Topics for the day

① Numpy and Pandas Intro

② Data preprocessing techniques

Numpy → Numerical Python

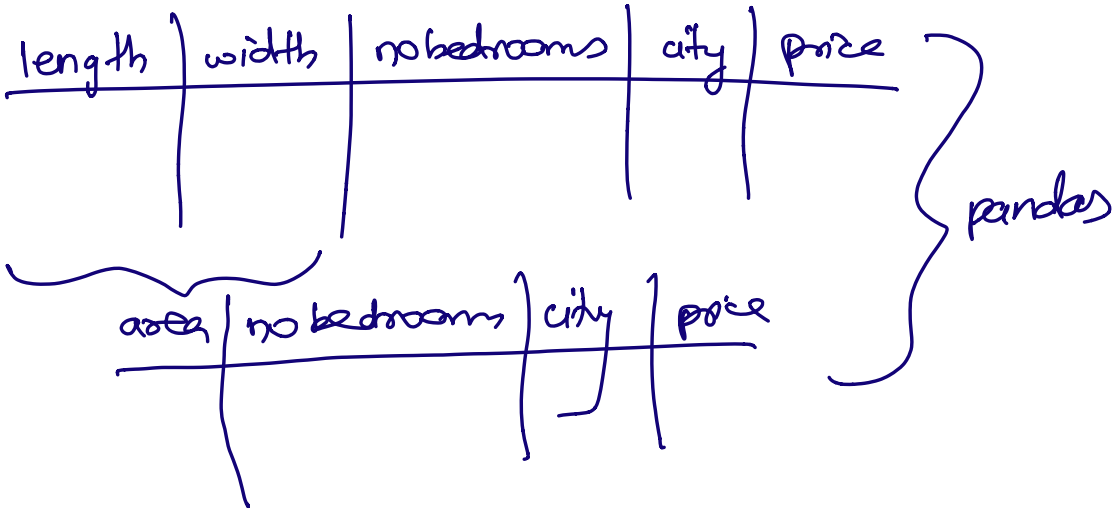
deal with arrays. (optimized - memory consumption, CPU speed) utilization

① One dimensional array

[1, 2, 3, 4]

② Two dimensional array

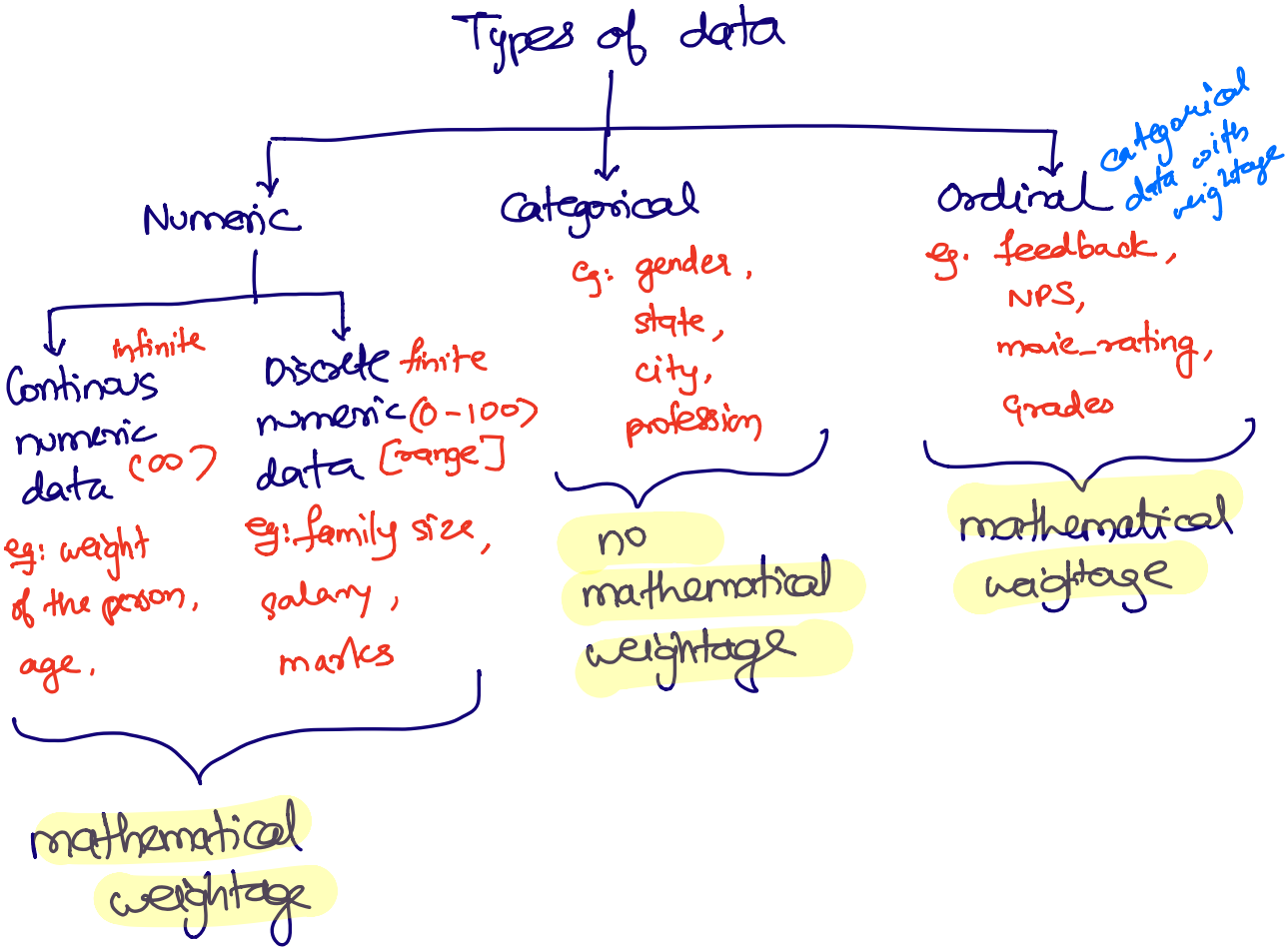
$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \end{bmatrix}$$



