{1}{(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 = \frac{n}{(\sigma^2)^2}\left(\sigma^2 - \frac{1}{n}\sum_{i=1}^{n}(x_i-\mu)^2\right).$$

Recalling that $\hat{\mu}(\mathbf{x}) = \bar{x}$, we obtain

$$\hat{\sigma}^2(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{x})^2.$$

Example Suppose that X is a discrete random variable with the following probability mass function: where $0 \le \theta \le 1$ is a parameter. The following 10 independent observations

X	0	1	2	3
$P(X)$	$2\theta/3$	$\theta/3$	$2(1-\theta)/3$	$(1-\theta)/3$

were taken from such a distribution: (3,0,2,1,3,2,1,0,2,1). What is the maximum likelihood estimate of θ.

Solution: Since the sample is (3,0,2,1,3,2,1,0,2,1), the likelihood is

$$\begin{aligned} L(\theta) = &\quad P(X=3)P(X=0)P(X=2)P(X=1)P(X=3) \\ \times &\quad P(X=2)P(X=1)P(X=0)P(X=2)P(X=1) \end{aligned}$$

Substituting from the probability distribution given above, we have

$$L(\theta) = \prod_{i=1}^{n} P(X_i|\theta) = \left(\frac{2\theta}{3}\right)^2\left(\frac{\theta}{3}\right)^3\left(\frac{2(1-\theta)}{3}\right)^3\left(\frac{1-\theta}{3}\right)^2$$

Let us look at the log likelihood function

$$\begin{aligned} l(\theta) &= \log L(\theta) = \sum_{i=1}^{n}\log P(X_i|\theta) \\ &= 2\left(\log\frac{2}{3} + \log\theta\right) + 3\left(\log\frac{1}{3} + \log\theta\right) + 3\left(\log\frac{2}{3} + \log(1-\theta)\right) + 2\left(\log\frac{1}{3} + \log(1-\theta)\right) \\ &= C + 5\log\theta + 5\log(1-\theta) \end{aligned}$$

where C is a constant which does not depend on θ. It can be seen that the log likelihood function is easier to maximize compared to the likelihood function.

Let the derivative of $l(\theta)$ with respect to θ be zero:

$$\frac{dl(\theta)}{d\theta} = \frac{5}{\theta} - \frac{5}{1-\theta} = 0$$

and the solution gives us the MLE, which is $\hat{\theta} = 0.5$.

randomly Prediction

$50 \; X \quad \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_{50} \end{bmatrix}$

$(50 \times 2) \quad (2 \times 1)$

$= (50 \times 1)$

$f(x) = \frac{1}{1+e^{-x}}$

$S = \sigma$ sigma $\quad S = \Sigma$

$$\sigma(3) = \frac{1}{1+e^{-3}}$$

$$\frac{d}{dx}\sigma(x)$$

$$\frac{d}{dx}\left(\frac{1}{x}\right) = -\frac{1}{x^2} = -x^{-2}$$

chain rule of derivati

$$\frac{d(uv)}{dx} = uv' + u'v$$

$$\sigma(x) = \frac{1}{1+e^{-x}}$$ (Sigmoid function)

$$\frac{d}{dx}\sigma(x) = \frac{d}{dx}\left[\frac{1}{1+e^{-x}}\right]$$

$$= \frac{d}{dx}\left(1+e^{-x}\right)^{-1}$$

$$= -(1+e^{-x})^{-2}(-e^{-x})$$

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

$$= \frac{1}{1+e^{-x}} \cdot \frac{e^{-x}}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \cdot \frac{(1+e^{-x})-1}{1+e^{-x}}$$

$$= \frac{1}{1+e^{-x}} \cdot \left(\frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}}\right)$$

$$= \frac{1}{1+e^{-x}} \cdot \left(1 - \frac{1}{1+e^{-x}}\right)$$

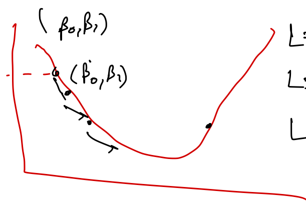
$$\boxed{\frac{d\sigma(x)}{dx} = \sigma(x) \cdot (1-\sigma(x))}$$

loss :

$$L = (\text{Predicted} - \text{True})^2$$

$$L = |\text{Predic} - \text{True}|$$

$$L = ((0,1) - 0/1)^2$$



loss function

For each training data-point, we have a features x_i and an observed class, y_i.

The probability of the class is p, if $y_i = 1$, or $1-p$ if $y_i = 0$

Head (H/T)

$y_i = 1$; p

$y_i = 0$; $(1-p)$

Cross Entropy loss L_{CE} =

$$P(Y|X) = p(x_i)^{y_i} \cdot (1-p(x_i))^{1-y_i}$$

$$P(Y|X) = \hat{y}^{y_i} \cdot (1-\hat{y})^{(1-y)}$$

$$\hat{y} = \sigma(\beta_0 + \beta_1 \cdot x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

estimate/prediction → $\hat{y}$

true / actual → y

$$\log l = y\log\hat{y} + (1-y)\log(1-\hat{y})$$ (using log-identities)

$$= y\log\sigma(\beta_0+\beta_1\cdot x) + (1-y)\log(1-\sigma(\beta_0+\beta_1\cdot x))$$

$$-\frac{\partial \log l}{\partial \beta_j} = \frac{y}{\sigma(\beta_0+\beta_1\cdot x)}\frac{\partial\sigma(\beta_0+\beta_1\cdot x)}{\partial\beta_j} + \frac{(1-y)}{(1-\sigma(\beta_0+\beta_1\cdot x))}\frac{\partial(1-\sigma(\beta_0+\beta_1\cdot x))}{\partial\beta_j}$$

$$= \frac{y}{\sigma(\beta_0+\beta_1\cdot x)}\frac{\partial\sigma(\beta_0+\beta_1\cdot x)}{\partial\beta_j} - \frac{(1-y)}{1-\sigma(\beta_0+\beta_1\cdot x)}\frac{\partial\sigma(\beta_0+\beta_1\cdot x)}{\partial\beta_j}$$

