DS203

https://tinyurl.com/DS-2024...

Books: Learning data science
Veridical data science

Stats - Business statics for contempory desicon making

ml - An intro to stat. learning

Test - Reasoning and
Prog - intern access
test = assignments

ds203.2023@gmail.com
Prog. for data science
0-2-2-6, 8-10 assign.
(Assignments - 8-10 - No fraud)
10%. 2 sup quiz, 30%. midsem,
30%. project, 30%. endsem.

31st July - (Note: understanding problem/domain knowledge is v. imp)

Data science: ① Broad field
② extracting insights
③ knowledge
④ collecting, cleaning, organizing, analysing, and int data.
⑤ to uncover patterns, trends and meaningful information.
from data

→ takeaway: Data science is a jumble of many things. we need skills in most of this.

Data science utilizes various techniques, methodology, and tools to extract valuable insights from data. ↳ communication is also an important part

Case study: The problem: Optimization of process(es) in a chemical plant.
goals: ① operating parameter's opti.
② throughput prediction
③ Breakdown prediction
④ Predictive maintenance of eqp. → is this most optimised way? How to predict if they are sick?

Let's start:
① data given (Kaggle bains) → 132 cols, 7 yrs of daily average → process:
↓
what tools to use? excel? deeplearn?
↳ difference

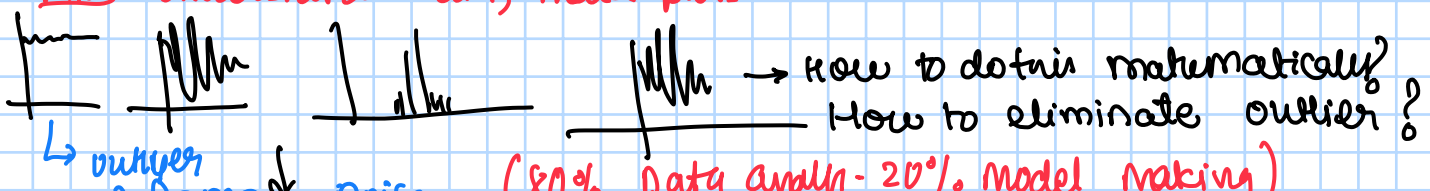
Problem framing → clear problem framing.
↓
Data acquisition → Right data (Operational repo - CRM)
↓
Data preparation → Exploratory data analysis (EDA)
↓
model planning → finalise training model for data set (Reg & Class)
↓
model building → Build model
↓
visual communication → using graphs
↓
Deployment-maintence → to maintain frame iteratively retrain the model

We do preliminary analysis of data:
↳ what is data size?
↳ most app tools?
↳ type of data items present?
↳ missing obs? their perc?
↳ how to handle missing obs?
↳ Are there any outlying obs?
↳ How to handle it?

gender? ↳ F/M → 0 / 1
↳ scaling?
↳ encoding?

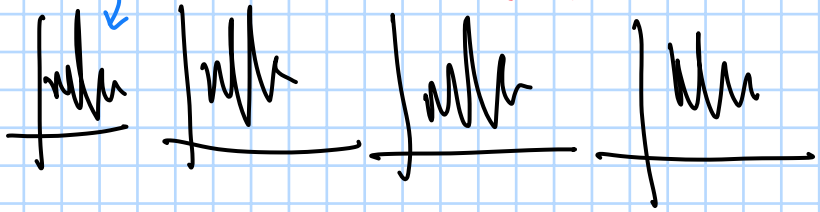
matrix leaving
$y = f(x)$
↳ input variables
↳ unknown
output variables
what is y what is x? cause-effect?

* First: understand data, make plots:

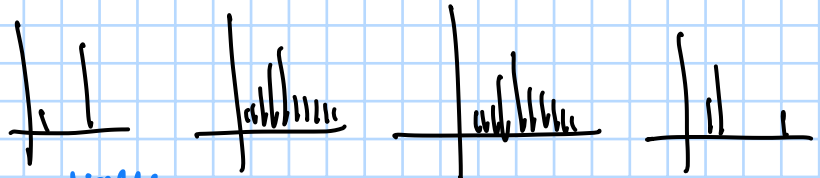


↳ outliers
↳ Remove noise → How to do this mathematically? How to eliminate outlier?

(80% Data analy - 20% Model making)



→ processed (trends with time)



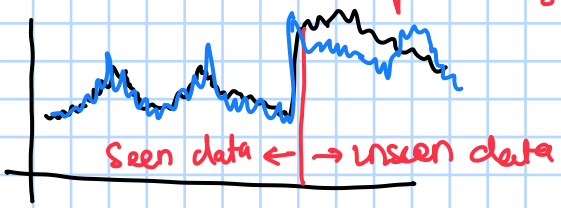
→ Histogram (frequency)

essentially two points

→ correlation)

(Heatmap)

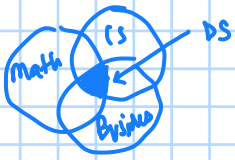
Basic model: Linear regression?



evaluation:
using math and stats, R value, f value ...

data → analysis → prediction / model

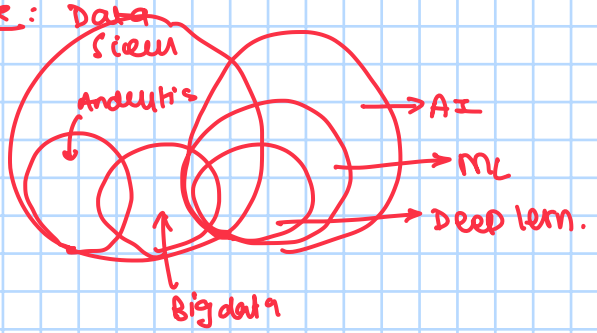
2nd Aug:
- mastering DS
  - ↳ statistical foundations ↳ machine learning fundamentals (DS 2037)
  - ↳ knowledge of tools, prog. (DS 203) ↳ Domain knowledge
  - ↳ comm. skills



Example:
good PPT vs bad ppt
↳ less text ↳ No use of gpt

course - intro
↓
Data
↓
cloud comp

Note: Data Science



Stages: ① Problem statement
↓
② Progg. for data science

Analytics:
① Systematic comp. analysis of data/stats
② Information resulting from analysis of data/stats
③ Discovery and comm of pattern in data.
④ raw data → draw conclusion.

AI: AI is ability of machine to perform fuctions similar to that of a human mind like perceiving, learning, problem solving, etc.

ML: AI technique to handle large and complex data
- → linear regression
- → logistic regression
- → clustering

(See Hotspot)

- supervised learning
- unsupervised learning
- Reinforcement learning

Deeplearning - neural network  Cat / dog

7th Aug : HW : Do notebooks (Moodle)

ML → stats ← we have to do stats

$y = f(x)$
↑ function

what is ML?
what do we have? - Data
- Structured data "Table" / "Excel sheet"

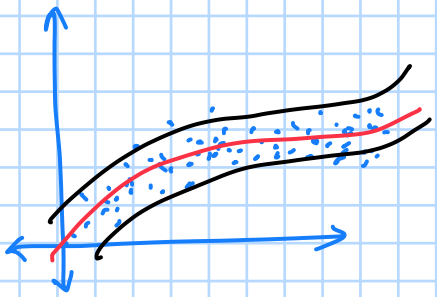
eg. y | x_1 ← only 1 variable

something we want to predict

we want to find patterns, trends, etc.

y | x_1 x_2 x_3 x_4 → $y = f(x)$

x is giving us to predict y.



so $y = f(x)$
↳ model
↳ line
↳ polynomial
↳ something else

levels of measurement:

- NOMINAL ..... DISCRETE
  NO ORDER
  DIFFERENTIATE
  } Gender — M / F
  Red, green, Blue
  classification
- ORDINAL ...
- INTERVAL
- RATIO

see prof notes ←

... Nominal + ordering possible
- grades
- tall, mid, short
- DISCRETE
- ORDER

1	tall 3
2	mid 2
3	short 1
⋮	

continuous problem?
- 20°C v/s 10°C
- unit issues
  Height 4 feet, 2 feet
- Addition / subtraction
- multiplication and division

captures the input correctly

{ x_3 into 3 columns

x_{3_1}	x_{3_2}	x_{3_3}
0	0	1
1	0	0
0	1	0

if

x_1	x_2	x_3
		0
		1
		2
		6
		2

} Nominal
↓
we cannot put this in
↓
we have to change this

How to handle these cases?

$y = f(x)$
↳ nominal / ordinal / intv / ratio

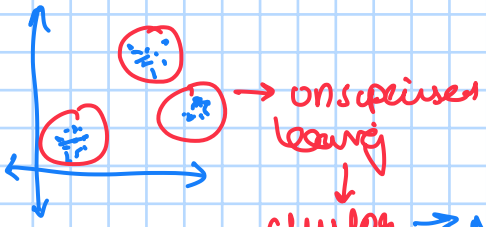
DISCRETE "classification" | CONT "Regression"

label / target / response dependent
↑
$y = f(x)$
↳ predictors / features

PURCHASING_POWER = f(AGE, GENDER, ...)
"DISTANCE" → Imp. role in ML.

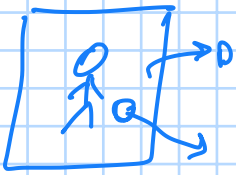
"DISTANCE" → imp role in ML.
↓
Based on particular distances, can we group stuff together?

