Develop algorithm field of study machine learning. → actionable knowledge

Imp
1. Data storage - HDD, flash m/y
2. Abstraction
   ↳ Data cleaning
   ↳ Feature extraction → 80% training data, 20% test data
   ↳ standardization
3. Generalization → future prediction
4. Evaluation

} different steps in ML

1. Data is very important which is stored in anywhere.
2. Abstraction give the meaning to the data.
   Data cleaning → remove the missing value.
   Feature extraction → eg: Bank loan gender not impt
3. Generalization → transform the data, training the data into a model. unlike models.
4. Evaluation → test the unseen data, the model is performing well or not.

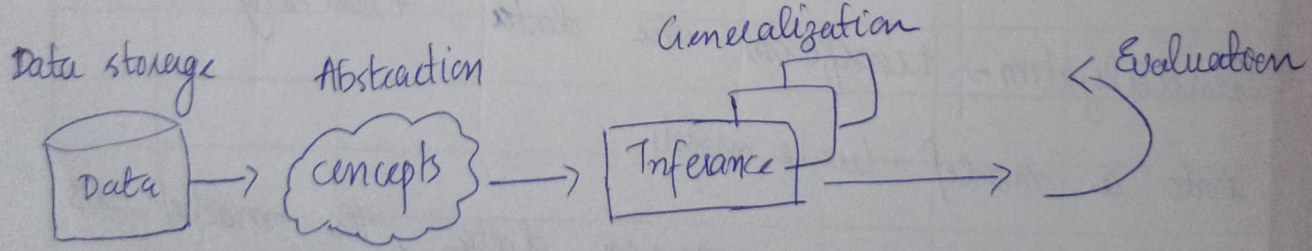
1. Overfitting
2. Underfitting

1. → • The model perform well on the training data but the model perform poor on the test data.
   • Model fails to analyze the underline pattern of data.
   • Model ~~does not~~ byheart the data

		sqfeet	rooms		
eg:	city	1500	3	2	4,0000
	city	1480	3	2	1,0000 x

2. The model perform both poorly on training and tut data

Good fit - Model perform both well on training & test data



Traditional

Data →
program →
computer → output

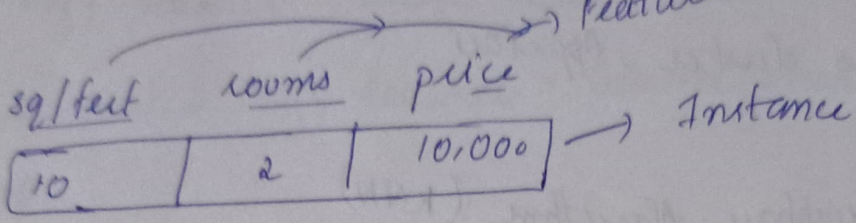
VS

Machine learning

Data →
output →
computer → program

Input Data

Dataset - is a collection of unit of observation - which is different in different context



Features can be different types: } different type of data

① Numeric eg: height, weight

② categorical / Nominal

eg: male/female - gender

③ Ordinal ~~eg:~~ education qualification - SSLC, 12th, UG, PG

↙ variant of categorical eg: size of shirt

There are 2 type ML Model:

Predictive & Descriptive:

→ supervised → classification

predictive: It is used for future predictions and the data consist of features and target variables (class label)

Descriptive: No feature is superior over other, which does not have a target variable.

(unsupervised, clustering

predictive

↳ SVM, KNN, Decision tree → classification different algo

Descriptive

↳ kmeans cluster, Apriori

K-Nearest Neighbour Algorithm (KNN)

- supervised classification algorithm
- imp concept is to find the similarity of neighbor data

eg: Tomato : Sweetness = 6
Crunchiness = 4

Ingredient	Sweetness	Crunchiness	Food Type	Distance to the tomato
Grape	8	5	Fruit	$\sqrt{(6-8)^2+(4-5)^2} = \sqrt{5}$
Greenbean	3	7	vegetable	$\sqrt{(6-3)^2+(4-7)^2} = \sqrt{18}$
Nuts	3	6	protein	$\sqrt{(6-3)^2+(4-6)^2} = \sqrt{13}$
Orange	7	3	Fruit	$\sqrt{(6-7)^2+(4-3)^2} = \sqrt{2}$

→ Identify distance measurement

→ k value should not too much low & too much high

→ k=1, choose nearest value so $\sqrt{2}$, so we predict the tomato is a fruit.

