Page No.
Date 6 8 24

☆ Process of Data Science

1. Defination of Problem Statement
   → AIM
2. Data Acquisition
3. Data preparation / preprocessing
4. EDA → Exploratotary Data Analysis / feature selection
5. Model Planning → Finalizing Model
6. Visual Communication
7. Deployment

HW
1- Kaggle download - Titanic DB
review & discuss in next class

Page No.
Date

☆ Know your data

i) Size of data

→ df.shape

ii) How data look like

→ df.head()

iii) Detail of data

→ df.info

iv) Calculate Null value

→ df.isnull(.sum)

v) How data look mathematically

→ df.describe()

vi) To check duplicate

→ df.duplicated(.sum)

vii) To check the corelation between columns

→ df.corr()

Page No.
Date

☆ EDA univariate

I) Categorical data

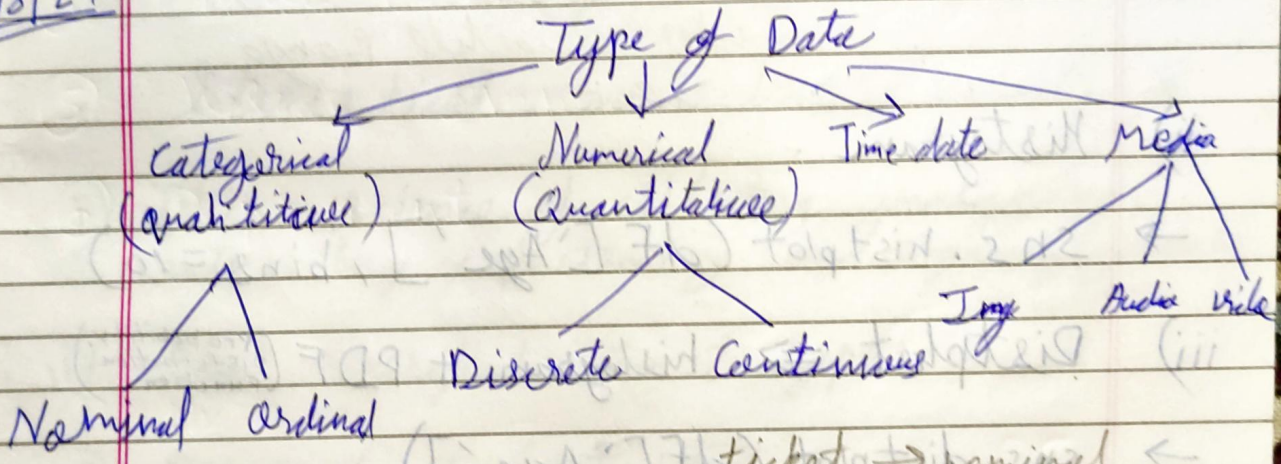
i) Pie chart

```
df['Age'].value_counts().plot(kind='pie', autopct='%.2f')
```

ii) Count Chart

```
sns.countplot(df['Sex'])
```

7/8/24



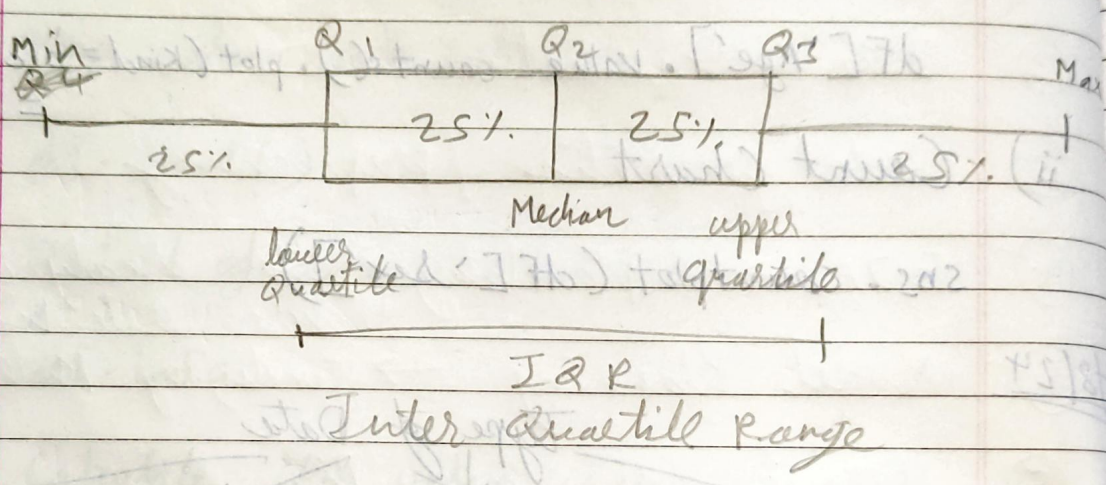
P_ID → Nominal
Survived → ~~Nominal~~
P_class → ordinal
Name → Nominal
Sex → Nominal
Age → ~~Discrete~~ Continuous
SibSp → Discrete
NW → Car Detail → toggle_data type

ticket → nominal
Parch →
Fare → continuous
cabin → nominal
embark → nominal

II) Numerical data

i) Box Plot

→ sns.boxplot(df['Age'])



ii) Histogram

→ sns.histplot(df['Age'], bins=10)

iii) Distplot ⇒ histogram + PDF (Probability Distribution Function)

→ sns.distplot(df['Age'])

☆ EDA Bivariate / Multivariate

i) Heat Map

HW diff probability vs likelyhood

Page No.
Date 8 8 24

- Probability – It is measure of likelihood of an event occuring

- Random Variable – variable

- Probability Distribution – description of possible value of each random variable

- Conditional Probability

$$P\left(\frac{B}{A}\right) = \frac{P(A \cap B)}{P(A)} = \frac{P(B)\,P(A/B)}{P(A)}$$

- Statistics

i) It is a science of collecting, analyzing, interpreting, presenting and organizing data

ii.) ~~Stats~~ classified into 2 Types

a) Descriptive Statistics

b) Inferential Statistics

iii) Also helps in hypothesis testing

iv) Correlation -

v) Regression - it is a model, the relationship between dependent variable and one or more independent variable