Note: If Y is known → "Supervised learning"
If Y is not-there → "Unsupervised learning" → clustering



→ unsupervised learning
↓
cluster → new person → clustered in it

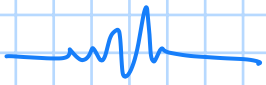
Set of people doing similar stuff ← Amazon Example



→ Pixels x_{100000}
make features

$\vec{x}_{img} = (x_1 \ldots x_{10000})$

These features are very important
But Big ML models can make features for us.

speak →  → features → $x_1\ x_2 \ldots x_n$

Population → Entire possible set of data points

Sample → we can always get - As data is sample space
what to do with sample?
↳ use it to understand
① understand the population
② predict the behavior of the pop

9th Aug:

0 on a scale of 10 test - closest mean no IQ, time 4 am vs 2 am issue (similar to temp - int vs ratio) } Basically 0 can have different meaning for different int vs ratio.

requires point?

0 means 0

$y = f(x)$

Classification v/s Regression

↳ Discrete ↳ cont values

Supervised vs unsupervised

Population v/s sample

↳ entire data (very abstract)

↳ subset of population

Number of Bank accounts - entire data! ↓ Population!

Steps: ① from data train (f) ② from f predict y. (Testing)

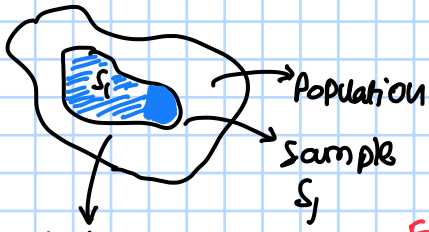
$y = \beta_0 + \beta_1 x$

* Neural network
* Regression
* Classification
* Clustering

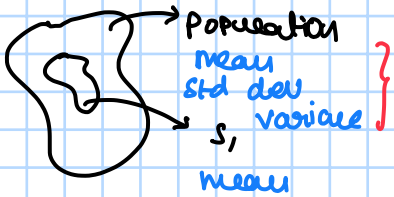
∴ Any data is 'always' a sample

use as much data as we can ← Note 80|20 is not comp

Training ↓ Testing



80% training
20% testing



Parameters

Statistics

Using statistics we want to ESTIMATE parameters.

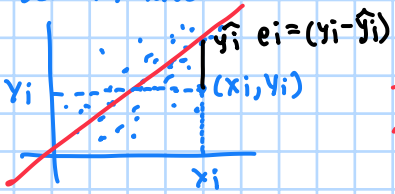
Expected value:

true will always be errors of estimation & it depends on the sample size 'n'

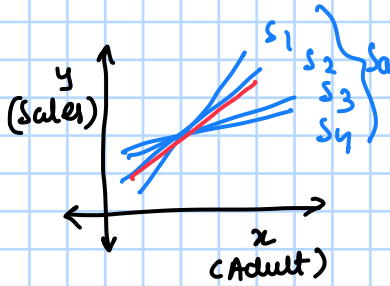
if $n_1 < n_2$

error(n_1) > error(n_2)

Best fit line



THINGS DON'T END WITH f(x)



$y = \beta_0 + \beta_1 x$ ↳ calculated from sample - can we estimate errors in β_0, β_1 → Yes.

$\sum e_i = 0$ (Best fit line)

$\sum e_i^2 = \sum (y_i - \hat{y}_i)^2$ → sum of square of errors

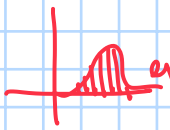
↑ do

why e_i^2 and not $|e_i|$?

$y = \beta_0 + \beta_1 x$

we have to find β_0 and β_1 s.t we minimise β_0 and β_1.

Note: If model is correct - random otherwise errors will show a trend.



$\sum e_i^2 = \sum (y_i - (ax_i + b))^2$

$\frac{\partial}{\partial a}(SSE) = 0 \quad \frac{\partial}{\partial b}(SSE) = 0$

$\sum 2(y_i - (ax_i + b))(-x_i) = 0$

$\sum (-y_i x_i + a x_i^2 + b x_i) = 0$

$\sum y_i x_i = a \sum x_i^2 + b \sum x_i$ —①

$\frac{\partial}{\partial b}(SSE) = 2 \sum (y_i - (ax_i + b))(-1) = 0$

$\sum y_i = \sum a x_i + \sum b$

$\Rightarrow b = \bar{y} - a\bar{x}$ —②

↓ mean of y ↳ mean of x

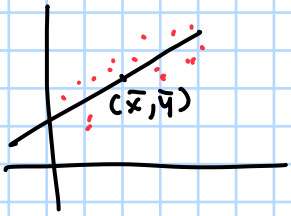
∴ reg. line passes $(\bar{x}, \bar{y})$

as $y = ax + (\bar{y} - a\bar{x})$

$\overline{y_i x_i} = a\overline{x_i^2} + (\bar{y} - a\bar{x})\overline{x_i}$

$\overline{xy} = a\overline{x^2} + (\bar{y} - a\bar{x})\bar{x}$

mean as a model!? → DO



can we use it?
i.e $\forall x$, $\bar{y}$ is solution.
$f(x) = \bar{y}$, $\forall x$.

→ finally it comes down to error

mean is a good enough model

So see the structure of data and then use / make functions.

If we are able to visuise data / small data - use basic models.
If huge / big data - use big / heavy models.

14th Aug : see Beta 0 vs Beta 1 — DO

$a = \frac{\bar{xy} - \bar{x}\,\bar{y}}{(\bar{x}^2 - \bar{x}^2)}$

$b = \bar{y}\,\frac{\bar{x}^2 - \bar{x}.\bar{xy}}{\bar{x}^2 - \bar{x}^2}$

$y = ax + b$ ← dependent
$= \beta_1 x + \beta_0$ → independent variable

S. Linear Regression

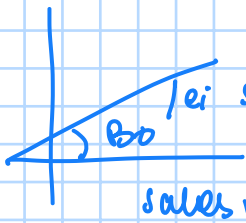
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$ ← dependent
} multiple ind var → M. linear reg

S.L.R

$\beta_0 = b$ = intersection → β_0 = collection of stuff left out
$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots$
we can never identify all β_i

Why errors are imp?
- if model correct → Residual errors have to be random

force fit
error plot



$SSE = \Sigma e_i^2$
$MSE = \sqrt{\Sigma e_i^2}$

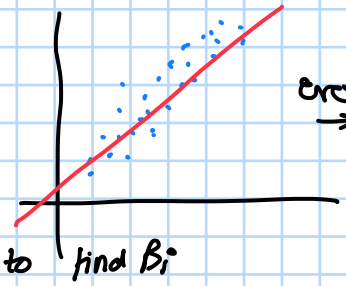
sales vs ad

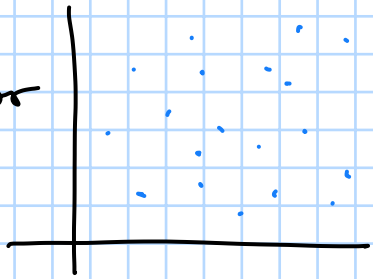
How to use error metric
- Absolute
- Relative

MAE
MSE
RMSE

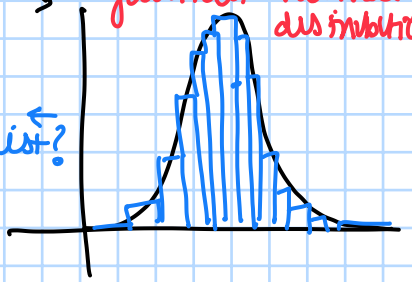
if RMSE = 634
or 2.35
then 2.35 ~ 6.34 for error

if random :





Histogram of error:
gaussian normal distribution



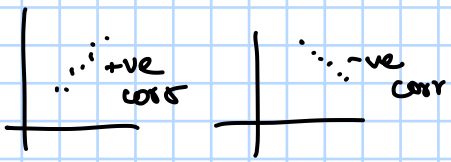
Gradient descent
a, b → algebraically but for M.L.R we use numerical method to find β_i
↑ gradient decent

dq: why normal dist?

see
Error math stats

R square -
SST = measure of total var

$\Sigma (x_i - \bar{x})(y_i - \bar{y})$ ~ correlation



$= \Sigma (y_i - \bar{y})^2$
$= \Sigma (y_i - \hat{y}_i + \hat{y}_i - \bar{y})^2$
$= \Sigma (y_i - \hat{y}_i)^2 + \Sigma (\hat{y}_i - \bar{y})^2 + 2(\ )(\ )$ → why zero?
$SST = SSE + SSR + 0$
↓ orthogonality (see)

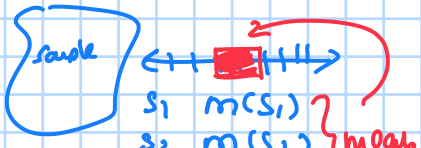
$1 = \frac{SSE}{SST} + \frac{SSR}{SST}$

$1 - \frac{SSE}{SST} = \frac{SSR}{SST} = R\text{-squared}$

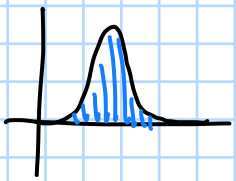
HW : - correlation
- correlation coeff

R^2 = sq of correlation coeff 'r' b/w x and y

- See all stat usage.