$$= \left[\frac{y}{\sigma(\beta_0+\beta_1\cdot x)} - \frac{1-y}{1-\sigma(\beta_0+\beta_1\cdot x)}\right]\frac{\partial\sigma(\beta_0+\beta_1\cdot x)}{\partial\beta_j}$$

$$= \frac{y - y\sigma(\beta_0+\beta_1\cdot x) - \sigma(\beta_0+\beta_1\cdot x) + y\sigma(\beta_0+\beta_1\cdot x)}{\sigma(\beta_0+\beta_1\cdot x)(1-\sigma(\beta_0+\beta_1\cdot x))} \cdot \sigma(\beta_0+\beta_1\cdot x)(1-\sigma(\beta_0+\beta_1\cdot x))\frac{\partial(\beta_0+\beta_1\cdot x)}{\partial\beta_j}$$

$$= \frac{y - \cancel{y\sigma(\beta_0+\beta_1\cdot x)} - \sigma(\beta_0+\beta_1\cdot x) + \cancel{y\sigma(\beta_0+\beta_1\cdot x)}}{\cancel{\sigma(\beta_0+\beta_1\cdot x)(1-\sigma(\beta_0+\beta_1\cdot x))}} \cdot \cancel{\sigma(\beta_0+\beta_1\cdot x)(1-\sigma(\beta_0+\beta_1\cdot x))} \cdot x$$

$$= (y - \sigma(\beta_0+\beta_1\cdot x)) \cdot x$$

$$\frac{\partial \log l}{\partial \beta_j} \implies (\hat{y} - y)x$$

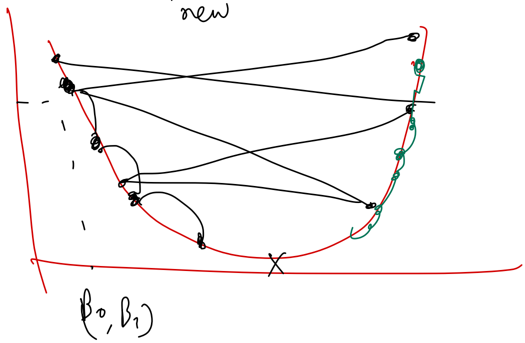
Stochastic Gradient Descent

$$\beta = \beta - \eta \cdot \nabla l(\beta)$$

Gradient Descent :

Loss

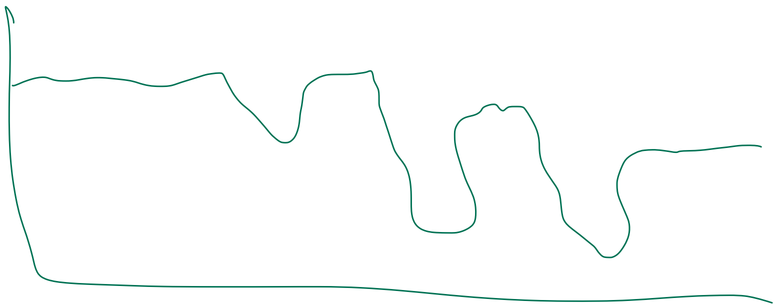
$$\beta_{new} = \beta_{old} - \eta \times \nabla(loss)$$



learning rate

high - functing loss. possibility of missing the min.

Loss ?



BLUE → Estimator

B → Best

L → Linear

U → Unbiased

E → Estimator

$$[0, 1]$$

Dataset → Logistic Regression → 0.6

Pass (1) vs Fail (0)

$$\begin{bmatrix} 2, & 1, & 3, & 4 \\ 3, & 1, & 4, & 7 \\ 9, & 11, & 5, & 9 \end{bmatrix}$$

$$\sqrt{(3-2)^2 + (1-1)^2 + (4-3)^2 + (7-4)^2}$$

Big Picture of ML Modelling

Business Dis ion → Question

"whom should be distribute our loans to"

cussi

Data

(Relationship # Yr. w/ firm)

(lacs)

Male - 1
Female 0

Age, Tenure, Annual Income, Credit Score, Gender, Yrs. of Educ, loanstatus

35,	3,	~~150~~ 15,	700,	1,	15,	0
40,	10,	10,	800,	1,	12,	1

1000

0, 1
(No) (Yes)

X → Y

loan_details

(PII)

mycode
ddt: '''
Select
Age, T, A, C, G, Y
from
loan_details
'''

"Dataset" (data.xlsx)

"Classification Model"

Logistic Regression → (85%)

Decision Tree → (90%)

SVM → (88%)

KNN → (70%)

NN DL Ensem → (95%)

df = sql $ mycode

("Shuffle the dataset")

← 70% → ← 30% →

Training dataset | Test Data

← Dataset →

Missing values

- Impute missing value using mean.

Ratio: 70%

700

Train data (X | Y) → logistic Regres. → [0, 1, ...] (β₀, β₁, β₂ ...)

Test Data

300

X	y
25, 5	

X

1, 1

Learned Logistic Regression

$\sigma(0.51 + 1.71 \times Age + 2.0 \times Tenure)$

25 + 5

$(0,1)$

$\sigma(0.66) \rightarrow 1$

M/O	GT
1	0
1	1
0	0
0	0

✓ → 3

X → 1

$\left(\frac{3}{4}\right) = 75\%$

5 +1

7 +4

9 +3

3 +3 / 6

Judicial Decision: Life sentence

Actual Event: Y, N

Threshold

sen

spec

(Test dataset)

"Ground Truth"

'0.5'

"model output"

		Real Label	
		Positive	**Negative**
Predicted Label	**Positive**	True Positive (TP)	False Positive (FP)
	Negative	False Negative (FN)	True Negative (TN)

$$\text{Precision} = \frac{\sum TP}{\sum TP + FP}$$

$$F1 = \frac{2 \times P \times R}{P + R}$$

$$\text{Recall} = \frac{\sum TP}{\sum TP + FN}$$

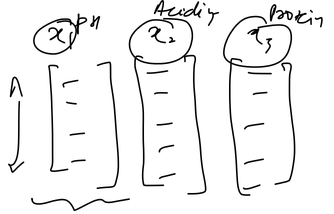
$$\text{Accuracy} = \frac{\sum TP + TN}{\sum TP + FP + FN + TN}$$

Clustering (unsupervised)

$\downarrow$

(no labels)

$S = \{x_1, x_2 .. x_n\}$ Dataset



(1500X1)

Given dataset and a number of clusters 'K'.

wants Divide the dataset into 'K' groups

$$\underbrace{A_1 \;,\; A_2 \;,\; A_3 \;..\;,\; A_K}_{K \text{ clusters / groups / partitions}}$$

Such that
want to have!