→ k=3, majority class has priority

→ k value should be optimal, we should select the no. of neighbors

KNN :- distance $(p_1, q_1) = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$ evaluation

→ KNN - lazy learner, when new data occur it only performs so it is called lazy learner. Generalization stage does not in KNN.

It only store the data, when a data occur/ comes it calculate and analyse.

$$dest(p, q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \dots (p_n - q_n)^2}$$

$k = \sqrt{4} = 2$, so we consider $\sqrt{2}$ fruit and $\sqrt{5}$ fruit. (majority) so it predict that food type is fruit.

- KNN Algorithm is a simple & effective algorithm that makes no assumption about the underlined data distribution. The letter k is a variable term implied nearest number of neighbors should be used.

In this example locating tomatoes nearest neighbors requires a nearest functions, that measures the similarity b/w 2 instances. Traditionally KNN algorithm uses 'euclidean distance' for measuring the similarity

- simple and effective classification algorithm
- New data arrives at testing phase so that learning phase is fast, learning phase >>> testing phase.
- No generalization and abstraction
- No model creation
- When missing values are there additional preprocessing is needed
- Make no assumptions

✓ k = No of neighbour to classify to unlabeled data

✓ ~~Education~~ Euclidean distance f → (most used for measurement)

Manhatten & Minkowski → (less used)

✓ lazy learning & instant learning & Rote learning → KNN

Q. Based on a survey conducted in an institution students are classified Based on the 2 attributes: academics excellence and other activities. Given the following data identify the classification of a student with x = 5 and y = 7 using KNN algorithm. (choose k as 3)

x (Academic Excellence)	y (other Activities)	z (classification)
8	6	outstanding
5	6	good
7	3	good
6	9	outstanding

Ans: distance measurement:

$\sqrt{(5-8)^2+(7-6)^2} = \sqrt{9+1} = \sqrt{10}$ outstanding

$\sqrt{(5-5)^2+(7-6)^2} = \sqrt{0+12} = \sqrt{1}$ good

$\sqrt{(5-7)^2+(7-3)^2} = \sqrt{9+16} = \sqrt{20}$ good

$\sqrt{(5-6)^2+(7-9)^2} = \sqrt{1+9} = \sqrt{5}$ outstanding.

k = 3 so that the neaest neighbor $\sqrt{10}$, $\sqrt{1}$ and $\sqrt{5}$ outstanding good outstanding ie, outstanding

Probilistic Learning

Naive Bayes classification

eg: probability that today might be a rainy day (70% chance) 10 days taken with the same and similar condition of today. then 7 days are rainy and concluded that 7/10 ie, 70%

application:

eg: weather forcasting, Text classification (spam / Not spam)

eg: probability that the incoming msg is spam or ham

: Anomaly detection in a network

: diagnosis of a disease, based on symptoms

- deals with probability of likelihood of an event.
  - Based on many diff terms (likelihood of an event)

eg: coin tossing → trial / event → head

trial: Incoming msg , event → spam

- 'event' and 'trail' main dependency of likelihood
- All the features have importance.