12/8/24

** Measures of Central Tendency

i) Mean (μ or $\bar{X}$) =

* Measure of dispersion

Page No.
Date 12 8 24

☆

i) Mean

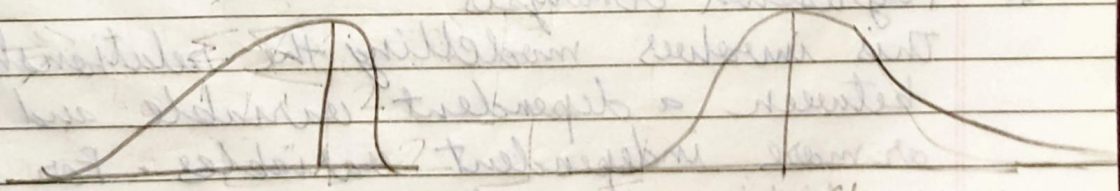
- central tendency
- $\bar{X} = \frac{\sum_{i=1}^{N} X_i}{N}$

ii) Variance

- Dispersion
- $\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \bar{X})^2}{N}$

iii) Skewness

- Symmetry (positive or negative)
- $\text{Skew} = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})^3}{\sigma} \right]$
-



-ve Skewness +ve Skewness

iv) Kurtosis

- Shape (Tall or flat)
- $Kurt = \frac{1}{N} \sum_{i=1}^{N} \left[\frac{(X_i - \bar{X})}{\sigma} \right]^4$
- Mesokurtosis - An excess kurtosis of 0
- Platykurtosis - A -ve excess kurtosis.
- Leptokurtosis - A +ve excess kurtosis

13/8/24

* ANOVA - Analysis of varience
This involves comparing means across multiple groups to determine if there are any significant differences. For ex, comparing the mean heights of individuals from different regions.

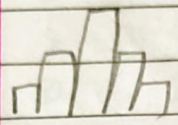
* Regression Analysis
This involves modelling the relationship between a dependent variable and one or more independent variables. For ex, predicting the sales of a product based on advertising expenditure.

Page No.
Date 13 8 24

* Chi-Square tests
This involves testing independence

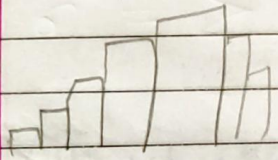
* Shapes of Histogram

i) Symmetric

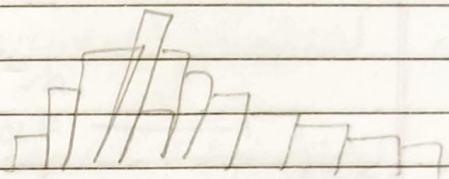


ii) Bimodal

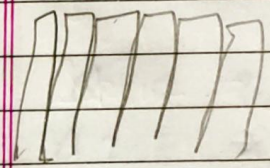
iii) Left Skew



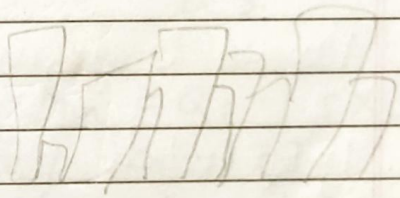
iv) Right Skew



v) Uniform



vi) No pattern



ANOVA - 1, 2 way
Distribution types

Page No.

Date 14 8 24

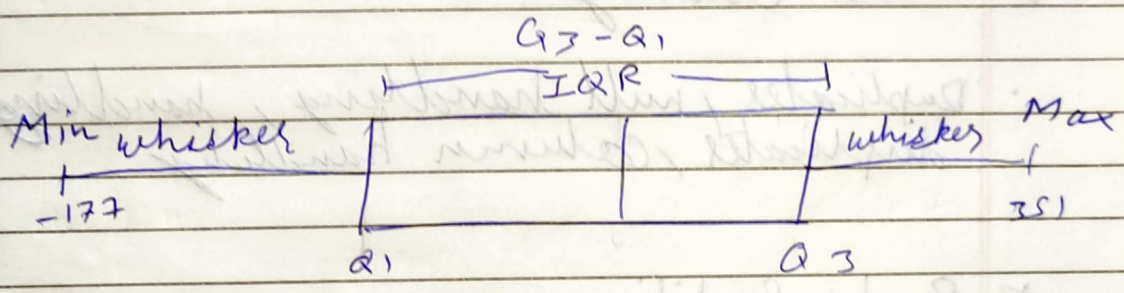
☆ ANOVA

Page No.
Date

☆ Boxplot example

1	6	-20
2	10	30
3	11	33 ← 34 Q1
4	31	35
5	35	~~37~~
6	50	40 ← 43 Q2
7	65	45
8	75	50
9	100	100 ← 200 Q3
10	105	300
11	201	500
12	700	700
13	1000	

IQR = 166
Min = -125
Max = 449



Q1 = 25% = 21
Q2 = 50% = 65
Q3 = 75% = 153

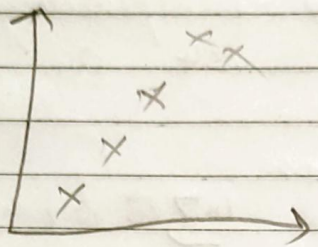
IQR = 132

Min = Q1 - 1.5 × IQR
= -177

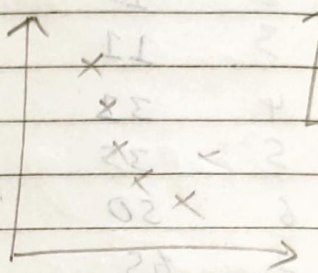
Max = Q3 + 1.5 × IQR
= 351

Page No.
Date

☆ Correlation



positively correlated



negatively correlated

$$Corr = \frac{Cov(x,y)}{x * y}$$

29/08/24

☆ Data Cleaning

- Duplicates, null handling, handling duplicates, column handling

☆ Pandas Profiling

```
!pip install pandas_profiling
from pandas_profiling import ProfileReport
prof = ProfileReport(df)
prof.to_file(output_file='output.html')
```