(I) $A_i \neq \phi$ (empty) [no clusters should be empty]

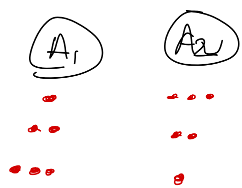
(II) $A_i \cap A_j = \phi \;\; \forall i, j$ [no two clusters should have common data]

(III) $\bigcup_{i=1}^{K} A_i = S$ (All points are associated with some or other cluster)

$$\boxed{\mathcal{L}(P(A_1, A_2 .. A_K)) = \sum_{i=1}^{K} \sum_{x \in A_i} d(x_i, A_i)}$$

take point in A_i and find the distance b/w x and A_i mean (i^{th} cluster) and do this for all such clusters

F No. of clusters = 2, no. of points 4



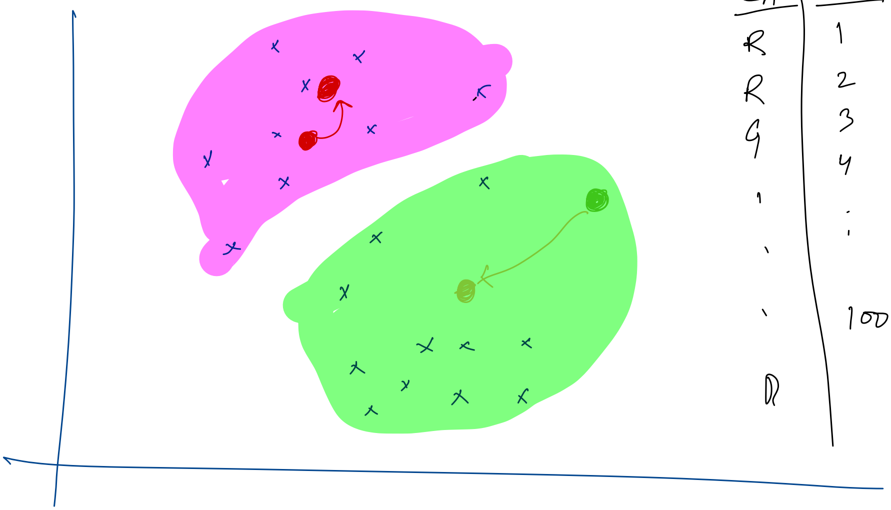
# of clusters

$$\frac{2^{n-2}}{2} \Rightarrow \frac{2^{4-2}}{2} = 2$$

$$S(n, K) = \frac{1}{K!} \sum_{i=0}^{K} (-1)^i \binom{K}{i} (K-i)^n$$

Cluster n data points in K clusters / partitions $= O(K^n)$

(Stirling's number of 2nd kind)



C_A	P_i	G_q	R_q
R	1	5	3
R	2	2	1
G	3	.	
.	4	.	
.	:	.	
.	100		
D			

Jaccard Similarity

$A = \{3, 4, 7, 9\}$ $A \cap B = \{3, 4\}$

$B = \{1, 2, 3, 4\}$ ~~$A \cup B =$~~

$A \cup B = \{1, 2, 3, 4, 7, 9\}$

$$\frac{|A \cap B|}{|A \cup B|} = \frac{2}{6} = \left(\frac{1}{3}\right) = 33\%$$

$= 1 - 33\%$

$= 66\%$

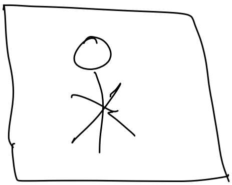
$A = \{2, 3, 4\}$

$B = \{1, 8\}$

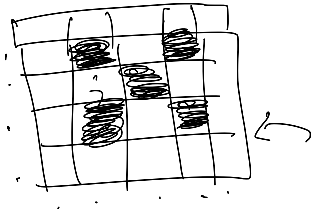
$= \frac{0}{\sim}$

$= 0$

$= 1$



Black / white



$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(5 X 5)

(25px)

(0 – 255)

dark — black

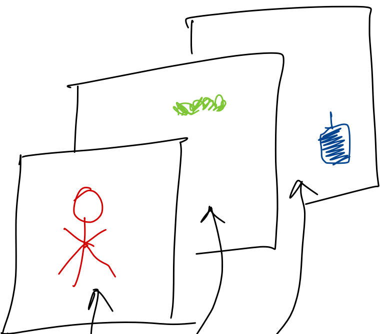
light — white

Gray scale images

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 0 & \cancel{80} & 50 & & \end{bmatrix}$$

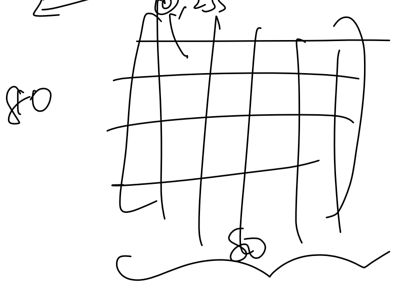
$\downarrow$

[0, 255]



R G B

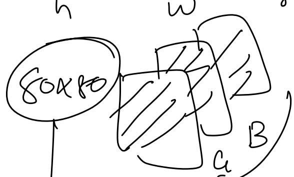
Pure red ← 0, 255 → absence of red

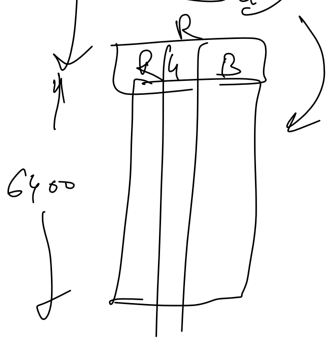


× 3

80 X 80 × 3

h w rgb





Dimensionality Reduction

$$\rightarrow \left| \begin{bmatrix} 2-\lambda & -1 \\ -1 & 2-\lambda \end{bmatrix} \right|$$

$$\Rightarrow (2-\lambda)^2 - (-1)(-1) = 0$$

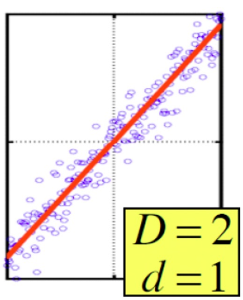
$$\Rightarrow$$

λ's are e.values

day / customer	We 7/10/96	Th 7/11/96	Fr 7/12/96	Sa 7/13/96	Su 7/14/96
ABC Inc.	1	1	1	0	0
DEF Ltd.	2	2	2	0	0
GHI Inc.	1	1	1	0	0
KLM Co.	5	5	5	0	0
Smith	0	0	0	2	2
Johnson	0	0	0	3	3
Thompson	0	0	0	1	1

Goal of dimensionality reduction is to discover the axis of data!