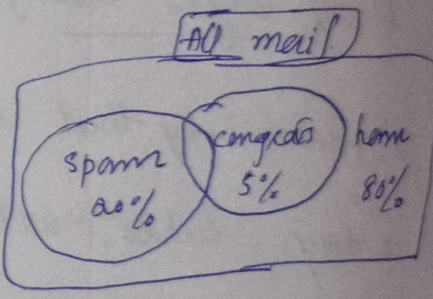
  Even if the feature posses small role even though less impact they also considered

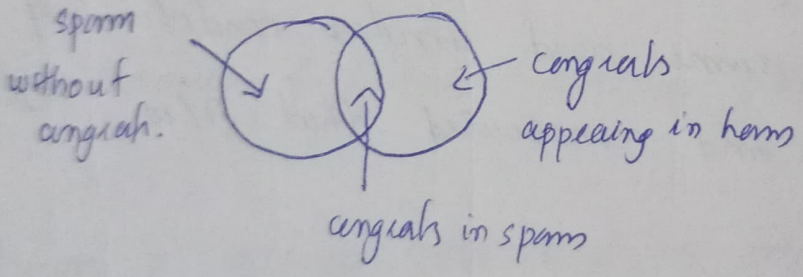
- Mutually exclusive & exhaustive event

  ie, Both events does not come together → Either spam/ham

example: ~~eg~~ spam mail - 20% ham mail → 80%

ie, 100 - 80 → 20%





- probability that msg is spam and msg contain congrats [dependent event]

"dependent event" in (probabilistic learning) considered as the basis for predictive modeling.

- Independent event.

- probability (spam ∩ congratulations)

independent event ↓

$P(\text{spam}) \times P(\text{congratulations})$

$.20 \times .05$

$= 0.01 \rightarrow 1\%$

spam msg provided congratulations occur → dependent event.

Bayes,

✓ $P(A|B) = \frac{P(A \cap B)}{P(B)}$

↓

conditional probability → A mail which is spam and which included congrats msg.

↙

one type

↳ posterior probability : A new instance occurs how the evaluation occurs.

✓ $P(A \cap B) = P(A|B) \times P(B)$

$P(A \cap \bar{B}) = P(B \cap \bar{A})$

$P(B|A) = \frac{P(B \cap A)}{P(A)}$

$P(B \cap A) = P(B|A) \times P(A)$

ie,

$P(A|B) = \frac{P(B|A) \times P(A)}{P(B)}$

prior probability → eg: In spam msg there is already congrats msg.

$$P(\text{spam} \mid \text{congrats}) = \frac{P(\text{congrats} \mid \text{spam})\, P(\text{spam})}{P(\text{congrats})}$$

likelihood ↓ ← prior probability

← marginal likelihood.

posterior probability →

Naive Bayes problem.

supervised

chills	Running Nose	Headache	Fever	Has Flu
Y	N	mild	Y	N
Y	Y	No	N	Y
Y	N	strong	Y	Y
N	Y	mild	Y	Y
N	N	No	N	N
N	Y	strong	Y	Y
N	Y	strong	N	N
Y	Y	mild	Y	Y

chills	Fever	Headache	Running Nose	Has Flu
Y	Y	mild	No	?

↗ posterior probability

solution:

Has Flu Yes = 5/8 = 0.625 ← prior probability

Has Flu No = 3/8 = 0.375

$P(\text{chills} = Y \mid \text{Flu} = Y) = \frac{3}{5} = 0.6$

$P(\text{Fever} = Y \mid \text{Flu} = Y) = \frac{4}{5} = 0.8$

$P(\text{Headache} = \text{mild} \mid \text{Flu} = Y) = \frac{2}{5} = 0.4$

$P(\text{Running Nose} = \text{No} \mid \text{Flu} = Y) = 1/5 = 0.2$

$\therefore (0.6 \times 0.8 \times 0.4 \times 0.2) \times 0.625$

$= 0.0384 \times 0.625$

$= 0.024$

conditional probability

$P(\text{chills} = Y \mid \text{Flu} = N) = 1/3 = 0.33$

$P(\text{Fever} = Y \mid \text{Flu} = N) = 1/3 = 0.33$

$P(\text{Headache} = \text{mild} \mid \text{Flu} = N) = 1/3 = 0.33$

$P(\text{Running Nose} = \text{No} \mid \text{Flu} = N) = 2/3 = 0.66$

$\therefore (0.33 \times 0.33 \times 0.33 \times 0.66) \times 0.375$

$= 0.02371 \times 0.375 = 0.009$

0.33 ×
0.33
0.99
0.990
0.990
0.11989

0.02371 ×
0.375

so here 0.024 > 0.009 ie, has Flu = Y

problem

- Estimate conditional probabilities of each attributes {color, legs, height, smelly} for the species classes {M, H} using the data given in the table
- Using these probabilities estimate the probability value for the new instance {color = green, legs = 2, Height = Tall and Smelly = No}