Page No.
Date 3 9 24

☆ Confusion Matrix

Actual \ Predicted	1	0
1	TP	FN
0	FP	TN

Actual
Same → True
Differ → False

Predicted
1 → Positive
0 → Negative

Ex

Actual y	Predicted $\hat{y}$
0	1
1	1
0	1
1	1
1	1
0	1
1	1
0	1
1	1
1	

T=10

	P=10	N=0
	6	0
	4	0

$$TPR = \frac{TP}{P} = 0.6 \text{ (Recall)}$$

$$TNR = \frac{TN}{N} = 0$$

$$FPR = \frac{FP}{P} = 0.4$$

$$FNR = \frac{FN}{N} = 0$$

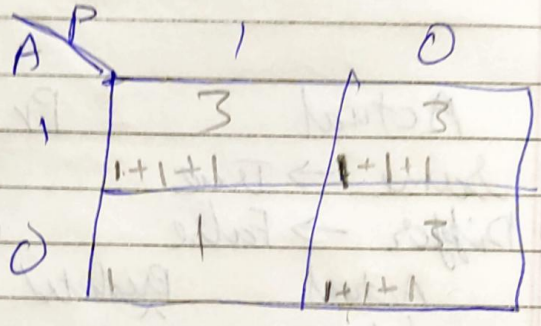
Page No.
Date

$$\text{Accuracy} = \frac{TP + TN}{T} = \frac{6+0}{10} = 0.6$$

$$\text{Precision} = \frac{TP}{TP+FN} = \frac{6}{6+0} =$$

Q

y	$\hat{y}$
0	0
1	1
0	0
1	0
1	0
1	0
0	0
1	1
0	1
1	1



$T = 10$

$$TPR = \frac{TP}{P} \quad 75\% \text{ (Recall)}$$

$$TNR = \frac{TN}{N} \quad 50\%$$

$$FPR = \frac{FP}{P} \quad 25\%$$

$$FNR = \frac{FN}{N} \quad 50\%$$

$$\text{Accuracy} = \frac{TP + TN}{T} = 60\%$$

$$\text{Precision} = \frac{TP}{TP+FN}$$

Page No.
Date

Q

y	$\hat{y}$
0	0
1	1
2	2
2	2
1	1
0	1
1	1
2	1
0	1
1	1

A \ P	0	1	2
3 − 0	1 1	2	0
4 − 1	0	1+1+1+1	0
3 − 2	0	1 1	1+1 2

Q

y	$\hat{y}$
1	1
1	0
1	0
1	1
0	1
0	0
0	1
0	0
1	0
0	0

A \ P	P	N
1	TP 2	FN 3
0	FP 2	TN 3

$$\text{Recall} = TPR = \frac{TP}{\text{Actual P}} = \frac{2}{5} = \frac{TP}{TP+FN} = 40\%$$

$$TNR = \frac{TN}{\text{Actual N}} = \frac{3}{5} = \frac{TN}{TN+FP} = 60\%$$

$$FPR = \frac{FP}{\text{Actual P}} = \frac{2}{5} = \frac{FP}{FP+TN} = 40\%$$

$$FNR = \frac{FN}{\text{Actual N}} = \frac{3}{5} = \frac{FN}{FN+TP} = 60\%$$

$$\text{Precision} = \frac{TP}{\text{predicted P}} = \frac{TP}{TP+FP} = \frac{2}{4} = \frac{1}{2} = 50\%$$

$$\text{Accuracy} = \frac{TP+TN}{T} = \frac{5}{10} = \frac{1}{2} = 50\%$$

Page No.
Date 12 9 24

Q $\underset{\begin{matrix}1\,2\,3\\5\,6\,7\end{matrix}}{X} \Rightarrow \underset{\begin{matrix}1-2-3\\5-6-7\end{matrix}}{X}$

Hypothesis Testing - Unit 1

Q1 Population of Gandhinagar is more than 1Cr

H0: population of gandhinagar is greater than 1 Cr

H1: population of gandhinagar is less than 1 Cr

Q2 Salary of Ram and Shyam is not same

H0: Salary of Ram & Shyam are not same

H1: Salary of Ram & Shyam are same

Q.3 Value of Area is more that 100 Km^2

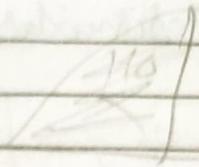
H0: Value of area is greater 100 Km^2

H1: Value of area is less than 100 Km^2

Page No.
Date

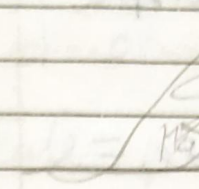
Q $H_0 >= 5$

$H_a \neq 5$



Q $H_0 : \mu = 10$

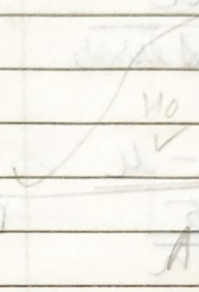
$H_a = \mu > 10$ or $\mu < 10$



Q

$H_0 \Rightarrow A \neq B$

$H_a \Rightarrow A = B$



13/09/24

Significant value (α)

CI = ~~EX~~ $1 - \alpha$

CI $\Rightarrow$ 95%.

$\alpha \Rightarrow 5.0\% = 0.05$

Ex Ex $\mu_{1a} = 100$

$n = 30$

$\bar{x} = 100$

$\sigma = 20$

$x_1 = 50 \quad n_1 = 30$

$x_2 = 60 \quad n_2 = 30$

$\sigma_1 = 20, \quad \sigma_2 = 50$

degree of freedom
$= n - 1$

Page No.
Date

H_0 : similes as equal

for one sample	for Two sample Independent	for Two sample dependent
t-value = $\frac{\bar{x} - \mu}{\text{standard error}}$ $= \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$	t-value = $\frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma^2}{n_2}}}$ DOF $= n_1 + n_2 - 2$	t-value = $\frac{\bar{x} - 0}{\frac{s}{\sqrt{n}}}$ DOF $= n - 1$
(1) $= \frac{110 - 100}{\frac{20}{\sqrt{30}}}$ (t-value) 1.90 $= \frac{10\sqrt{30}}{20}$ $= 2.73$		
(2) $\frac{95 - 100}{\frac{20}{\sqrt{30}}}$ (t-value) 1.10 $= -1.36$		

C.I	90%	95%	99%	99.9%
1 tail	0.1%	0.05	0.01	
2 tail	0.05%	0.025	0.005	

	True	False
Rejected	×	✓
Not Rejected (Accepte)	✓	×

17/9/24

Type 1 error : when H_0 is true we may reject it

$P(\text{Reject } H_0 \text{ when it is true}) = P\left(\frac{\text{Reject } H_0}{H_0}\right) = \alpha$