Histogram of means:



16th Aug:

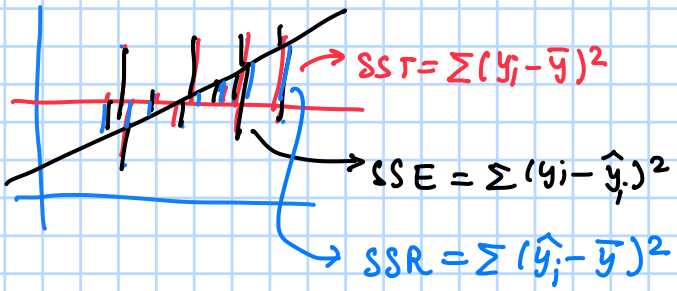
=RAND() → variable by definition is random 'ei vs xi'

Statistical numbers:

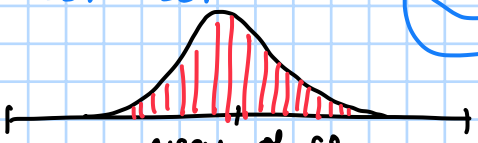
R^2 = ratio of $\frac{SSR}{SST}$ → do: minimum value

$r = \frac{\Sigma(x_i-\bar{x})(y_i-\bar{y})}{\cdot\cdot\cdot}$ ~ correlation coeff ~ x ranges wrt its mean how does y change.

r^2 is sq of correlation coeff

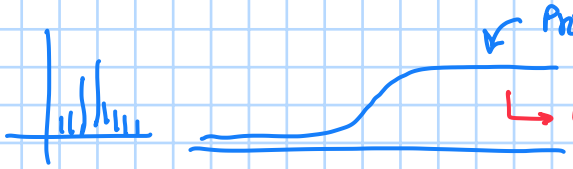


$1 - \frac{SSE}{SST} = \frac{SSR}{SST} = R^2$



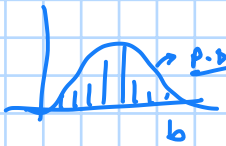
mean of S_i will form this kind of distribution

Given - 1000 data points - distribution



↳ cumulative probability function = $\int_{-\infty}^{x} PDF\,dx$

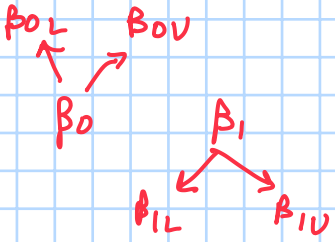
Note: prob at a point = 0 but p at a particular region can be found.



$\int_{-\infty}^{\infty} PDF = 1$

$\int_{a}^{b} PDF.dx$ = area = prob b/w a and b

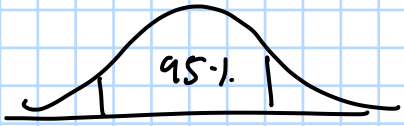
$y = \beta_0 + \beta_1 x$ Statistics → used to estimate parameters



parameter - associated with pop
statistic - associated with sample

Confidence interval:

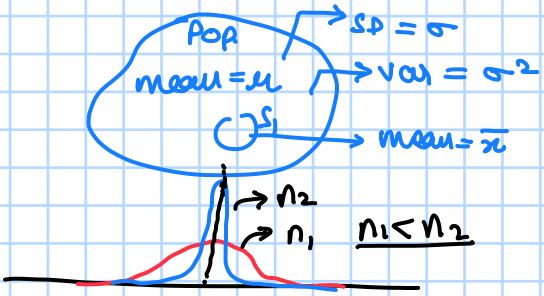
95% confidence interval:



mean b/w 2 values

$S_i = \bar{x}$ mean statistically significant

Central limit theorem:
for a suff. large sample



"dist of sample means → normal dist as n → large"

$\mu_{\bar{x}} = \mu$

$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

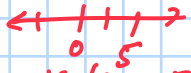
if $n \to \infty$ then straight line

$\beta_1 = 0$ or 5

↓

if 5 then is it different from 0?

Hypothesis statistics



if |(0 − 5)| > (LL − UL)/2

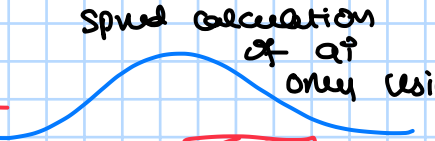
None reason

21st aug :

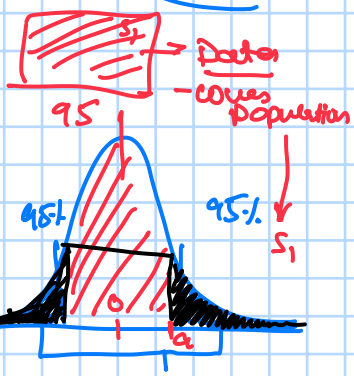
S_1 ← Random representative sample $y = ax + b$ → use → do this using methods discussed, get a, b.

S_i ← we will get different a_i, b_i values everytime.

Spread calculation of a_i only using S_1



We don't want $0 = a_i$ as then no regression ←



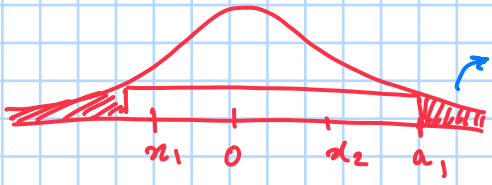
standard deviation / standard errors using a_1, b_1 : $\frac{\sigma}{\sqrt{n}}$

$$SE(a) = \sqrt{\frac{\sigma^2}{\sum (x_i - \bar{x})^2}} \qquad SE(b) = \sqrt{\frac{\sigma^2}{n}\left(1 + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}\right)}$$

(σ from S_1 (close))

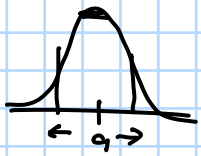
null hypothesis - is a statistically different from ZERO

Pvalue :



Red area is Pvalue

Two tailed vs one-tailed <0.05 is good



F value : multiple linear regression

S_1

$$F = \frac{SSR \uparrow}{\frac{n-k-1}{SSE \downarrow}{n-1}}$$

F statistic / F

n = number of observations

k = no of predictors (ind variables)

more F, better

significance F — P-value associated with F-static (good < 0.05)

AIC (Akaike Information Criteria) and BIC (Bayes Information)

① AIC − BIC = 2 good

↓ is good for AIC and BIC

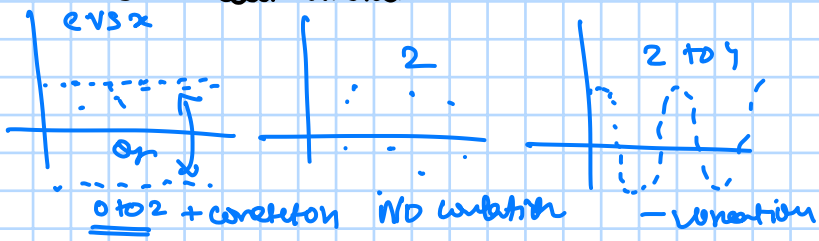
Jarque-Bera test - e vs x (OLS - ord least sq)

low value - <2 } good - follow Dist
with high Pvalue -> 0.05

high value -> 6 } Bad - do not follow Dist
with <0.05

Durbin-Watson test - OLS

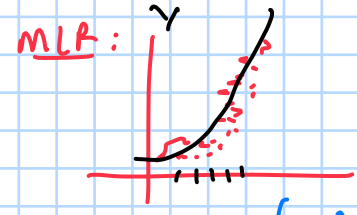
- check autocorrelations

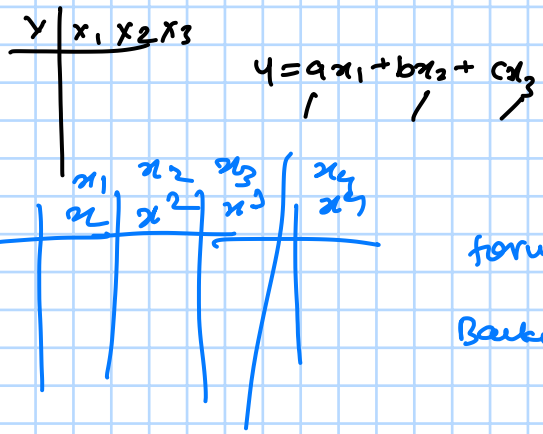


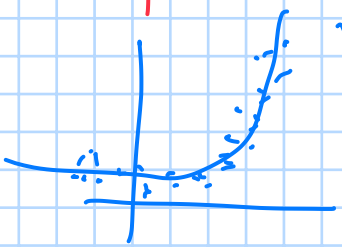
Condition numbers: - OLS
- sensitivity of my model
↓ - coeff not minimal <10
↑ - coeff not maximal >30
>100 very bad

Omnibus: combination of all test (low value)
- P > 0.05

MLR:

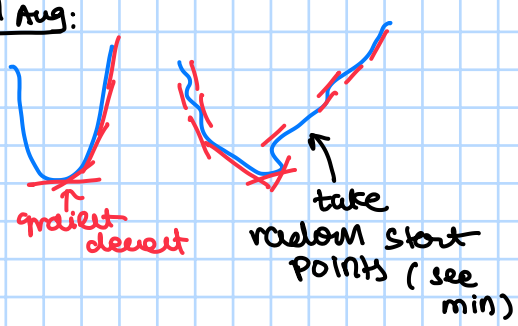


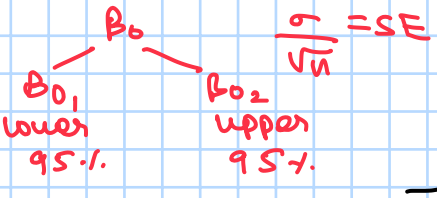


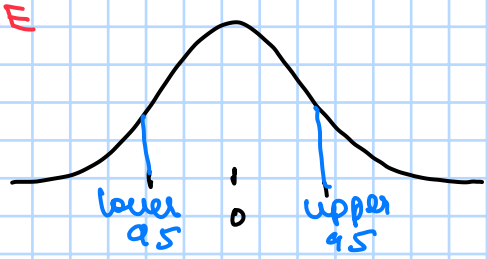


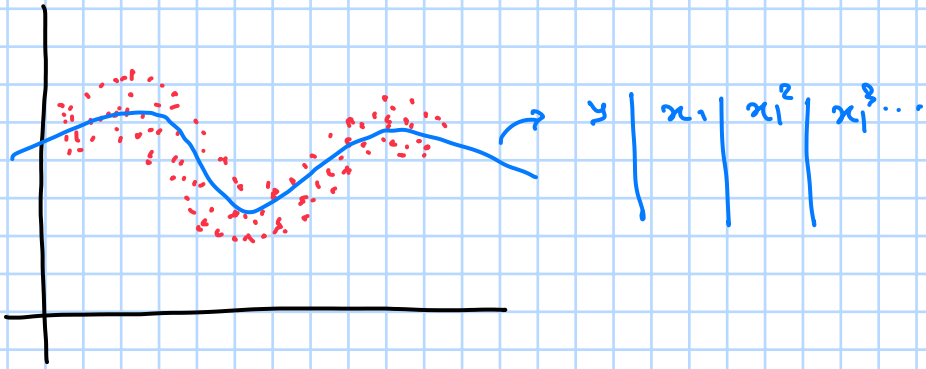
forward feature elim.
VS
Backward featu elim.