No	Color	legs	Height	Smelly	Species
1	white	3	short	yes	M
2	green	2	Tall	No	M
3	green	3	short	yes	M
4	white	3	short	yes	M
5	Green	2	short	No	H
6	white	2	Tall	No	H
7	white	2	Tall	No	H
8	white	2	short	yes	H
9					

solution

color	legs	Height	Smelly	Species
green	2	Tall	No	?

Has species M = 4/8 = .5

Has species H = 4/8 = .5

$P(M=Y \mid color = Green) = 2/4 = .5$
$P(M=Y \mid legs = 2) = 1/4 = .25$
$P(M=Y \mid Height = Tall) = 1/4 = .25$
$P(M=Y \mid smelly = No) = 1/4 = .25$

$\therefore (.5 \times .25 \times .25 \times .25) \times .5$
$= 0.0078125 \times .5$
$= 0.00390$

$P(H=Y \mid color = Green) = 1/4 = .25$
$P(H=Y \mid legs = 2) = 4/4 = 1$
$P(H=Y \mid Height = Tall) = 2/4 = .5$
$P(H=Y \mid smelly = No) = 3/4 = .75$

$\therefore (.25 \times 1 \times .5 \times .75) \times .5$
$= 0.09375 \times 0.5$
$= 0.046875$

ie, $0.046875 > 0.00390$ $\therefore$ species = H

Name Byes Algorithms gives all the features has important.

Laplace Estimator / Laplace Smoothing

Fruit	Yellow	Sweet	Long	Total
Mango	350	450	0	650
Banana	400	300	350	400
Others	50	850	400	150
Total	800	850	400	1200

Laplace estimator also known as Laplace smoothing is a technique used in probability & statistics to handle the problem of zero

probability in categorical data. It involves adding a small constant value (often 1) to each count of observed icons which ensures that no probability is ever zero. Even if some events have not observed in examples. Laplace estimator ensures that all possible events have a non zero probability making the model more robust especially when dealing with more dataset.

$$P(\text{mango}) = \frac{650}{1200} = \cancel{0.333}\ 0.542$$

$$P(\text{Banana}) = \frac{400}{1200} = 0.333$$

$$P(\text{other}) = \frac{150}{1200} = 0.625$$

$P(\text{fruit} / \text{yellow, sweet, long})$

$$P(\text{yellow} / \text{mango}) = \frac{3050}{650} =$$

$$P(\text{sweet} / \text{mango}) = \frac{450}{650}$$

$$P(\text{long} / \text{mango}) = \frac{0}{650}$$

when we multiple it become zero

So avoid this,

$$\frac{0}{650} = \frac{0+1}{650+k} \qquad k = 2$$

↓

Binary category.

$\therefore$ p(yellow | mango) = $\frac{350}{650}$ = $\frac{350+1}{650+2}$ = $\frac{351}{652}$

= ~~0.000050~~ 0.539

$\therefore$ p(sweet | mango) = $\frac{451}{652}$ = ~~0.539~~ 0.692

$\therefore$ p(long | mango) = $\frac{1}{652}$ = 0.002

p(yellow | banana) = $\frac{400}{400}$ = $\frac{401}{402}$ = 0.998

p(sweet | banana) = $\frac{301}{402}$ = 0.749

p(long | banana) = $\frac{351}{402}$ = 0.873

$\rightarrow$ p(mango | yellow, sweet, long)

= p(mango) = p(yellow | mango) $\times$
p(sweet | mango) $\times$
p(long | mango)