Type 2 error : when H_0 is false we may accept it

$P\left(\frac{\text{Accept } H_0}{H_1}\right) = \beta$

Q2. $\mu_0 = 5g$

$\bar{x} = 5.05g$

$\sigma = 0.1$

$n = 12$

$H_0 : \mu = 5$ ✓

$H_1 : \mu > 5$ ✓

Sol $$t_c = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

$$t_c = \frac{5.05 - 5}{0.1/\sqrt{12}}$$

$$= \frac{\pm 0.05}{0.1/2\sqrt{3}}$$

$$= \pm \frac{0.05}{0.1} \times 2\sqrt{3}$$

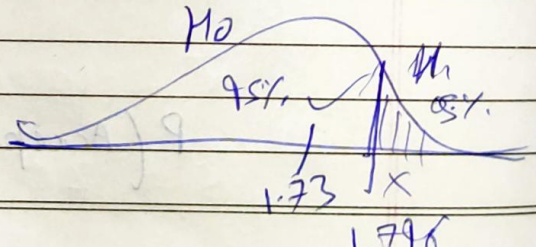
$$= \pm 0.5 \times 2\sqrt{3}$$

$$= \pm \frac{8}{10} \times \pi \times \sqrt{3}$$

$$= \pm\sqrt{3}$$

$t_c = \pm 1.73$

Accept H_0



Page No.
Date

Q2 $\mu = 42$ min

$n = 12$ journeys

$\bar{x} = 50$ min

$\sigma = 15$ min

CI = 5%

$H_0 : \mu = 42$ min

$H_1 : \mu = 50$ min ~~$= 50$ min~~ > 42 min

Sol $t_c = \frac{50 - 42}{15/\sqrt{12}}$

$= \frac{8}{15/2\sqrt{3}}$

$= \frac{8}{15} \times 2\sqrt{3}$

$= 1.84$

Reject H0

failed to re

Page No.
Date

Q.3 $n = 25$

$\mu_0 = 2.45$ $\quad H_0 : \mu = 2.45$

$\bar{X} = 2.65$ $\quad H_1 : \mu > 2.45$

$\sigma = 0.35$

$CI = 99\%$.

Sol $\quad t_c = \dfrac{2.65 - 2.45}{0.35/\sqrt{25}}$

$= \dfrac{0.2 \times 5}{0.35}$

$= \dfrac{2 \times 8}{3.5_{0.7}}$

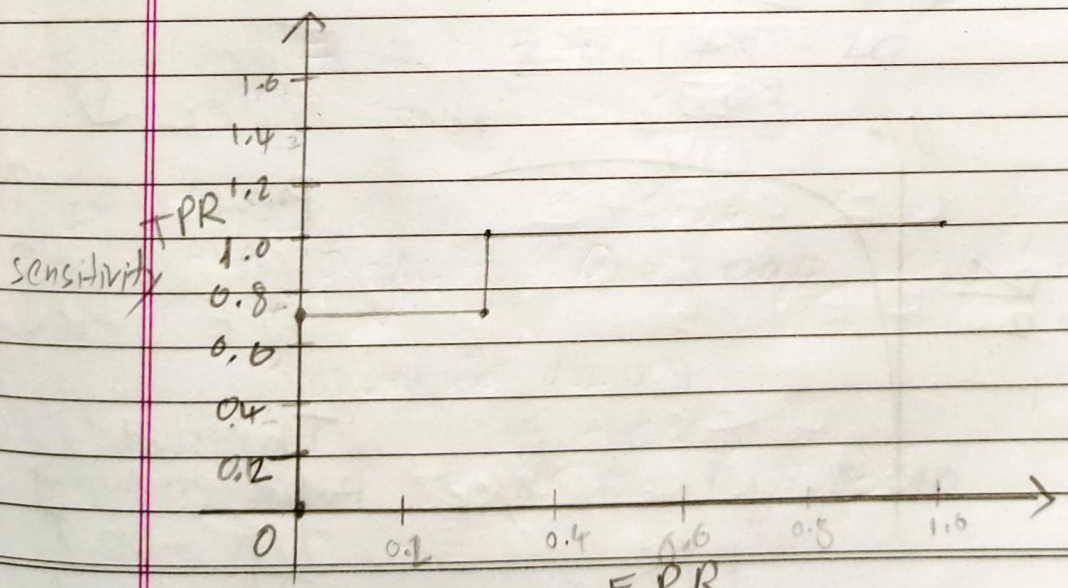
$t_c = 2.85$

Page No.
Date 18 9 24

Score	Prediction (1)	" (0.7)	" (0.6)	" (0.5)	" (0) $\hat{Y}$	Y
0.25	0	0	0	0	1	0
0.45	0	0	0	0	1	0
0.55	0	0	0	1	1	1
0.67	0	0	1	1	1	0
0.82	0	1	1	1	1	1
0.95	0	1	1	1	1	1
FP	0	0	1	1	3	
TP	0	2	2	3	3	
TPR	0	0	$\frac{1}{1+2} = \frac{1}{3}$	$\frac{1}{1+2} = \frac{1}{3}$	1	
TPR	0	$\frac{2+1}{2+1} = \frac{2}{3}$	$\frac{2}{2+1} = 2/3$	$\frac{3}{3+0} = \frac{3}{3}$	1	

$$FPR = \frac{FP}{FP + TN}$$ = Actual 0

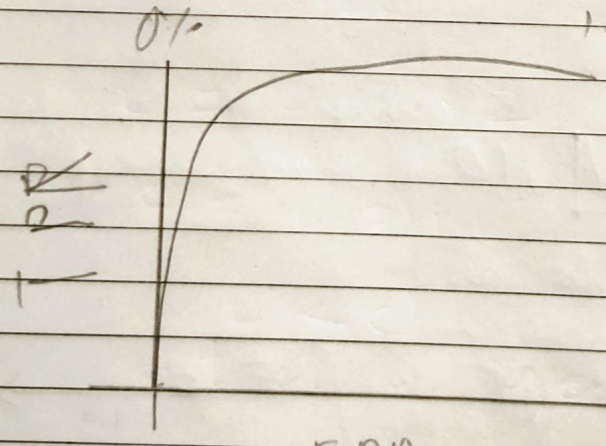
$$TPR = \frac{TP}{TP + FN}$$ Actual 1



Page No.
Date 18 9 24

★ ROC

The Receiver operating Characteristics



Threshold value dec ↓

P+ve predict ↑

True/False +ve ↑

T/F -ve ↓

Threshold ↑

-Ve prediction ↑

T/F -ve ↑

T/F +ve ↓

Page No.
Date

Ex2

Sol $\bar{x} = 20.175$

$\mu = 20$

$\sigma = 3.02$

$n = 12$

$H_0 : \mu = 20$

$H_1 : \mu > 20$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{20.175 - 20}{\frac{3.02}{\sqrt{12}}}$$