Rather than representing every point with 2 coordinates we represent each point with 1 coordinate



Rank: No. of linearly independent $\left(\frac{\text{columns}}{\text{rows}}\right)$ of "A".

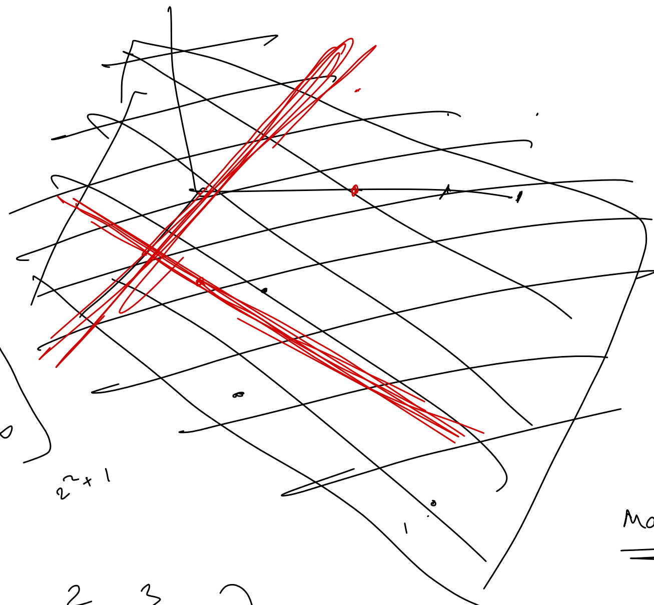
$$A = \begin{bmatrix} 1 & 2 & 3 \\ 0 & 1 & 2 \\ 1 & 3 & 5 \end{bmatrix} \qquad R_3 \leftarrow R_1 + R_2$$

$$\text{Rank}(A) = 2$$

Basis:

$$\underset{X}{[1 \quad 2 \quad 3]} \qquad \underset{Y}{[0 \quad 1 \quad 2]}$$

new cordinates will be: $[1 \quad 0]$ $[0 \quad 1]$ $[1 \quad 1]$



$$\begin{bmatrix} 2 & 0 & 0 \\ 0 & 1.5 & 0 \\ 0 & 0 & 0.6 \end{bmatrix}$$

0

2

2^2+1

proof:

$a^2+b^2 = (a+b)^2 - 2ab$

$LHS = a^2+b^2 \times 0$

$= 0$

$RHS = \underline{\quad} \times 0$

$= 0$

Matrix Decomposition

$$\underbrace{\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 0 & 1 & 2 \end{bmatrix}}_{A} = \underset{3\times 3}{\begin{bmatrix} A \end{bmatrix}} \times \underset{3\times 3}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}} \times \underset{\cancel{3\times 3}}{\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}}$$

patches

one

$$\begin{array}{c} v_1 \\ v_2 \\ v_3 \\ v_4 \end{array} \begin{bmatrix} - & - & - \\ - & - & \\ - & - & \\ - & . & \end{bmatrix} = \begin{bmatrix} \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots \end{bmatrix} \begin{bmatrix} \quad \end{bmatrix}$$

$$\begin{bmatrix} 3 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 0.5 \end{bmatrix}$$