= ~~0.542 $\times$ 0.00050 $\times$ 0.531 $\times$ 0.002 = 0.000~~

0.542 $\times$ 0.002 $\times$ 0.692 $\times$ 0.539 = 0.000050

p(Banana) = 0.333 $\times$ 0.998 $\times$ 0.749 $\times$ 0.873

=

p(other) =

p(yellow | other) = ~~401~~ $\frac{51}{152}$ = 0.335

p(sweet | other) = $\frac{851}{152}$ = 5.598

p(long | other) = $\frac{401}{152}$ = 2.657

$p(other) = 0.125 \times 0.335 \times 5.595 \times 2.657$

knn naive bayes svm ML

(2)

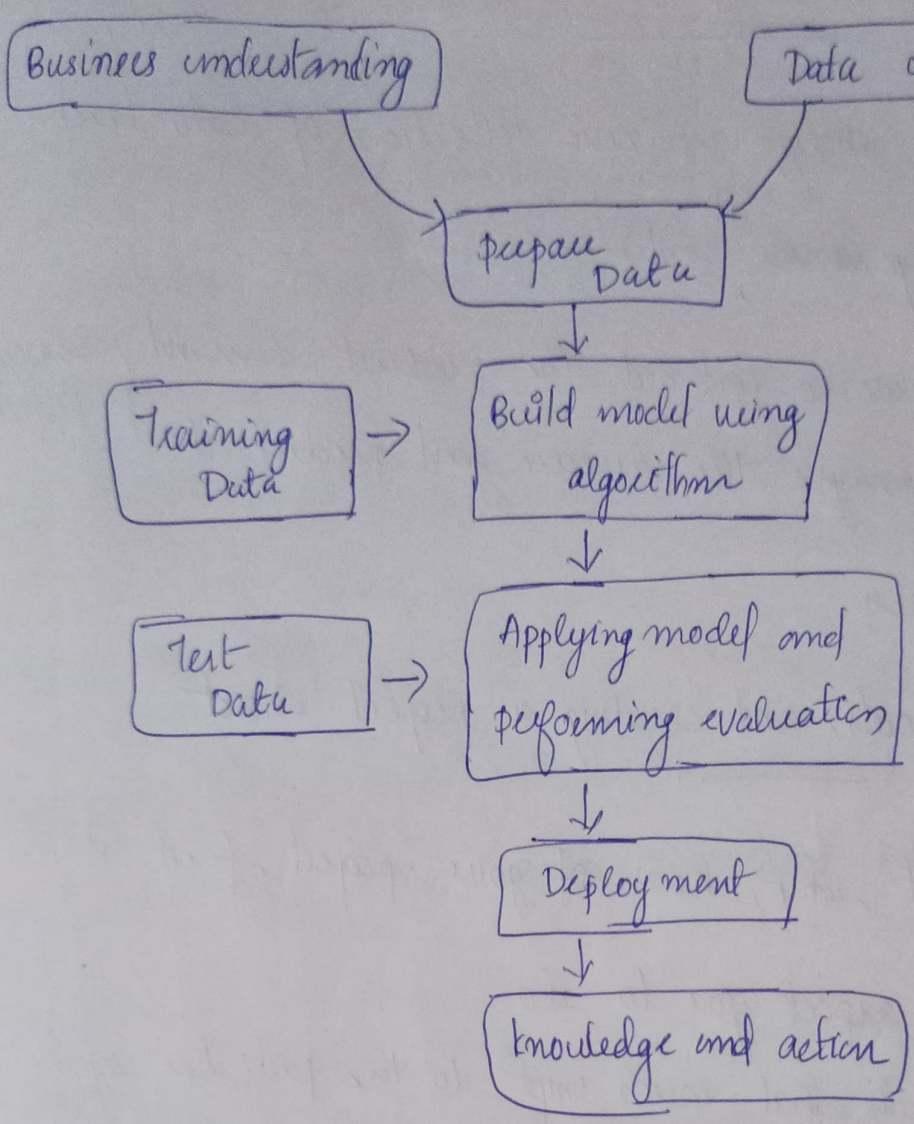
Data Science Process

It is a collection of technique used to extract values from data.