$$t = 0.2007$$

$c = 1.79$

$t < c$

$\therefore$ Fail to Reject H_0

Ex 1

Sol $\bar{x} = 21$

$\mu = 20$

$n = 25$

$\sigma = 7$

$H_0 : \mu = 20$

$H_1 : \mu > 20$

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

$$t = \frac{21 - 20}{\frac{7}{\sqrt{25}}}$$

$$t = 0.7142$$

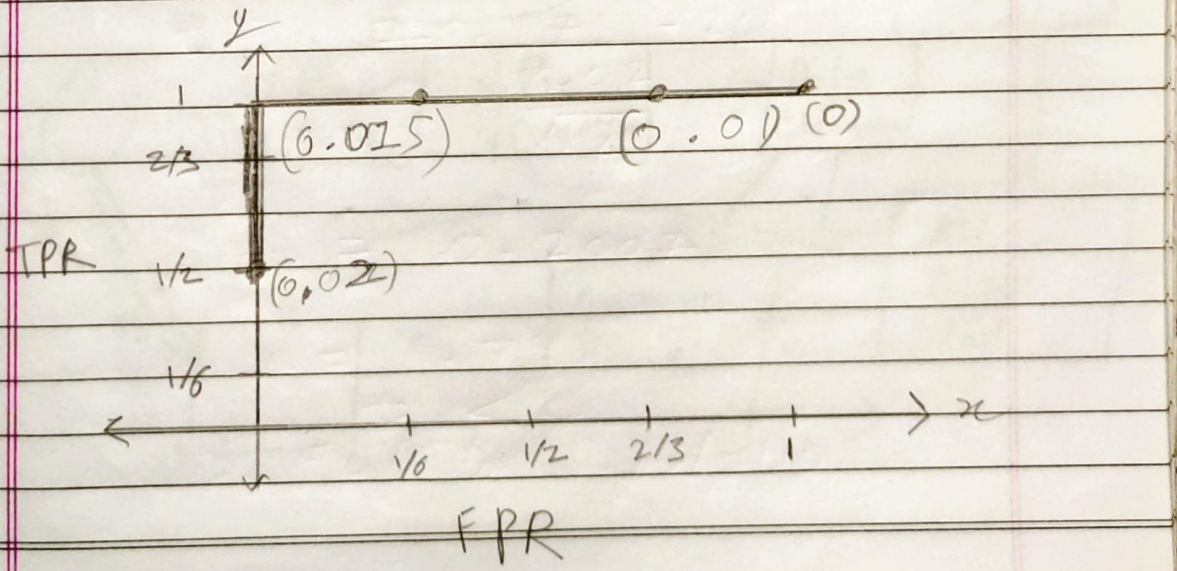
$$c = 1.711$$

$$t < c$$

$\therefore$ Fail to reject H_0

Ex ROC

Instance	Cut-off	0.020	0.015	0.01	0	Actual
1	0.008	0	0	0	1	0
2	0.011	0	0	1	1	0
3	0.021	1	1	1	1	1
4	0.009	0	0	0	1	0
5	0.014	0	0	1	1	0
6	0.015	0	1	1	1	0
7	0.012	0	0	1	1	0
8	0.015	0	1	1	1	1
TP		1	2	2	2	
TN		6	6	2	0	
FP		0	1	4	6	
FN		1	0	0	0	
TPR		1/2	1	1	1	
FPR		0	1/6	2/3	1	



Page No.
Date 19 9 24

Quiz

1) FP
2) Both T/F +ve ↓
3) Both T/F -ve ↑
4) A
5) F +ve ↓
6) No
7) Recall ↓ if precision ↑
8) False

Page No.
Date 19 9 24

☆ EDA (Exploratory Data Analysis)

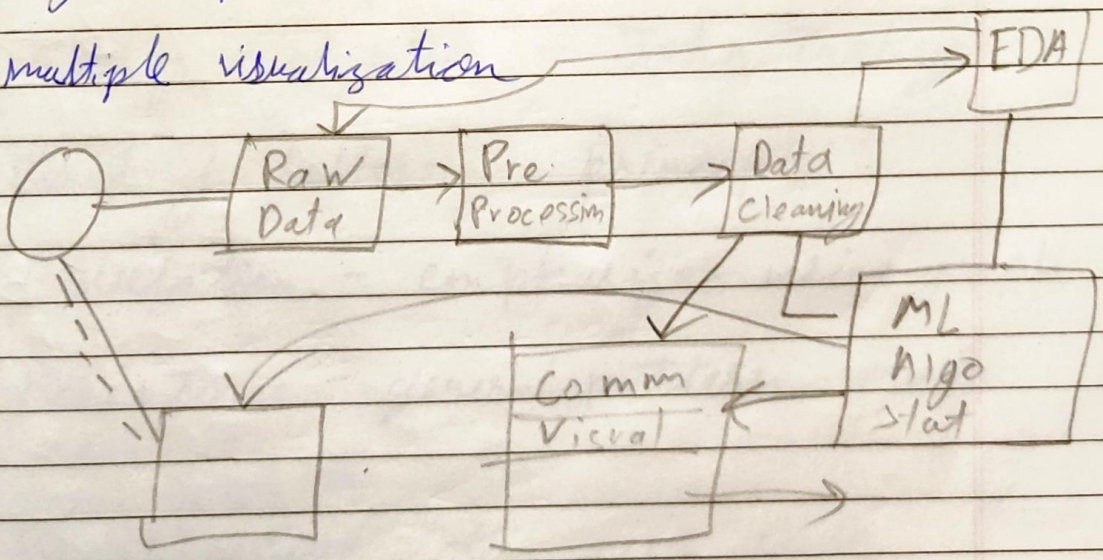
Q Is EDA & DA the same

Sol DA - descriptive, diagnostic, predictive and prescriptive

EDA - synonymous with descriptive analysis, where one explores hidden relationships and patterns in the ~~data~~ available data