$$= 3^2 + 2^2 + \left(\frac{1}{2}\right)^2$$

$$= 9 + 4 + 0.25$$

$$= 13.25$$

50

$$\frac{(3)^2}{13.25},\ \frac{(2)^2}{12.25},\ \frac{(0.5)^2}{11.25}$$

$$\frac{9}{13.25} \quad \frac{4}{12.5} \quad \frac{1}{4 \times 12.5}$$

67% 30% 3%

Img → rank (80) SVD → $U\ \Sigma\ V^T$

50 MP (5MB) 95%

80%

15%

2%

20

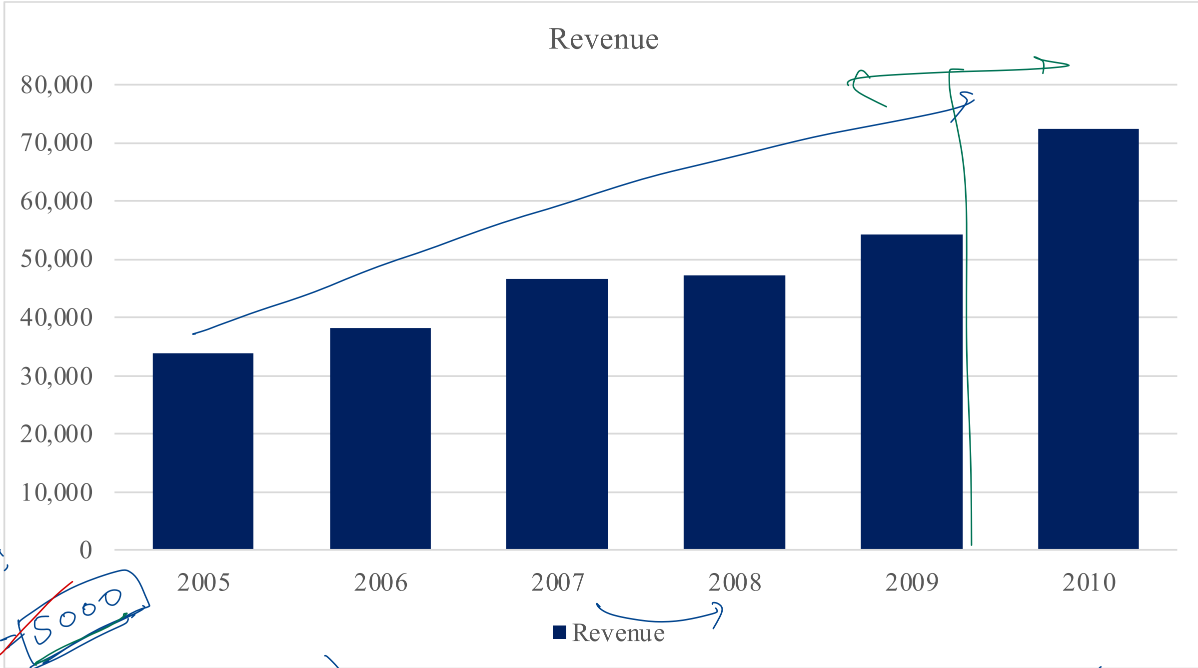
80

15L
Turn
36%
$15 \times (1.30)^4 = 42L$ 21L 24L 1000

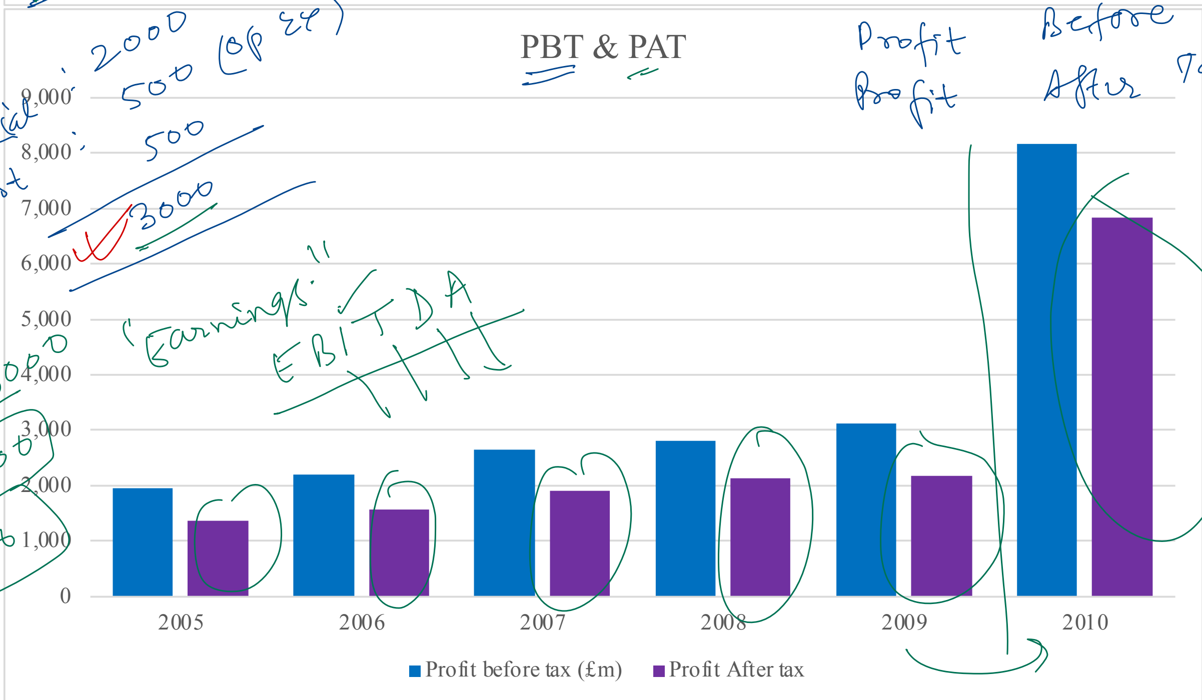
Data Driven Consumer Analytics at Superstore

Superstore is a multinational groceries and general merchandise retailer headquartered in Garden City, England. Shops of **Superstore** are larger, mainly out-of-town hypermarkets that stock nearly all product ranges, although some are in the heart of town centers and inner-city locations. Recently they have started a home shopping service through their website. **Superstore** observed a significant growth by opening the website operations.

0 → 24%



raw material
Revenue: 50 x 100 = 5000



Profit Before Tax.
Profit After Tax

PBT costs: raw material: 2000
labour cost: 500 (op ex)
Misc: 500
3000
"Earnings" EBITDA
Profit Income: 3000
Tax: 600
PAT: 2400

Figure 1 Revenue & Profit (in £ millions)

According to NY Bank retail analyst, "**Superstore** has pulled off a trick (online store) that I'm not aware of any other retailer had thought earlier". However due to internet penetration & democratization, many competitors supermarket chains have opened their website operations and this is driving lots of consumer leaving for other competing stores. This is known as consumer **churn**. Formally, Churn is the phenomenon where customers of a business no longer purchase or interact with the business.