23rd Aug:



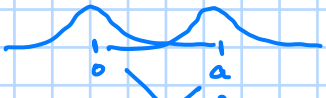






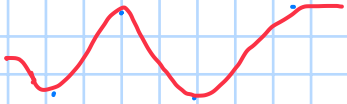
$y \mid x_1 \mid x_1^2 \mid x_1^3 \cdots$

Hypothisis testing
- Null hypothisis



- we do F test to find out.

overfitting:

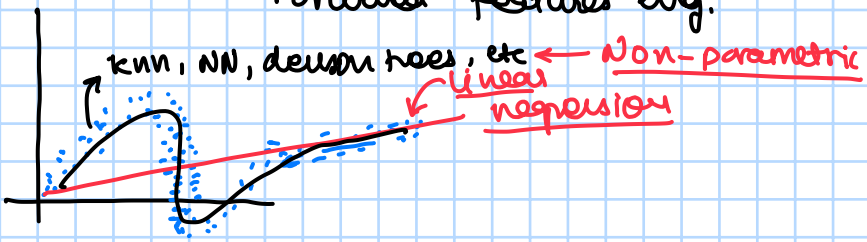


if Errors are drastically different from training data, then overfitting is occuring.

y	x	sinx	cosx	tan(x)
.	.	.	.	.
:	:	:	:	:

← feature engineering

Backward feature eng.
vs
Forward features eng.

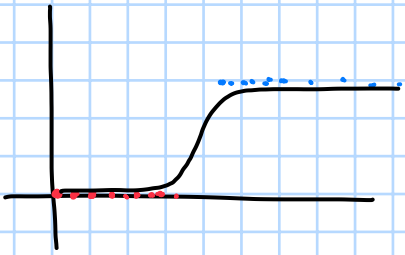


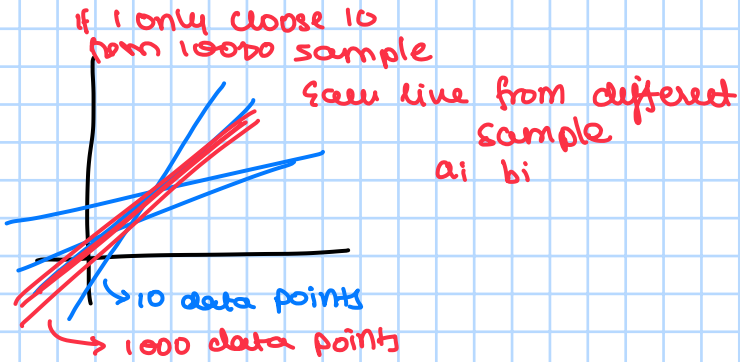
Decision tree / Ensemble method
- combine trees together

Parametric - what if questions
non-parametric - predict value but don't give sensitivity analysis directly.

SKlearn - many models available, see models and make them.
- we have have to decide a good model.
- 80% training 20% testing.
- we are interested in both.

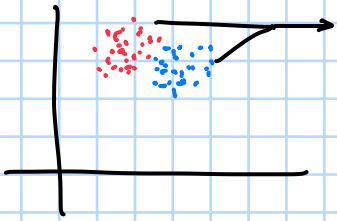
Logistic regression :





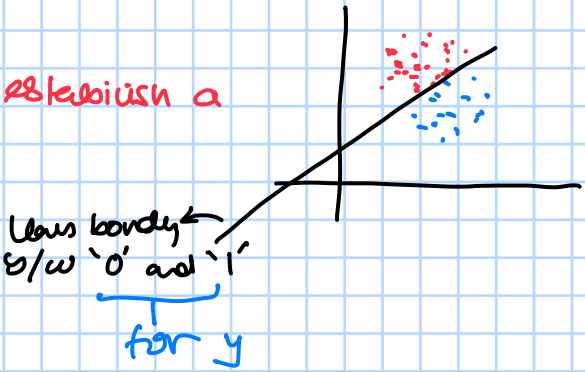
Nominal, ordinal } 'y' Classification → Example, whose face is this?

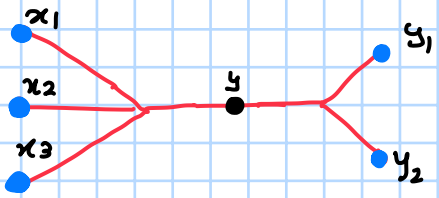
Interval, ratio } 'y' regression → we saw this



Problem: given observations, we want to establish a boundary between both.

$y = f(X)$ ← given x_i's, we want to know if it should be classified 0 or 1.





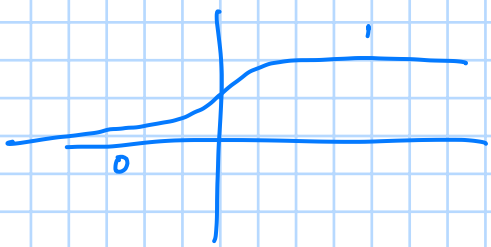
or y in general.

0 state and 1 state

$a = w_1x_1 + w_2x_2 + w_3x_3 + b$ ← bias / w_0

0 or 1 ← sigmoid function

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



$f(a) < {}^{0}_{1}$

→ sigmoid function

$f(a) = \frac{1}{1+\frac{1}{e^a}}$ ← we treat this as probability

$s(a) = \frac{1}{1+e^{-a}}$ < 0; $s(a) \in (0, 0.5)$; 1; $s(a) \in (0.5, 1)$

So we have to find $w_0, w_1, w_2 \ldots$, pass it to sigmoid function and get result.

$f(a) = \dfrac{1}{1+e^{-a}}$

$P(y=1|x) = \dfrac{1}{1+e^{-a}}$ $\quad (0 \le P \le 1)$

x_1	x_2	x_3	t (or y_{act})	y_{pred}
.	.	.	0	1
.	.	.	0	0
.	.	.	1	1

goal: if $t=1$ → we want to maximise P
$t=0$ → we want to maximise $1-P$

t = observed outcome

L = likelyhood $= \prod_{n=1}^{N} P_n^{t_n}(1-P_n)^{(1-t_n)}$

↑ we want to maximise this → we get weights

$P = \dfrac{1}{1+e^{-a}} = \dfrac{1}{1+e^{-(w_1x_1 + w_2x_2 + \cdots)}}$

$$L = \prod_{n=1}^{N} \left(\frac{1}{1+e^{-(w_1x_{1n} + w_2x_{2n} + \cdots)}} \right)^{t_n} (1-P_n)^{(1-t_n)}$$

$$\log(L) = \sum_{n=1}^{N} t_n \log P_n + (1-t_n)\log(1-P_n)$$

↳ log likelyhood, $\max L \Rightarrow \max \log L$
$\max L \Leftarrow \max \log L$

$J = -\log(L)$

→ minimising J

$$\frac{\partial J}{\partial w} = X^T(P-T) \quad \text{or} \quad \frac{\partial J}{\partial w} = (P-T)^T X$$

} gradient of objective function

$w_{new} = w_{curr} - \eta \cdot \dfrac{dJ}{dw}$

learning rate

28th Aug :

$\sigma(a) = \frac{1}{1+e^{-a}}$ ← sigmoid function

helps to arrive at weights, that we make boundary

$\frac{dJ}{dw} = x^T(P-T)$ $\frac{dJ}{dw} = (P-T)^T x$ $y = w^T x + w_0 T + E$

Actual obs (0,1) — predicted probs

η : value
decently small value

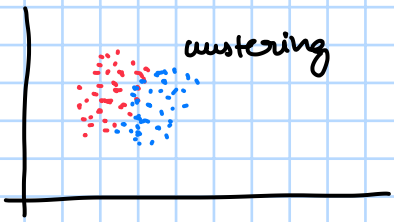
gradient descent

$w_{new} = w_c$ $w_0, w, \ldots$

$w_{new} = w_c - \eta \frac{dJ}{dw}$

learning rate

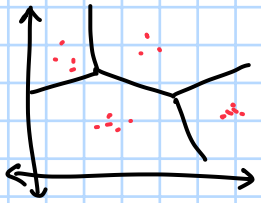
Observations → train model → learn



sklearn linear model ← your logistic reg. model

Errors in logistic logistic regression: ① we can use .score function

② confusion matrix :



	0	1	2	3	support
0	348	9	6	0	360
1	7	342	8	3	360
2	8	6	338	8	360
3	0	6	8	346	360

this diagonal is very important

Note : if we load only 25 instead of 360
Accuracy ↓ This is imbalance

solution:

↓ support	
360 —	360
360 —	360
25 —	200
360 —	360

↑ support

360 —	60
360 —	60
25 —	25
360 —	60
360 —	60

CONFUSION MATRIX:

Actual	1	0
1	TP	FN
0	FP	TN

Belong class 1 and class 1 True positive

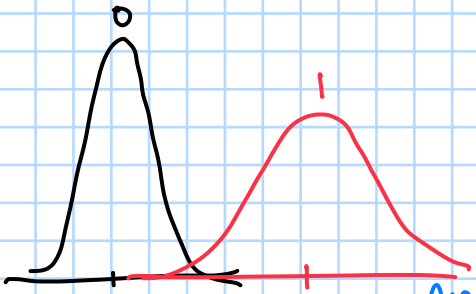
Accuracy = $\frac{TP+TN}{Total}$

Precision = $\frac{TP}{TP+FP}$ Positive prediction value of the classification, how many are correct

Recall = $\frac{TP}{TP+FN}$ Sensitivity / True positive rate of classification how many are correctly recalled.