Q How one should write an EDA report

Sol Thorough explanation of dataset variables
- their correlation
- any preprocessing perf on ds for ML algo
- multiple visualization



Page No.
Date

Types of EDA

EDA

Univariate

graphical non graphical

Multivariate

graphical non graphical

Page No.
Date 23 09 24

★ Tools of EDA

1. Python

2. R

3. MATLAB

★ Philosophy of EDA

Father of EDA - John Tukey

Based of following principles

1. Revelation - emphasizes using graph
2. Resistance - general pattern which co
3. Reexpression: to new scale

Page No.
Date

1. Curiosity and

Page No.
Date 23 09 24

Unit 3

* Basic ML Algos

1) Association Rule mining

- unsupervised learning technique
- tries to find rel^n between / among the variables of dataset
- relation / association
- Ex - Market Basket analysis, Web usage mining, etc
- Working
  - If and Else statement concept

- Measure of associations bet^n multiple of data

i) Support

$$\text{supp}(X) = \frac{\text{Freq}(X)}{T}$$

ii) Confidence

$$\text{confidence} = \frac{\text{Freq}(X, Y)}{\text{Freq}(X)}$$

iii) Lift

$$\text{Lift} = \frac{\text{Supp}(X, Y)}{\text{Supp}(X) * \text{Supp}(Y)} \quad \left(\because \text{Supp}(X,Y) = \frac{\text{Freq}(X,Y)}{T}\right)$$

Q Ex

$B \rightarrow C$

TID	Itemsets
T1	A, B
T2	B, D
T3	B, C
T4	A, B, D
T5	A, C
T6	B, C
T7	A, C
T8	A, B, C, E
T9	A, B, C

i) $\text{Supp} = \frac{6}{9}$

ii) $\text{Confidence} = \frac{4}{6}$

iii) $\text{Lift} = \frac{4/9}{6/9 \times 6/9} = \frac{4}{\frac{36}{81}} = 1$

- If Lift = 1

- If lift > 1

- If lift < 1

★ Bo ML A

- Machine Learing (ML)
- Subset of AI
- involves training of computers to learn
- from data without being programmed

- Identify patterns, make predictions

- performs tasks that usually require human intervention

★★ Key Components of ML Algo

1 Data - raw material . un/structured (ex - csv)
2 Features - relevent characteristics
3 Model -
4 Algorithm -

Types of ML Algos

1. Supervised Learning

- Regression
  - predicting a continuous numerical value (ex - house price)
  i) Linear
  ii) Logistic
  iii) Decision Tree
  iv) Random Forests

- Classification
  -

  i) Decision Tree
  ii) Random Forest
  iii) Neural Networks

2. Unsupervised Learning

- Clustering - grp data pts into similar clusters

  i) K-means clustering
  ii) hierarchical clustering
  iii) DB SCAN

3. Reinforcement Learning - learn through trial and error, interacting with an environment and receiving rewards or penalties.

Page No.
Date

Machine Learning

Supervised Learning | Unsupervised Learning | Reinforcement Learning

Classification | Regression | Clustering

Page No.
Date 24 9 24

☆ Types of Regression Algos

1. Linear
2. Random Forests
3.

Usage

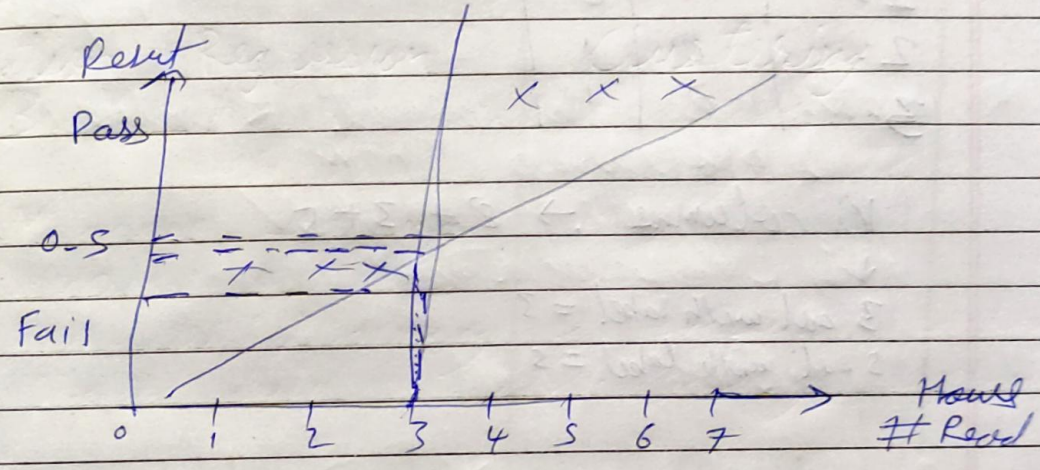
1
2
3

(1) Linear Regression

Mean Square Error

Page No.
Date 25 9 24

# hours Read	Result
1	Fail
2	Fail
3	Pass
4	Pass
5	Pass
6	Pass
7	Pass



$y = h_\theta(x) = mx + c$
$= \beta_0 x + \beta_1 x$

$g(z) = \frac{1}{1+e^{-z}}$ Sigmoid function

Logistic Regression (x)

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

Page No.
Date 26 09 24

* Encoding Categorical Data

① Label Encoding

Low = 1
High = 2

High
Low
Low

② One - hot Encoding (N categories adding new n columns)

	Female	Male
1.	1	0
2.	0	1
3.	0	1

old → new

10 columns → 8 + 3 + 5

↓

3 col with label = 3
5 col with label = 5

id	name	grade
1	X	A
2	Y	B
3	Z	A

⇒

id	name	A	B
1	X	1	0
2	Y	0	1
3	Z	1	0

Page No.
Date

* Model Evaluation

① R-Squared Value — close fit (nearer to 1)

② Mean-Square Error — (lower MSE)

Q Example 1

old - x =	neuron (w/loop)
1) 10 col in which 2 col house 5 category 10	8+5+5 18
old	neuron (w/o op)
2) 2 columns with 3 categories 1 col with 10 columns 20	17+3+3+10 33

30/9/24

☆ Classification Techniques

2. logistic regression

2. K nearest neighbors (Knn) (K - odd)

② KNN

If K is too low

If K is very high

hyperparameter = K - no. of nearest neighbor

* ~~bootstrap~~ how to decide K value?
  i) Bootstrap
  ii) Square root of N
  iii) cross validation