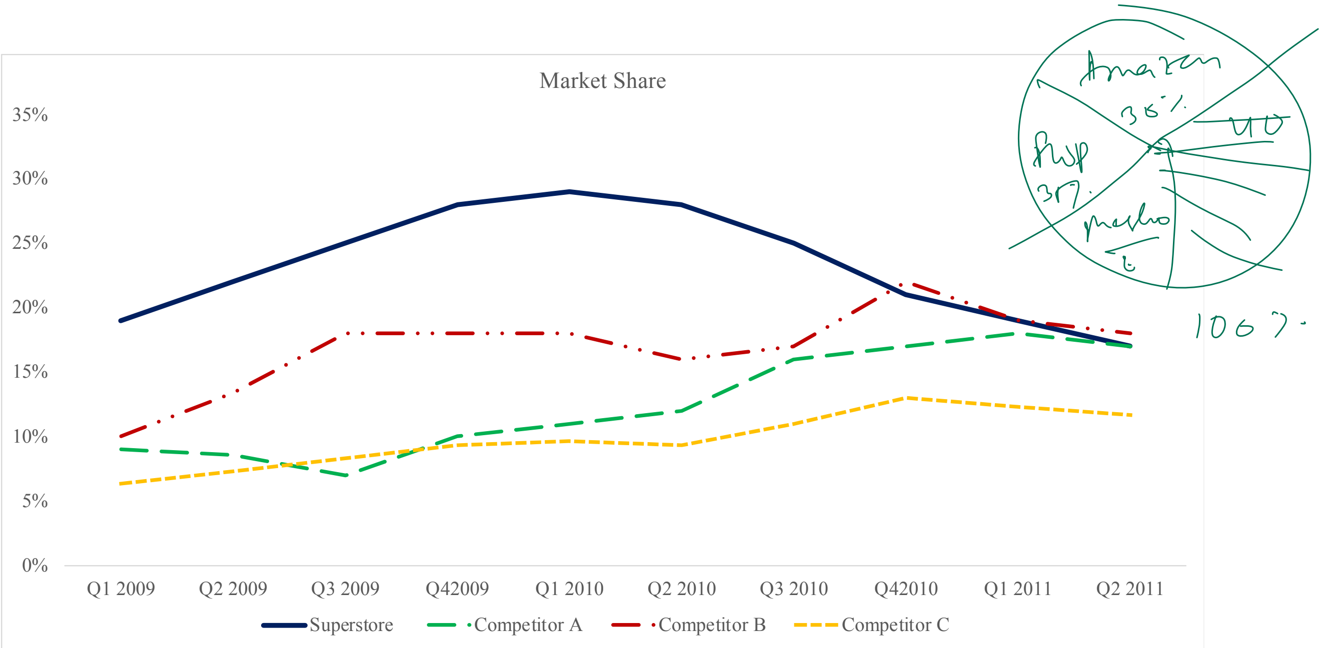


Figure 2 Market Share

Superstore consulted the MBB management consulting firm to help with their falling market share. MBB remarked "**Superstore** has plenty of consumer data. It is now imperative to use a data driven strategy to define a marketing campaign".

You joined the company in the aftermath of the adaptation of a standardized solution that the consulting company suggested, and are tasked with **defining data driven marketing campaign to mediate consumer attrition**.

Your fundamental task is to ---

1. Understand the consumer base according to purchase & spending patterns and call out any specific trends/spikes.
2. Group your customers into high engaged, mid and low engaged. (Hint: There is a very intuitive RFM framework. See more details in Appendix 1.)
3. Build a data driven mechanism to identify which customers are likely to churn.
4. You are also given £25000 to spend for marketing campaigns. You would like to give some coupons as ~~incentive for~~ the customers to perform their next transaction (hence reducing the churn). Which all customers would you like to target.

Here is the purchase transaction data containing 540k rows https://github.com/vntkumar8/musical-spoon/raw/main/Online%20Retail.xlsx

Snapshot of data

InvoiceNo	StockCode	Description	Quantity	InvoiceDate	UnitPrice	CustomerID	Country
536365	85123A	WHITE HANGING HEAI	6	01/12/10 8:26	2.55	17850	United Kingdom
536365	71053	WHITE METAL LANTER	6	01/12/10 8:26	3.39	17850	United Kingdom
536365	84406B	CREAM CUPID HEARTS	8	01/12/10 8:26	2.75	17850	United Kingdom
536365	84029G	KNITTED UNION FLAG	6	01/12/10 8:26	3.39	17850	United Kingdom

Appendix 1

RFM Segmentation
RFM segmentation is a method to identify groups of customers for special treatment. It is used as a tool to improve customer marketing.

What is RFM Segmentation?
RFM analysis allows marketers to target specific clusters of customers with communications that are much more relevant for their particular behavior – and thus generate much higher rates of response, plus increased loyalty and customer lifetime value. Like other segmentation methods, an RFM model is a powerful way to identify groups of customers for special treatment. RFM stands for recency, frequency and monetary – more about each of these shortly. Marketers typically have extensive data on their existing customers – such as purchase history, browsing history, prior campaign response patterns and demographics – that can be used to identify specific groups of customers that can be addressed with offers very relevant to each.

While there are countless ways to perform segmentation, RFM analysis is popular for three reasons:

- It utilizes objective, numerical scales that yield a concise and informative high-level depiction of customers.
- It is simple – marketers can use it effectively without the need for data scientists or sophisticated software.
- It is intuitive – the output of this segmentation method is easy to understand and interpret.

What are Recency, Frequency and Monetary?
Underlying the RFM segmentation technique is the idea that marketers can gain an extensive understanding of their customers by analyzing three quantifiable factors. These are:

- **Recency:** How much time has elapsed since a customer's last activity or transaction with the brand? Activity is usually a purchase, although variations are sometimes used, e.g., the last visit to a website or use of a mobile app. In most cases, the more recently a customer has interacted or transacted with a brand, the more likely that customer will be responsive to communications from the brand.
- **Frequency:** How often has a customer transacted or interacted with the brand during a particular period of time? Clearly, customers with frequent activities are more engaged, and probably more loyal, than customers who rarely do so. And one-time-only customers are in a class of their own.
- **Monetary:** Also referred to as "monetary value," this factor reflects how much a customer has spent with the brand during a particular period of time. Big spenders should usually be treated differently than customers who spend little. Looking at monetary divided by frequency indicates the average purchase amount – an important secondary factor to consider when segmenting customers.

YouTube tutorial: https://youtu.be/i-HNJZeOOMY

Churn: Netflix, Spotify, Airtel Black, DTH, OTT

Subscription based

cust_id → 1 x

Amazon Ecommerce, PayTM, Assumption

$$\frac{\text{Tl orders} \mid 5000}{\text{\# Customers}} \quad \text{- 2 yrs (1000)} \quad = \frac{5000}{1000} = 5$$

(2 yrs)

$$2021 \rightarrow \frac{(500)}{(100)} = 5$$

← 3 → ← 2m ← 5m → ← 6 →

$$\frac{3, 2, 5, 6}{} = 4$$

$$\underbrace{2, 3, 5, 6} = 4$$

5 %

Marketing Analytics

— Acquisition — Retention — Churned

↑ (1st visitors)

↑ 2nd time (visitors)

↓ (not coming)

C1, C2, C3, ..., C10

cust | Colgate | 20/- | M | TAD money

10 mg

400 mg

	1	2	3	4	5	6
May - 2010	✓					
Apr - 2010	✓	✓				
Mar - 2010	✓	✓	✓			
Feb - 2010	✓	✓	✓	✓		
Jan - 2010	100%	25%	20%	15%	4%	×
Dec - 2009	100%	30%	20%	15%	5%	2%

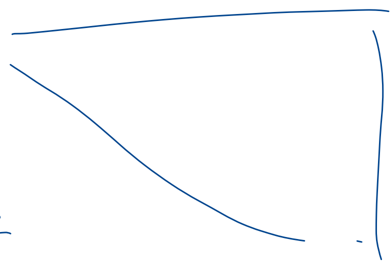
100 →

30% → 20% → 15%

Jan Feb Mar

Dec

How many active →

× 

RFM :

method to group / segment your customers into bins

100

- Very loyal
- Essential
- Early life
- Churned

Recency: when did customer made its last purchase txn.