③ "ROC": Reciever operator characteristic:
curve created for a classifier (and in the content of class)

F_1 score = $\frac{2 \times Prec \times Recall}{Prec + Recall}$ Harmonic mean



True pos. rate TPR $\frac{TP}{TP+FN}$

FPR false positive rate $\frac{FP}{FP+TN}$

Actual 0 - detected as 1 is Fp → not fatal

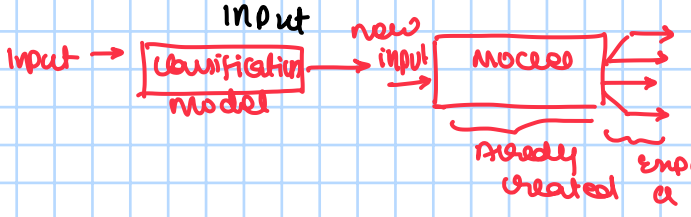
FN → fatal

30th Aug:

Classification:

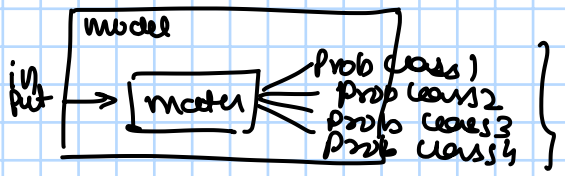
observations (label, inp var_values) → $\hat{y}$ = SOFTMAX (all probabilities)

input

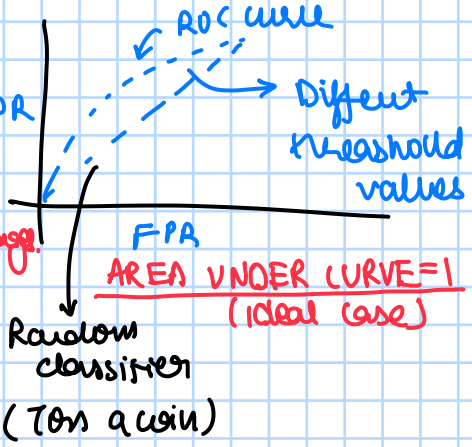


Note: sigmoid - binary classification problem.

Note: softmax function is used for multi-class



$$\text{SOFTMAX} = \frac{e^{pi}}{\sum e^{pi}}$$



When you change threshold value - ROC curve changes.

AREA UNDER CURVE = 1 (ideal case)

Clustering:

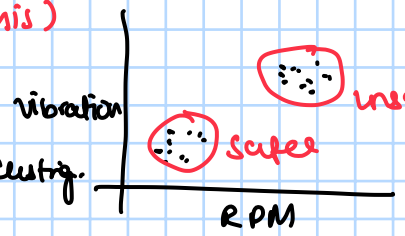
unsupervised learning

goal: discover new aspects of data
mathematically visualise data
subgroups among variables

Note: we can always convert regression problem to classification problem. S, M, L from numbers.

Types:

① PCA (principle component analysis)
② Clustering



Issues: ① subjective process
② noise related to clustg.

Sol

} clustering - part of pre processing as after classifying we will make two classes.

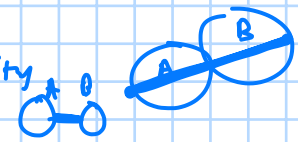
① kmeans - one-specified no. of clusters — 3 if true randomly 3 is selected
② Hierarchical - one ad only one cluster

each observation - own cluster

if 1000 data points → 1 cluster
1000 cluster 1 cluster

Based on distance

or linkage — complete max dissimarity
— single min dissimarity
— Average
— Centroid



given data → problems with y - 18 problems
→ problems with x - 49 problems

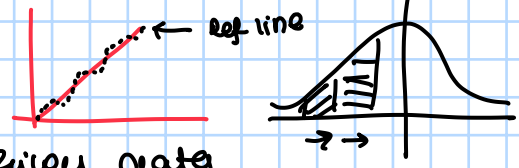
- Not available - clustering
- Incomplete - Noise → manual checking, Automated methods
- insufficient data → few data (collect more, simulate and grate)
- too much data → Statistical sampling, use big data tools, Binning
- imbalance (classification) → over sampling, undersampling, combo technique

x problems → within columns, across columns

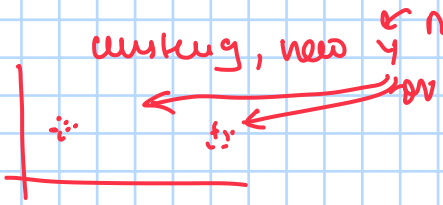
4th sept :

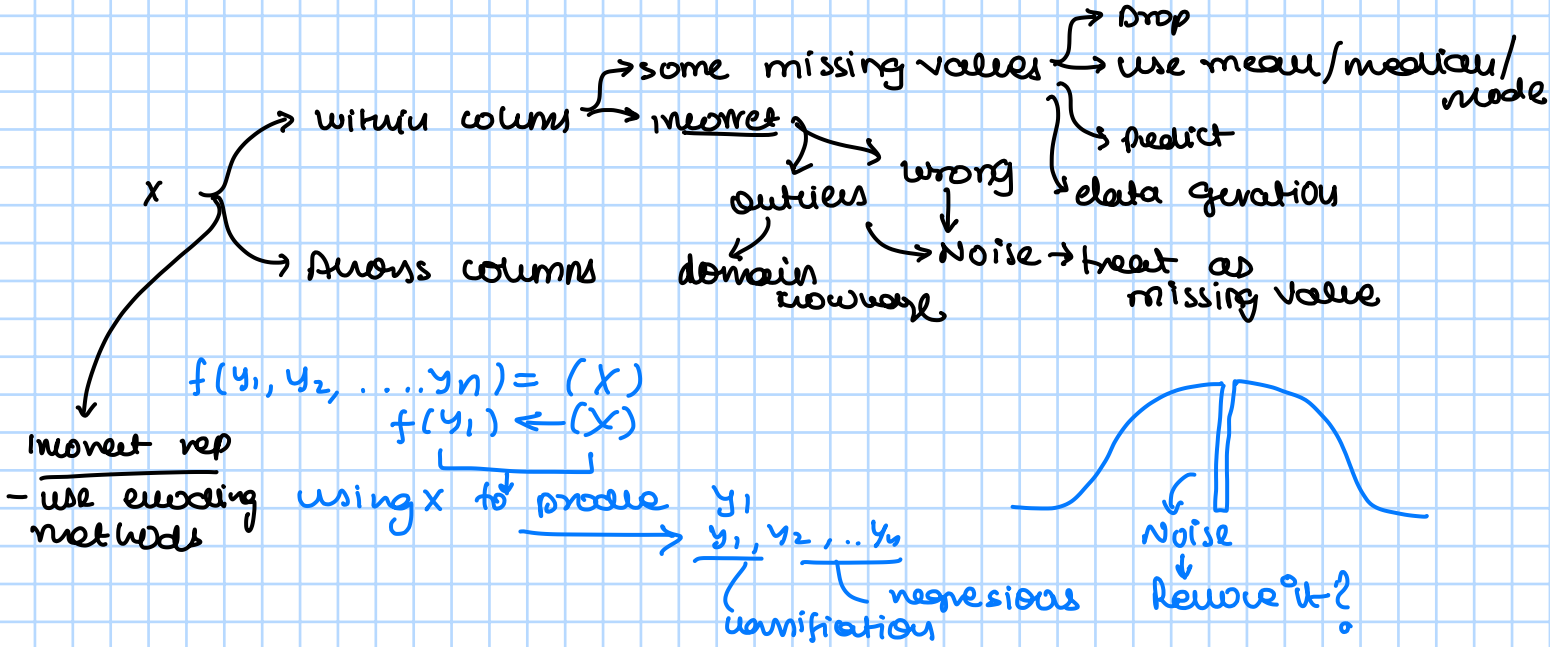
Data analysis
- graphs like ① heat maps, box graph
② descriptive stats

EDA - exploratory data analysis
- gain insights
- spot anomalies
- Q-Q plots → Q-Q plot



given data w.r.t gaussian dist





Across cols
- features not there
- too many features → PCA (Principle component analysis)
- Feature scaling problems
- Correlation / multi-colinearity

6th sept :

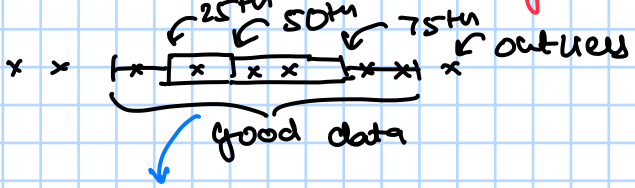
example of data cleaning project
① Birds eye view
② finding bad data
③ trasforming timestamps to require features
④ feature eng
⑤ predict data for bad data

descriptive statistics :

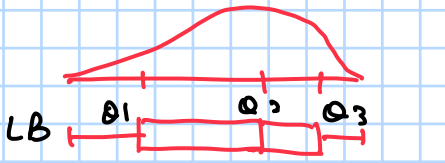
Outliers :

25th percentile, 50th percentile, 75th percentile
median

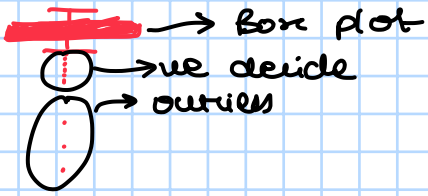
median close to mean is a good thing



50th percentile not shown at centre
is this wrong / right?