Machine Learning.

① File must hold logically structured data.

② All your data must be strictly numeric.



Machine learning program implementation steps

① Loading the data

② Pre processing activities & EDA

ⓐ General statistical summary.

ⓑ Seperate data as features and label. (ensure data is in the form of numpy array)

ⓒ Deal with missing data.

→ Statistics as a base

① Replace NaN with mean (continous numeric data)

② Replace NaN with median (discrete numeric data)

③ Replace NaN with Mode (Categorical data)

→ Domain knowledge as base

① Statistics Technique (mean, median, mode)

② Decide the default value for replacement.

GUIDELINES

if column data is string, ideally it is recommended to delete that record.

if label has missing data, delete that specific record

String Data (Handling missing data)

- Categorical / ordinal → use mode
- Pure string
  - ① Delete it
  - ② mode .... complex so not recommended

② Dealing with Categorical Data (creation of Dummy variables)

Eg:

eid	esal	ccity
1	1000	mumbai
2	2000	chennai
3	3000	mumbai

mathematical

① get the unique values of categorical column.

['mumbai', 'chennai']

⇒ ② Sort the list in ascending order

['chennai', 'mumbai']

= 0 1

eid	esal	ccity
1	1000	1
2	2000	0
3	3000	1

③ Replace the values of the column with the index values of the list created in step 2.

LABEL ENCODING

dummy variables

0	1	eid	esal
0	1	1	1000
1	0	2	2000
0	1	3	3000

④ Convert the column values into column itself such that the columns are loosing mathematical weightage.

ONE-HOT ENCODING

m, c, h, b → b, c, h, m
0 1 2 3

0	1	2	3	eid	esal
0	0	1	0	10	1000

Trainer Name	Subject (ML-DS)	Rating (1-5) discrete	Next class
Prashant	ML	4	y
Arun	DS	2	n
:	:	:	:

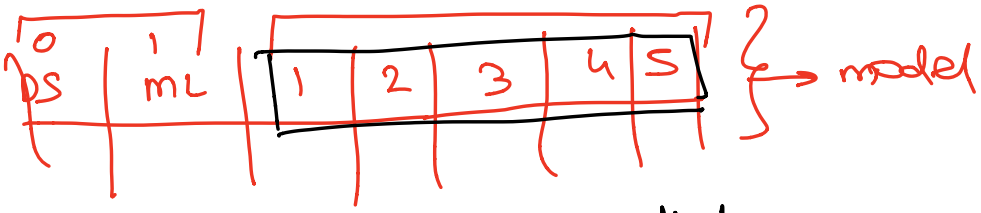
pure string — categorical (LE, OHE) — ordinal (OHE) — discrete

features → TW, S, R

Label → Next class

follow statistics — ✓

domain — ✗



ordinal

Subject	Rating
DS	Excellent
ML	Good
DS	Bad
ML	OK
:	Better
:	Excellent

Bad – 1
OK – 2
Good – 3
Better – 4
Excellent – 5

domain

replace string with ints

⇓

OHE