- We can find patterns from data
- " " " connection within data } through data science
- " " " Relations within data

(The method is based on evidence/empirical knowledge/ past historical ~~knowledge~~ observations) ⇒ science

Data science is commonly refered as knowledge discovery, data mining, predictive analytics, machine learning



1. prior knowledge
2. preparation
3. Modeling
4. Application
5. knowledge.

1. prior knowledge

- the inf which is already known about subject

  10,0000 → 5000 → difficulties

  ↓ income, credict history

  eg: predict intrest rate for a person
- The inf about data.
- prior knowledge about the subject area and the data.

Data Science and Data Science Process

It involves methods to analyse massive collection of data and extract knowledge it contains.

eg: In a retail chain we try to find out the product demand } Data science
: In hospital doctor analyse the disease and predict it } Project

Data Science process

1. Defining research goals, and creating a project chart

- understanding what, why, how of your project, first step.

→ what the company expect you to do
→ why the scenario is that much impt to the particular project
→ how it is going to be / Is it a biggest strategic picture or a lone wolf (single individual idea)

eg: Research defaulters

By following the ~~final~~ three steps we come up with a well defined research goal, well defined deliverables and action plans → we can achieve.

Creating a project chart: → which helps mostly the clients.

① A clear research goal
② The project mission and context

It is a schedule or document which helps the client to give more clarity

(3) How you are going to perform your analysis

(4) What resources you expect to use

(5) proof that is an achievable project or a proof of concepts

(6) Deliverables and measures of success

(7) Timeline

First step is time consuming.

2. Retrieving Data

- Start with data stored within the company. If it is not enough they collect the data from the third party.

Data Sources
- → Database - (storage)
- → Data marts - (subset of data warehouse)
- → Data house - (to read and analysing the data)
- → Data lakes - (natural data / raw data storage)

If the data is not enough

(Data is used for companies)

- Don't be afraid to shoparound

Data is considered as precious, there is regulations & policies to retrieve the data.

- → Data as an asset
- → Do quality checks
- → Investigate Data

3 phases
- Data retrieve (datatype is ok, data is needed)
- Data preparation (avoid null, duplicate data, errorcheck, missing values)
- data exploration - Statistical properties (Bar chart, histogram, mean, median etc)

3. Data Preparation - Cleansing, Integrating & Transforming data.

Cleansing

- we need to sanitize (clean) the data.

→ Data should be sanitized, so that model performs well on the data.

- Interpretation error (eg: febuary 30 → that is logically incorrect,
  : human have 360 age)

- Inconsistency (eg: Take 2 column which represent female, in one column it write female other column write F)

→ Data should be represented in a consistent way and should be interpreted correctly. There occurs 2 types of errors

a. Interpretation error:

In this case the data is represented in such a way that it is not logicaly correct. eg: age > 350, febuary 30
(this kind of data should not happen)

b. Inconsistency error:

The data is not represented in a consistent manner eg: representing 'female' in one column and 'F' in another column.

Error Description	possible solutions
• mistakes during data entry	• Manual overrules
• Redundant whitespace	• use string functions (strip(), lstrip(), rstrip())
• Impossible values	• Manual overrules
• Outlayers ↓	• validate and if errors, treat missing values

(An observation that seems to be distand from other observations)

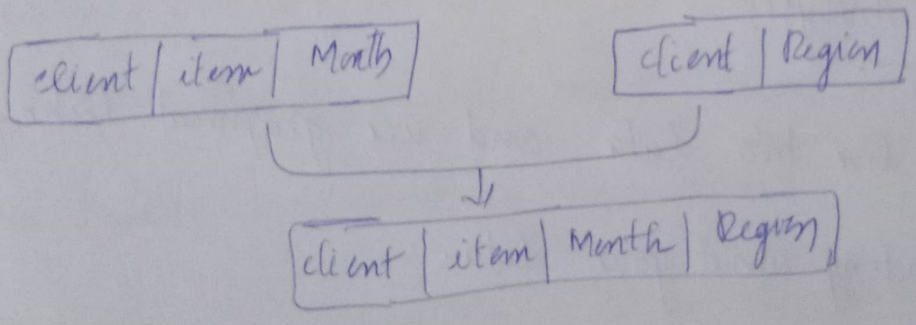
→ Different units of measurement (height in cm & m)

eg: height of student
5-6 F
↓
entry occur 10 F
↓
first we validate if it is wrong then we correct