I.Q.R = $Q_3 - Q_1 = (V_{75}) - (V_{25})$
inter Quartile range

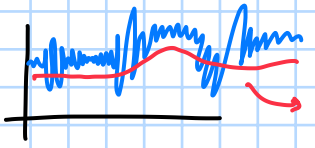


$LB = Q_1 - 1.5 \times (I.Q.R)$
$UB = Q_3 + 1.5 \times (I.Q.R)$

all data b/w LB and UB is our 'good data' (generally, this 1.5 can also be changed)

Trick : we can technically apply box plot to filter data

Scale before algorithm - ① K-nearest neighbors
② SVM
③ Linear regression
④ NN



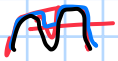
solution
- moving averages or weighted avg

exponential / weighted average

simple average / simple moving average

Strategies for bad data :

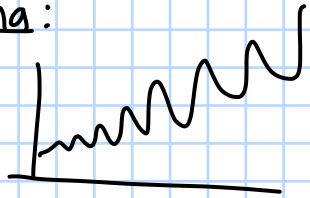
① Winsorization - replaces extreme value with good value (nearest)
② Imputisation - replace with mean / median
③ trimming - removing outliers
④ capping - cap the data
⑤ other predictive methods



CRISP-DM → Industry Standard for data mining

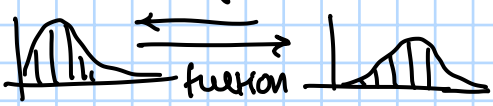
Scaling - some algorithms are sensitive to scaling
- some are not

Arima :

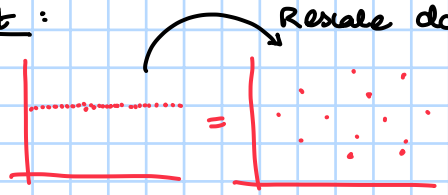


→ multiplicative → log transformation
→ additive → normal / no transformation

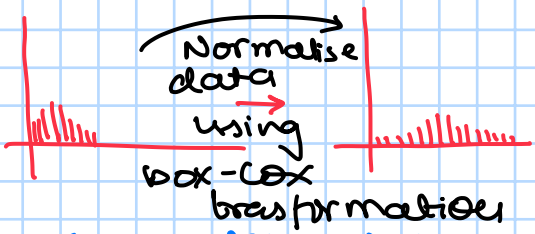
Box-cox transformation - maximum likelyhood estimation



11th sept :

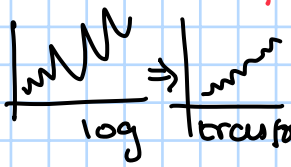


we do this for better clustering



for models which has normal input/ supposes normal input

Heteroscedasticity - we need to eliminate this as many algos have assumed this not present



Data imbalance:

confusion matrics

	0	1	
0	253	5	= 258
1	5	12	= 17

→ highly imbalance data

accuracy = 96.3%

$= \frac{12+253}{\dots}$ ← class 1 has no value

class 1 $\frac{12}{17} \times 100 \approx 70\%$.

} problem → solution → inc data / dec data

SMOTE - Synthetic minority oversampling technique

technique:

① randomly choose ⊗
② least distance point ⊗
③ connect both points and on this line make new ⊗ in between lines.

Note: There is an issue to this, as we make points that are inter-mingling with our data.

ENN - (Noise reduction) - Edited noise-reduction

Algorithm ① k nearest points, we find their class
② take a vote of majority class, eliminate the instance with minority as major class.

Tomek links (Undersampling) -

different classes

- remove the majority b/w each link

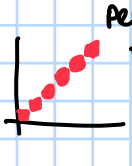
Features / columns - Across the feature problem
- more the number of features, better model → Not true
- overfitting might occur
- As we inc features, too many variables causing too less data

we need to detect direct correlation and multicorelearity in the system

x_3, x_4 $\quad x_1 = f(x_1, x_2, \dots)$

① heat maps
② multicolearity - VIF - variance inflation
③ feature encoding
④ feature eng.
⑤ feature red - PCA

Person co.factor heat map



VIF

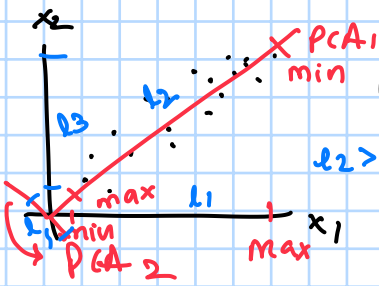
$VIF(x_i) = \frac{1}{1-R_i^2}$

R_i^2 is R^2 by regr with x_i vs other x

⋮ ← we can see what one col

① UNDERSTAND STRUCTURE OF DATA
② MINIMISE FEATURES

} PCA : principle component analysis

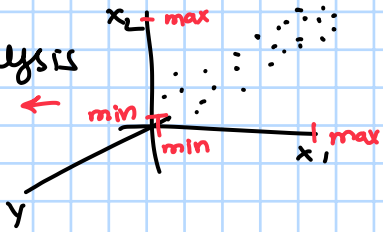


in PCA, we find axis (⊥) as PCA_1, PCA_2

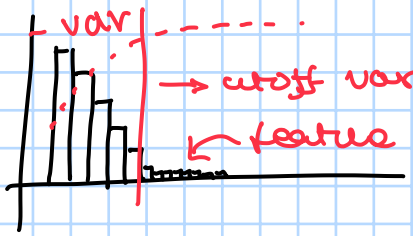
$l_2 > l_1$

$l_3 > l_4$

Are these true ← most perfect axis for maximally capturing variance.



If we plot var by $PC_1, PC_2, \ldots PC_n$ ↑



if 200 columns ← we can ignore this ← too less
↳ we get first PC_n $n = 199$ ∴ 1 dim less

} igen value / vector problem ∴ sales v/s ADS → PC_2, PC_1 } 200 columns not needed

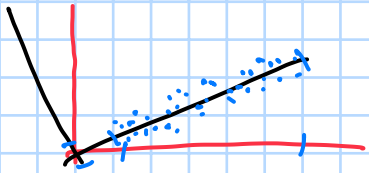
Note - we loose explaination of model.

as $PC_1 = \lambda x_1 + \lambda_2 x_2$

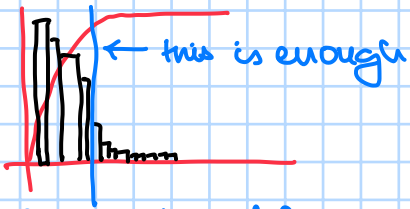
$PC_2 = \lambda_1' x_1 + \lambda_2' x_2$

13th Sept:

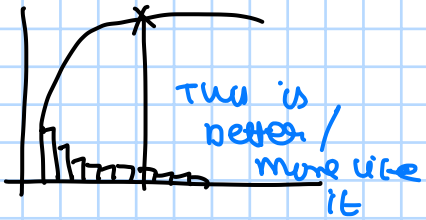
PCA:



loses features, but we loose explainability



if this occurs then data was not normalised, Normalise data



multidim data — 2 to 3 visulise data

MNIST Dataset:

28x28 pixels / Rows

← Number cluster using PCA

New method for 748 dim → 2 dim

t-SNE:

← Number clusters

creates similarity but does not have distance

← if problem has / uses distance we cannot use T-SME

see — Dendrogram

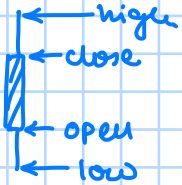
Feature encoding -

- Label encoding
- one-hot encoding
- Binary encoding
- Integer encoding
- frequency encoding
- Target encoding

25th Sep:

1. Feature creation methods
2. Project
3. Assignment — midsem how to solve

feature eng: How to convert text, image, Audio, video, etc.