* Application
- Null Handling → Simple Imputer (preprocessing)
- Classification

* Advantage
- Easy to implement
- less hyperparameter
- Adjustable

* Disadvantage

- Scale
- Curse of Dimensionality
- overfitting / Underfitting

2/10/24

(3) Naive Bayes

→ Theorem - Bayes theorem

$$P(A/B) = \frac{P(B/A)\ P(A)}{P(B)}$$

*

i) Prior

ii) Conditional

* Disadvantages

i) Basic assumption like independent event are not real life/world data

ii) Zero frequency for a new keyword, gives zero probability as outcomes

Page No.
Date

* Advantages
  - easy & efficient
  - effective with large no. of features
  - can handle high dimensional data
  - less complex
  - scales well
  - overfitting

* Applications
  - Spam Email Filtering
  - Text Classification

Example

W1	W2	W3	Spam
Sale	offer	Fast	Y
Sale	Click	Quick	Y
Test	Offer	Money	N

Predict

W_1 = Sale	W_1 = Sale
W_2 = Money	W_2 = offer
W_3 = Money	W_3 = Fast
Y = 0	Y = 1
N = 0	N = 0

(1) Prior probability

$P(Y) = 2/3$
$P(N) = 1/3$

(2) Find conditional Probability

W_1	Y	N
Sale	2/2	0/1
Test	0/2	1/1

W_2	Y	N
offer	1/2	0/1
Click	1/2	0/1

W_3	Y	N
Fast	1/2	0/1
Quick	1/2	0/1
Money	0/2	1/1

(3) $P(Y/W_1 = \text{Sale}, W_2 = \text{Money}, W_3 = \text{Money})$

$= P(Y) * P(W_1 = \text{Sale}/Y) * P(W_2 = \text{Money}/Y) * P(W_3 = \text{Money}/Y)$

$= \frac{2}{3} \times 1 * 0 \times 0 = 0$

$P(N/W_1 = \text{Sale}, W_2 = \text{money}, W_3 = \text{money})$

$= \frac{1}{3} * 0 \times 0 \times 1$

$= 0$

Example 2

$P(Y) = 5/10$

$P(N) = 5/10$

	Y	N
Red	3/5	2/5
SUV	1/5	3/5
domestic	2/5	3/5

$$P\left(\frac{Y}{C=Red, T=SUV, O=domestic}\right) = \frac{5}{10} \times \frac{3}{5} \times \frac{2}{5} = \frac{3}{125}$$

$$P\left(\frac{N}{C=Red, T=SUV, O=domestic}\right) = \frac{5}{10} \times \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} = \frac{9}{125}$$

i) * Entropy

$$E = -\sum_{i \in C} P_i \log_2 P_i \qquad \frac{\log_{10}^{x}}{\log_{10}^{2}} = \log_2^{x}$$

Ex

Play TT
Yes
Yes
No
No
No
No

$C = \{Yes, No\}$

$$P(Yes) = \frac{2}{6} = \frac{1}{3} \quad \log_2 \frac{1}{3} = \frac{1}{3}(-0.584)$$

$$P(No) = \frac{4}{6} = \frac{2}{3} \quad \log_2 \frac{2}{3} = \frac{2}{3}(-0.176)$$

$$E = -\sum_{i \in C} P_i \log_2 P_i$$

$$= -\left[(-0.472) \times \frac{1}{3} + \frac{2}{3}(-0.176)\right]$$

$$= -[-0.528 - 0.384]$$

$$= -[-0.917]$$

$$= 0.917$$

Page No.
Date

ii) * Gini Impurity

$$= 1 - \sum_{c \in C} P_c^2$$

Yes
Yes
No
No
Maybe

Ex $C = \{Yes, No, Maybe\}$

$P(yes) = \frac{2}{6} = \frac{1}{3}$

$P(No) = \frac{3}{6} = \frac{1}{2}$

$P(Maybe) = \frac{1}{6}$

$$GI = 1 - \left[\left(\frac{1}{3}\right)^2 + \left(\frac{1}{2}\right)^2 + \left(\frac{1}{6}\right)^2\right]$$

$$= 1 - \frac{1}{9} + \frac{1}{4} + \frac{1}{36}$$

$$= 1 - \frac{4 + 9 + 1}{36}$$

$$= 1 - \left[\frac{14}{36}\right]$$

$$= 1 - \frac{7}{18}$$

$$= \frac{11}{18}$$

E =

Page No.
Date

24/10/24

☆ Decision Tree

iii) * Information Gain

A - attribute
S - collection of dataset

$$Gain(S, A) = Entropy(S) - \sum_{v \in Value(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

$$= -(0.64\ (-0.63) + (-0.35(-1.48))$$
$$= +0.403 + 0.51$$
$$= +0.913 \rightarrow \approx 0.940$$
$$\approx +0.940$$

18/11/24

* Recent Trends in various data collection and analysis techniques

☆ Unit 5

1) Data Visualization

art & science of communicating info visually

* Types of Visualization

1. Histogram / Line chart
2. Bar chart
3. Piechart
4. Bar chart
5. Scatter plot
6. Heatmap
7. Map
8. Network Diagram
9. Tree
10. Boxplot

* Technologies

D3.js
Plotly
Matplotlib

Page No.
Date

* Tools

- Tableau
- Power BI
- QlikView
- Google Data Studio
- Excel

* Techniques

i) Colour Coding
ii) Shading
iii) Symbol Size
iv) Line Thickness
v) Animation
vi) Interactive Elements

18/11/24

Page No.
Date

* Unit 3

Feature Engineering

→ Techniques of feature selection

i)
ii)
iii)

* Unit 4

* Common Distance Measures

i) Euclidean Distance

$\sqrt{x_1^2 - x_2^2}$

ii) Manhattan Distance

$= |P_1 - P_2|$

iii) Minkowski

$= ((x_1 - x_2)^x)^{1/x}$

Q In Sklearn, which distance technique ~~they~~ it is using?