Integrating Data

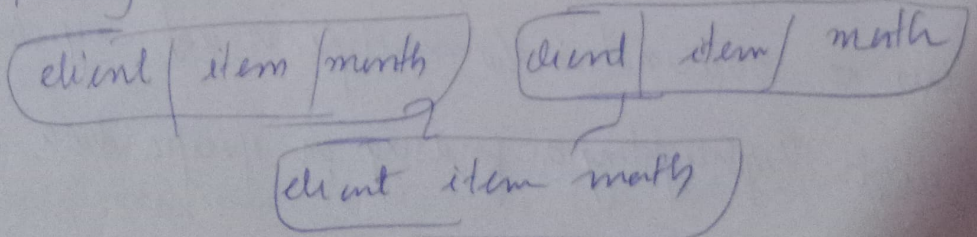
→ Data joining

→ Data appending



} Data joining (common attribute client)

Data appending - fields are same, additional storage

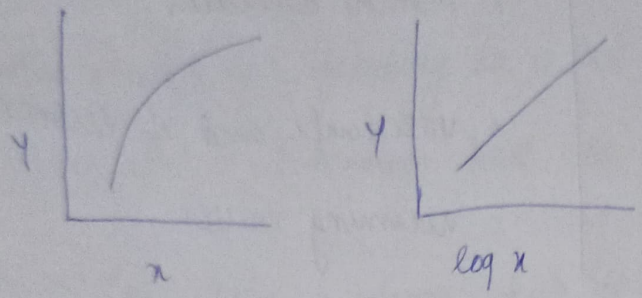


Data storage is problem
↓
overcome by virtual storage

Transforming Data

→ Reducing the no of variables is one way of transformation

→ Identifying the relationship b/w variables.



→ Turning variables to dummies. Dummy variables can take 2 variables True or False.

eg:

customer	year	Gender	sales
1	2015	F	10
2	2015	M	8
1	2016	F	11

⇒

customer	year	sales	Male	Female
1	2015	10	0	1
1	2016	11	0	1
2	2015	8	1	0

4. Exploratory Data Analysis

→ We take a deep dive into data and use graphical techniques to gain understanding about your data.

→ Linking and brushing:

- Linking refers to coordination b/w multiple visual representations
- Brushing refers to highlighting or focusing on specific data point within one visualization.

5. Building the model

model should be based on content → eg: weather forcasting suitable algorithms

① first we need to select the modeling technique and selecting the variables (features)

eg: we just import the library we dont need to build from the ~~stra~~ scratch.

② Execution of the model

③ Diagonising and compairing the model

other factors:
→ whether the model is to be used in the production environment
→ whether the model requires maintanence

2- Execution of the model

→ Model fit: (underlined patterns should be understand)

A good model fit means that the mode accurately captures the underlying patterns

eg: $r^2 = 85\%$ (best fit → which means model understand that housesize and location is depend on house price)

house price → housespize / house location

$r^2 = 0$, didn't analyze the model

$r^2 = 1$

R-squared indicates the proportion of variants in the dependent

variable that is predictable from the independent variable

eg: If you are predicting house prices not all houses are priced the same, there is a variation b/w prices based on factors like size, location etc.

$R^2 = 1$, indicates a better fit

$R^2 = 0$, indicates a model doesn't explained much of the variation

→ Model Comparison: It involves evaluating multiple models to determine which model is best fit for data or best at making the predictions.

6. Presenting Findings and Building applications on top of them

Presenting results to the stakeholders and industrializing your analysis process for repetative reuse and integration with other tools.