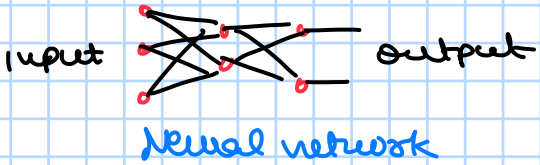


one more example of data

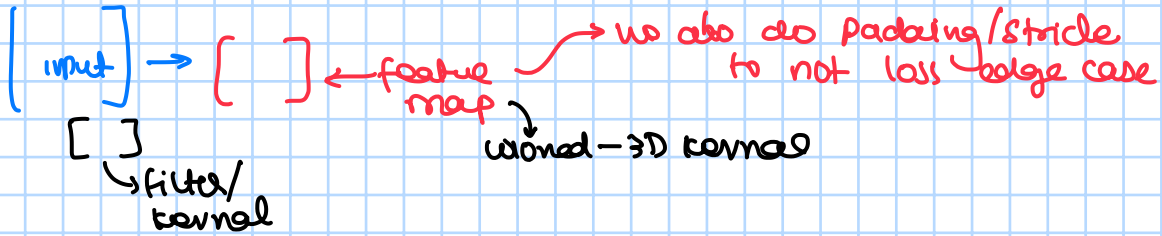
← model/dataset for photos → Best - convolutional neural network

can be split into pixels - and treat as it

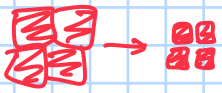
filter: 0 1 0 / 1 0 1 / 0 1 0



Neural network



convolution are further pooled - Pooling

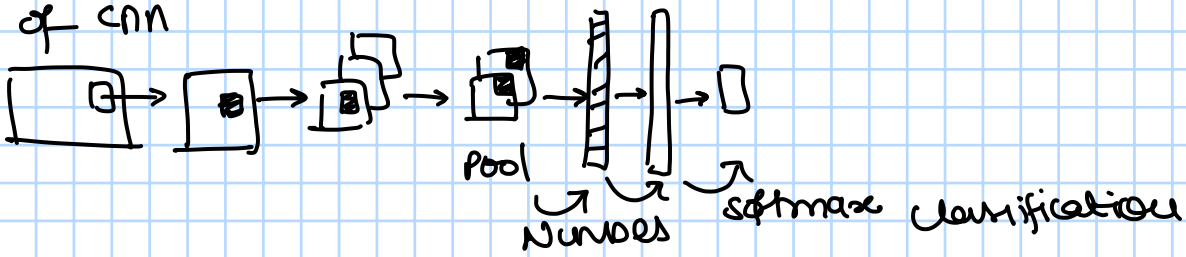


(we do all this to reduce feature / feature eng)

27th sept -

feature creation:

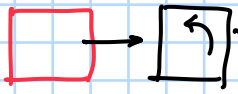
In example of CNN



Example:

cars ① car 1 [1, 0, 0]
② car 2 [0, 1, 0]
③ car 3 [0, 0, 1]

Less training:


rot, flips and all to train CNN

Test processing -

Translation, Auto-corr, Predict next words, text from handwritten note

4th Oct :

vectorisation
'encoding'

Diving — vector weightage (1,0,0...0)

Word2Vec - A type of word embedding which covers words to vectors

- distances / similarity scale etc can be done

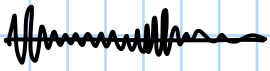
9th Oct:

Audio feature:

↳ Audio is finally a signal

↳ Interested in $Y = f(x)$
so we can convert audio into vectors

↳ Our ear is sensitive to 20Hz ~ 20KHz, we need to have this feature

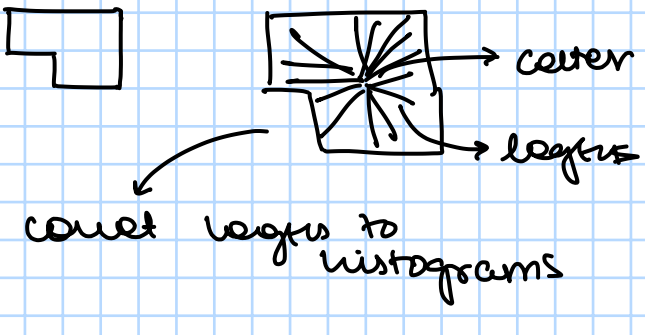
↳ Images also get generated 

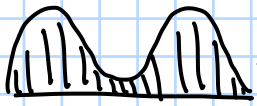
Note: project: Image processing with audio signals

Mel-frequency cepstral coefficients (MFCC)

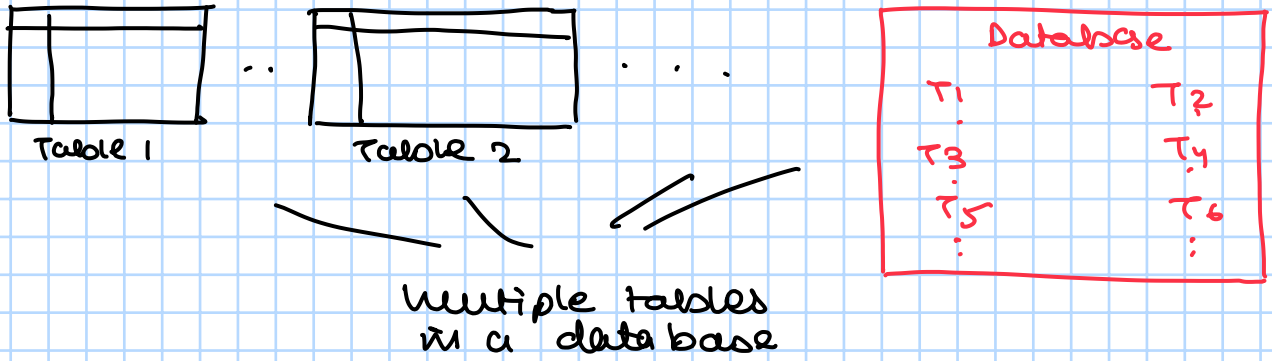
features: ① Mel-f cc
② Spectrogram – 2D – useful for peculiarity of beats of music
– convert Audio to Image – then do image proces
Image overlap
③ Cepstral – power transform of log

Shape in computer:




→ invariant of rotation as histogr

11th oct:



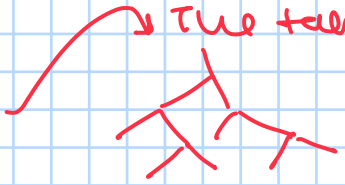
multiple tables in a data base

Example:

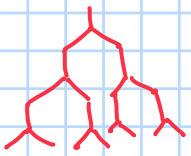
T1: Student

Schema

- ~~Roll no~~ - Alpha → key
- Name
- DOB
- Address
- :
- :

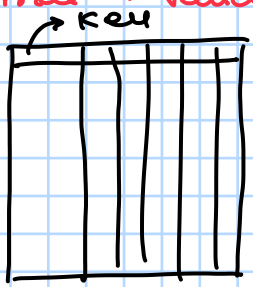


The tables T is one connected to tree in order to search better



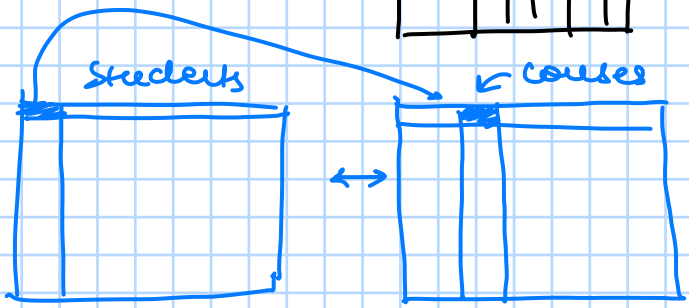
normalised form

- I st normal → every cell - only one data
- II nd normal → particular data format
- III rd normal → validity more than 2nd normal



After 3rd normal
- unique key
- keys don't repeat
- every cell - one data

eg:



given roll no, want a name and year 2024-Autumn all courses

Data Base

R.D.B.M.S - system

Relational Management

Everything at Backend links all of this

Relation

MS SQL / MySQL etc ... RDBMS

Question: How do we define schema for photos? videos? Audio?
(Apriori / Beforehand)

SQL – Structured Query language
↳ Query language supported by RDBMS

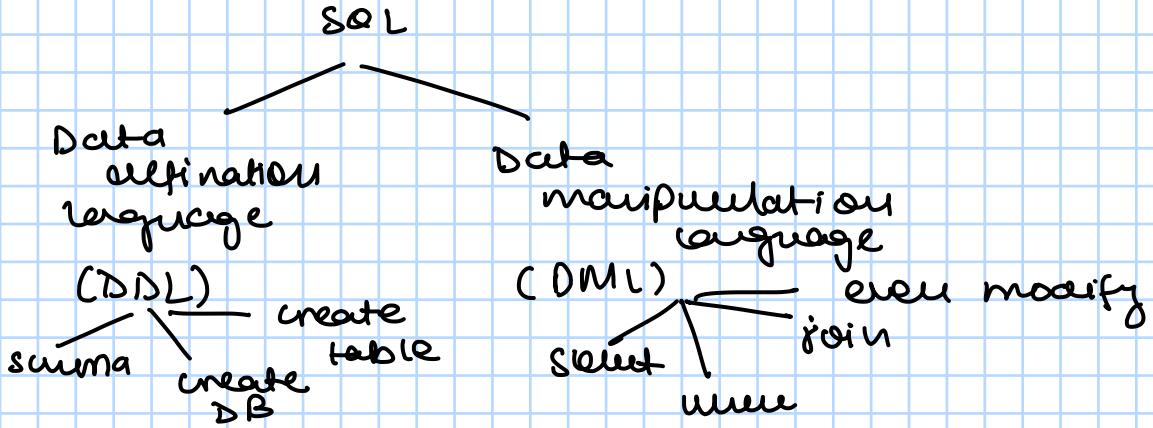
- 5th gen
- Declarative – lower than chat gpt
  but higher than python (3rd gen)
- Describe what to do but there is a syntax

Eg: select RollNo, Name from db.Students where DOB > "2000/01/01"

Syntax: Table is a Set

σ	select	select * from table where CONDITION;
π	projection	select column1, column2 from table;
∪	union	select * from table1 union select * from table2;
∩	intersect	select * from table1 intersect select * from table2;
−	diff	.
×	Cartesian product	.
⋈	join	.