Ans Euclidean

* Different Clustering ~~Tech~~ techniques

1.) Hierarchical

2.) Partitional

3.) Fuzzy

4.) Model-Based

* Cluster distance measures

- Single linkage

$D(c_1, c_2) = \min_{\substack{x_1 \in c_1 \\ x_2 \in c_2}} D(x_1, x_2)$

- Complete linkage

$$D(c_1, c_2) = \max D(x_1, x_2)$$

- Average linkage

$$D(c_1; c_2) = \frac{1}{|c_1|} \frac{1}{|c_2|} \sum\sum D(x_1, x_2)$$

- Centroids

$$D(c_1, c_2) = D\left(\left(\frac{1}{|c_1|}\sum_{x \in c_1} \bar{x}_1\right), \left(\frac{1}{|c_2|}\sum_{x \in c_2} \bar{x}_2\right)\right)$$

- Ward's Method

$$TD_{c_1 \cup c_2} = \sum_{x \in c_1 \cup c_2} D(x, M_{c_1 \cup c_2})^2$$

Jump level 2 off

Page No.
Date

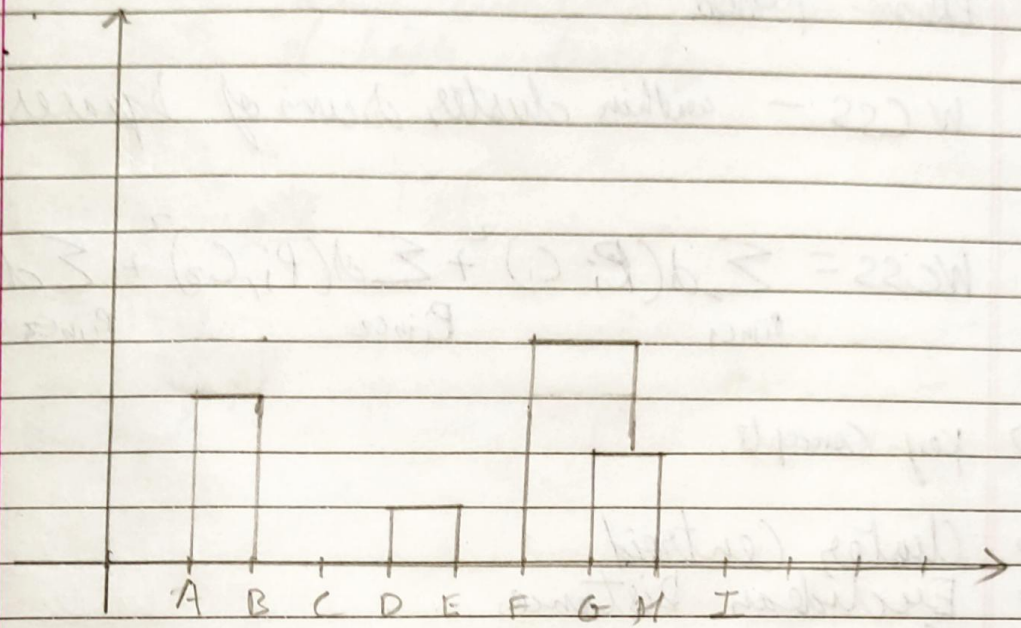
(1) PL

(2) GH

(3) AB

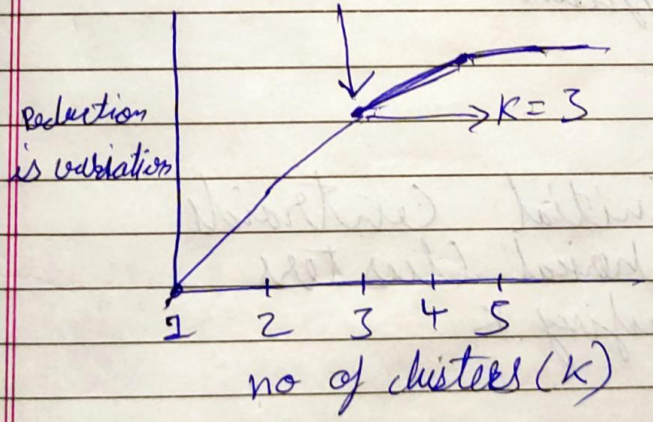
(4) GH - F

★ Dendogram



★ Kmeans - Clustering

elbow plot



stat Questionth Tosh Stat

YT

Page No.
Date

27/1/24

* Kmeans Clustering

Elbow Method

WCSS - within clusters sum of squares

$$WCSS = \sum_{P_i \text{ in } C_1} d(P_i, C_1)^2 + \sum_{P_i \text{ in } C_2} d(P_i, C_2)^2 + \sum_{P_i \text{ in } C_3} d(P_i, C_3)^2$$

→ Key Concepts

- Cluster Centroid
- Euclidean Distance
- Convergence

Strength

- Simple & Efficient
- Scalable

Weakness

- Sensitive to Initial Centroids
- Assumes Spherical Clusters
- Requires Specifying K

Page No.
Date

22/11/24

* DB Scan - density based Spacial Clustering of applications with Noise
This algo defines clusters as continuous regions of high density

Key concepts | rd

- core pt
- Border pt
- Noise pt

min = 4
ε - pen cap

** Strength
- Handles clusters of Arbitrary Shape
- Noise Tolerance
- Does not require the no of clusters

* Weakness
- Sensitivity of Parameters
- Clustering Density Variations

blog . DailyDoseofDS.com

25/7/24

☆ Choosing the right clustering approach:

- No of clusters
- cluster shape
- presence of noise & outliners
- computational complexity
- Interpretability.

* • Distance Matrics
* • Feature ~~Selection~~ Scaling

(1) Heirarchial Clustering

- strength - not defining no of clusters any shape, visualize & interpret
- weakness - combining all pts into one clusters less columns / small data set

(2) K means clustering

S - large dataset
hyper parameter
- Simple & efficient

W - defining no of clusters
non spherical
sensitive to initial centroid loc

(3) DBScan

S non spherical, ~~large data sets~~, not define no of clusters, handle noise & outliers

W- sensitive, for min pts & ε (distance / radius) density
computationally expensive for large dataset

(4) Fuzzy Clustering

S - can handle overlapping clustering
- more realistic representation of data

W - can be computationally expensive
- Interpretation of results can be challenging

Model-Based clustering (Gaussian Mixture Model)

S- handle - model complex data dist
overlapping clusters
estimate no of clusters

W- sensitive to

Page No.
Date

☆ Clustering Tendency

Y?
- avoids futile clustering
- optimizes algo selection

how?
- visual Inspection - scatter & density plots
- state test - hopkins test
- model-based approaches

* Specific application in Data Mining

customer segmentation
Anomaly detection
Document clustering