16th nov.
Frequency: How often does customer make their purchase

17 nov
Monetary val: How much money customer has spent

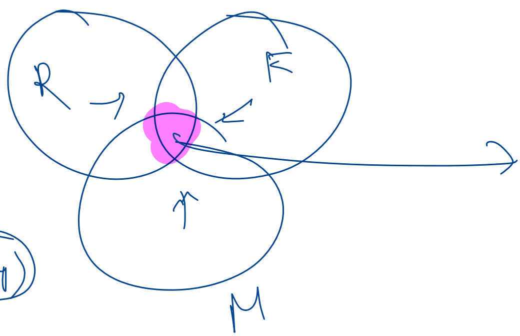
C1 15th nov

C2 3rd nov

18 day

3rd Dec.

30 day



High val. Customer

High spend, very frequently and last purchase was very recently

Recency ↓ Frequency ↑ Monterey ↑

cust_id, ~~Amo~~ item, date, Price.

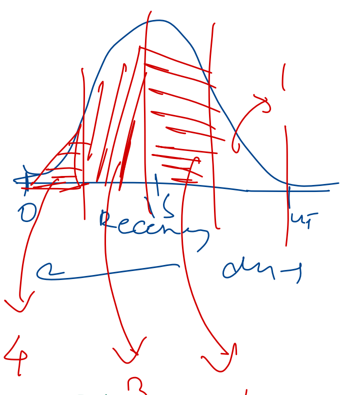
3
5
6

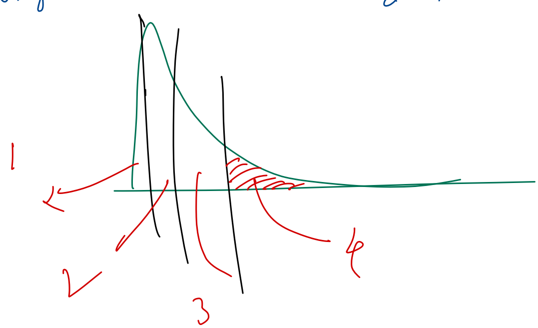
Recency
15 day

today (1)

Last order day

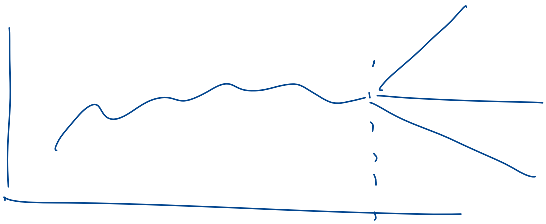
(frequency)





Exponential Smoothing:

"Forecasting"



2 past values

$\alpha = 0.6$

(0, 1)

previous, curr

current past

[1, 3, 3, 6, 7, 8, 9 . . .]

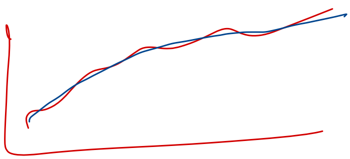
$\alpha \times$ Current value $+$

$(1-\alpha) \times$ Past value

$0.6 \times 3 + 0.4 \times 3$

$= 1.8 + 1.2$

$= 3$



$0.6 \times 6 + 0.4 \times 3 = 3.6 + 1.2 = 4.8$

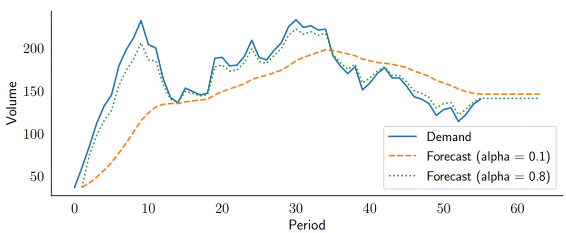


Figure 3.2: Simple smoothing

Simple/single exponential smoothing: This smoothing can be used for making forecasts based in a time series that has no trend and seasonality. Simple exponential smoothing does not do well when there is a trend in the data. Double exponential smoothing: This type of exponential smoothing comes with the support for trend components of time series.

simple exponential smoothing.

In the math notation x_t is `ts[t]`, and $\hat{x}_t$ is our prediction for x_t:

Initial conditions

- $s_0 = x_0$. This is our initial guess.
- $\hat{x}_0$ is undefined. We can't call the first guess a prediction since it's actually the first observation.

For $t > 0$

- $s_t = \alpha(x_t) + (1-\alpha)s_{t-1}$
- $\hat{x}_t = s_{t-1}$

When α is closer to 1 the model is more sensitive to recent observations. When α is closer to 0 the model is more sensitive to past observations.

Simple/single exponential smoothing: This smoothing can be used for making forecasts based in a time series that has no trend and seasonality. Simple exponential smoothing does not do well when there is a trend in the data. Double exponential smoothing: This type of exponential smoothing comes with the support for trend components of time series.