Example: (TSTOCK × TSTOCK) ⋈ (close WIPRO & close TATA)



16th oct :

Project : ① audio files (songs)

② unsupervised data

③ more project - feature engg & feature production

Songs - clusters

Note : what is MFCC - understand that first

Azure : do yourself

Big data : what a big ?

- velocity
- volume
- variety

Amount of moving through data pipes

structured
schema, tables, R.D.M.S

unstructured
text
video
Audio
Images

>99.999 %.
very important/very useful

How to store/process it :

→ How to store?
→ How to process?
→ what kind of info can be extracted

18th oct :

Login to VM ware
tuesday
① db inside VM
② LSE data table in stock

Big data :
- size (relative)

Stuctured V/S Unstructured
Note : semi-structured data is data-schema

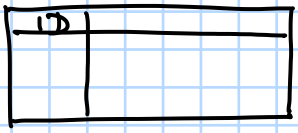
Unstructured - Image
Audio
Video
Text

Picture data — in RDBMS

Storing large data
- How to process it
- where
- How

23rd oct :

R.D.B.M.S

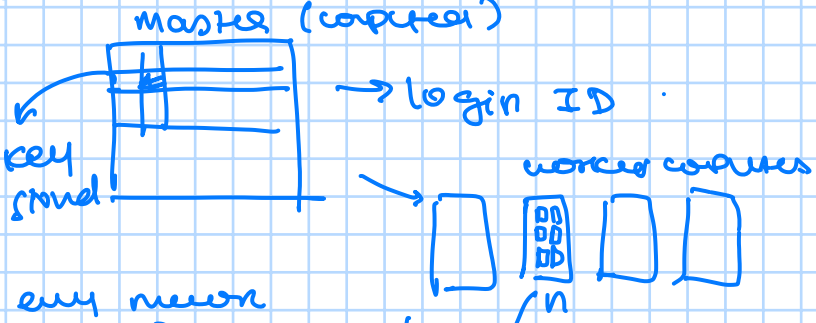


key, schema

Database → key + value

Database is collection of key value pairs

Example : facebook → <K, V>
login key / value place



every worker
① store key/values
② pointers to some computer say n

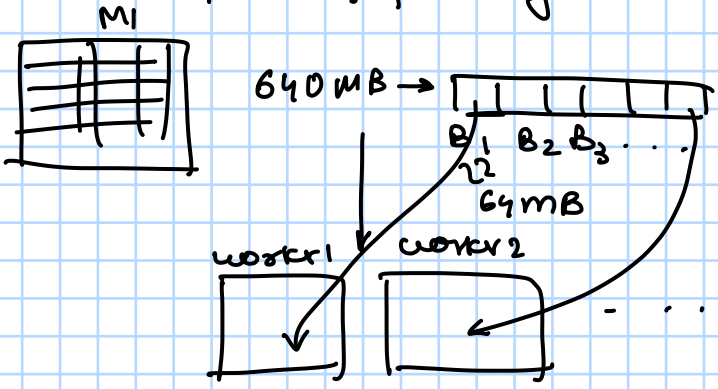
All my ① Images
② videos
③ comments
④ etc

Queries example ① How many videos got uploaded

Note - we don't need database, we need to use RAM

Then my worker computers gets answer and sends it to master

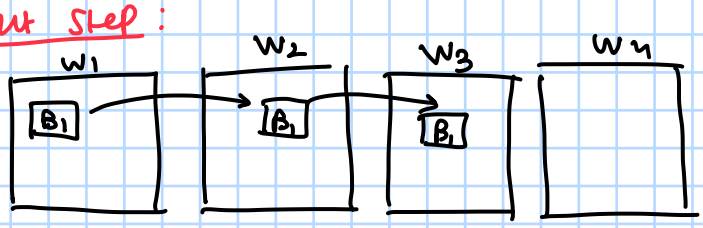
HADOOP - Parallel process for big data



It stores blocks into worker computers

64 Mb - As data is read for 64 mb data chunk

Next step :

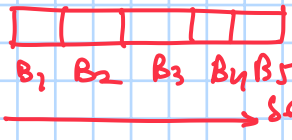


- min 3 blocks (same) stored for safety
- communication within RACKS is much higher than across racks

∴ 1 file ⇒ files and stored ← information available in master

what if master crashes → so there is one more master whose job is to copy master 1

Outages : Recovery takes place



Parts of file : instead of sequential we can also run them in parallel as every chunk on different computers

HADOOP is a cluster (of m+m+ w_1 + ...)

High Availability (HA)

Depends on how much data is needed

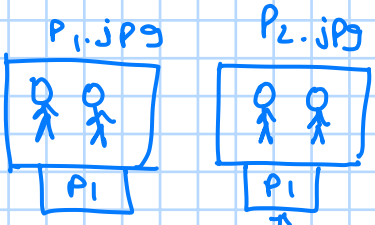
Processing :

programme P_1 → key/value pair → master

now P_1/P_2 programme gives/takes Queries and now can be solved to stored data or stored in RDMS.

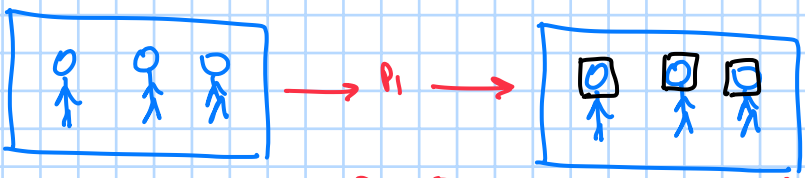
(AKA map)

ex :



Answer ① How many people
② who are in the photograph

image process programme (map)



Here P_1 ① identifies no of faces (Image processing)
② identifies who

map & reduce - Teaches you to make things brick by brick
- not traditional way
- we will be able to process almost all kind of data

project:

① every file has 20 rows { varying cols depending on length of songs

② creating right kind of features to solve the problem

③ Note: No of cols not uniform but cols per sec = 86
as $\frac{44100}{512}$ → $\frac{44100}{512}$ samples per sec
samples per col

map / Questions ① identify song
② identify who sung it

Questions to solve ① How to have same number of cols
↓
solve by clustering (Think of features)
② seems to be 7 group of songs / Mixt

Heatmap → Rearrange it to see clusters together

for first problem need feature set ← same number of features

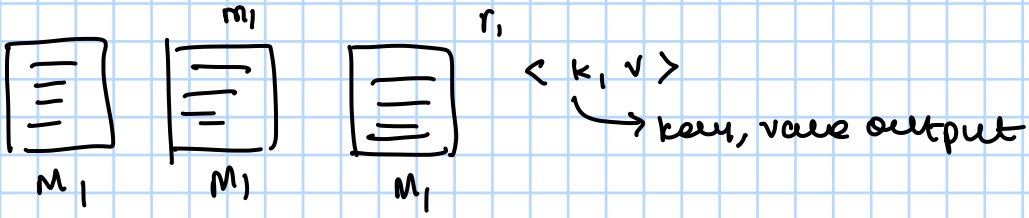
③ identify songs sung by P_1, P_2, P_3 ← so now we have categories

for this take your own songs → make a training set / train data

∴ we need training data → make MFCC coefficients
↓
Convert into relevant set of features
↓
use classification as a tool to identify which cluster does what

25th oct :

Map reduce - Input data → Map → SORT → REDUCE



use for SQL Queries — we can have key/value using map/reduce

- Helps us deal with big data
- Should/would/text/Audio/Video all doable

∴ does not matter if data is structure or not-

Input data (100 TB) → ALIVE

SPARK

Process Any kind / Any size

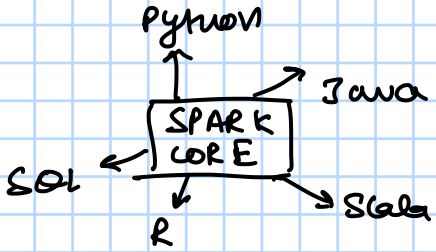
30th oct :

map reduce — strategy - start from output
- input to reduce is what

Spark - SQL
- ML
- graph processing

Speed of Spark - in memory memory
↓
Cloud computing - cluster

lazy evaluation - data evaluated only when required
- trade off - error can't get during analysis



1st Nov :

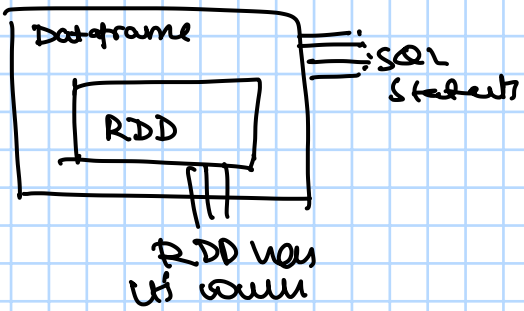
Sparks
- ↳ 01
- ↳ 02
- ↳ 03
- ↳ 04
- ↳ 05

} Notebooks to practice ← Moodle

Python list → RDD
↑
special variable in spark

6th NOV :

Part-4-Spark, Load the CSV TO Dataframe



Extract RDD to df :

```
df=dd.map(lambda x:1).reduce(lambda x,y:x+y)
```

we can create a temp view

```
df =CreateTempView("stocks")
sqlres = SS.Sql ("~~~")
sql res.show()
```

we can also define the schema using spark

NOR - 5th ipynb ← custom schema

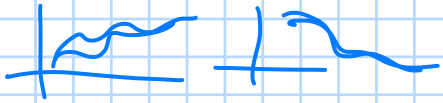
6th ipynb : linear reg with spark

```
df1.corr("open","close")
```

Note : we can do linear reg in SPARK

convert to python list

use matplotlib



8th NOV :

Recap

① Some basic
then done