Now we will implement double exponential smoothing. For our implementation the formula is as follows:

- $s_0 = x_0$
- $b_0 = 0$
- $\hat{x}_0$ is undefined

$s_1 = \alpha x_1 + (1-\alpha)(S_0 + b_0)$

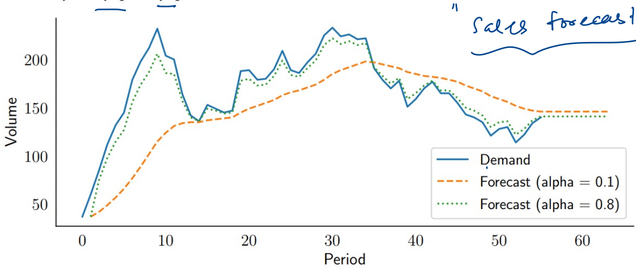
$\boxed{S_1 = \alpha x_1 + (1-\alpha)(x_0 + 0)}$

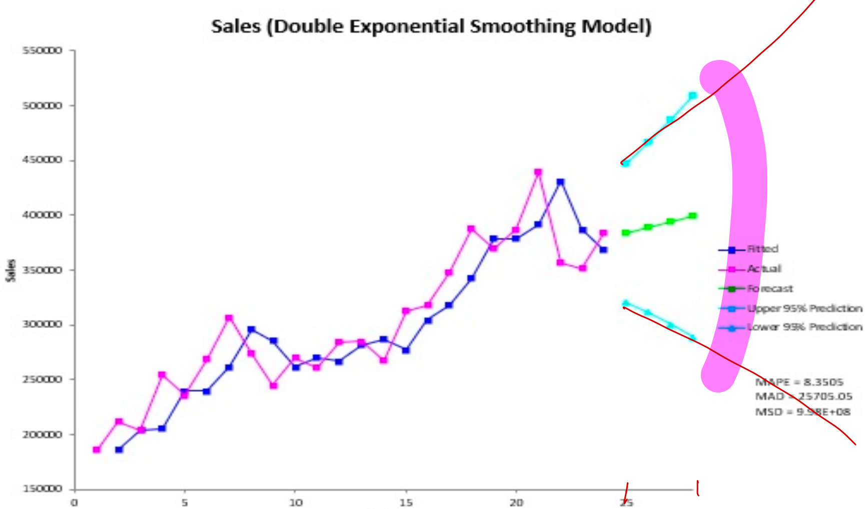
For $t > 0$:

- $s_t = \alpha x_t + (1-\alpha)(s_{t-1} + b_{t-1})$
- $b_t = \beta(s_t - s_{t-1}) + (1-\beta)b_{t-1}$
- $\hat{x}_t = s_{t-1} + b_{t-1}$

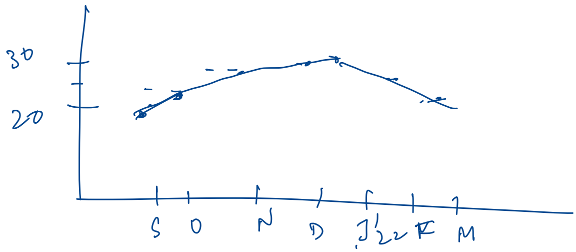
$S_{17} + b_{17} \Rightarrow S_{18}, b_{18}$

"Sales forecast"





100



Consumer price index

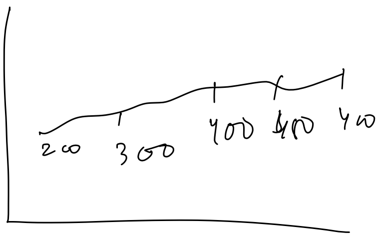
[20, 24, 28, 29, 30, 29]

$\frac{24-20}{20}$; $\frac{28-24}{24}$; $\frac{29-28}{24}$

$\frac{28-20}{20}$, $\frac{29-24}{24}$, $\frac{30-28}{28}$

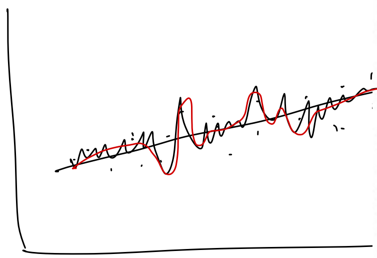
$\frac{1}{5}$, $\frac{1}{6}$, $\frac{1}{29}$, ...

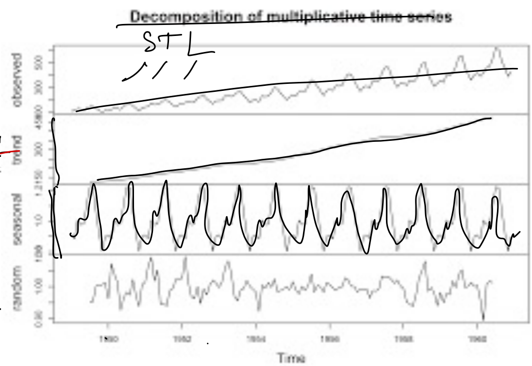
0.20% ; 12% ; ___ ;

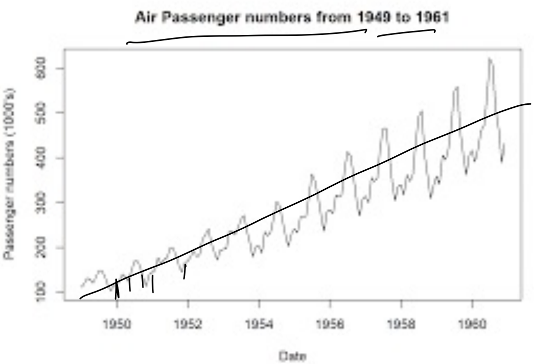


S
– trend

Seasonality

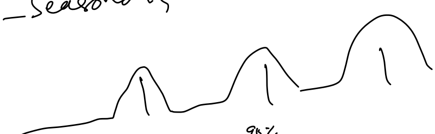






– Seasonality

MA



$R^2 \rightarrow (0,1) \rightarrow 90\%$

MSE → lesser the better

Actual: 90, 150, 250, 400

Predicted: 100, 200, 250, 300

MSE: $= (100-90)^2 + (200-150)^2 + (250-250)^2 + (300-400)^2$

Data → $\boxed{\text{Model}}$ → o/p

Double Exp. Smoothing:

$(\alpha, \beta) \rightarrow (0.2, 0.4, 0.6, 0.8)$ — 4

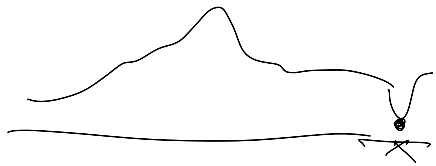
$(0.1, 0.2, 0.3, 0.4, 0.8, 1.)$

$5 \times 4 = \boxed{10}$

(Predict - Actual -)

$\alpha =$

α	β	MSE
0.1	0.2	110
0.1	0.4	210
0.1	0.6	300
0.1	0.8	80



(0.1, 0.8)

β

α

α 0.1 0.2 0.5