29/7/25

Machine Learning & Data Science

→ Basics of ML, different from traditional programming, types, of challenges

→ Visualis^n, plotting — Numpy
— Pandas
— Matplotlib
— sklearn

→ Statistics/statistical — train, test
— Validation
— K-cross valid^n
— Performance evalu^n for classific^n

→ Classific^n — KNN
— Naive Bayes
— Decision Tree → Ensemble learning.

→ Regression → Mathematics
→ Linear
→ Logistic
— Confidence & Interval
— ANOVA

→ Regularization & Kernal method.

→ Clustering — EM
— Mixture models
— K Means
— DBSCAN
— Hierarchical Clustering
— Optics

29/12/25 → DR — tSNE
— PCA

→ Deep learning — NN
— Back propagation
— Training methods

Book: ①

② Hands on ML with scikit-Learn, Keras, Tensor flow

③ Data mining - Concepts & Techniques

Evaluation:

→ Mid Term 25 M

→ End Term 40 M

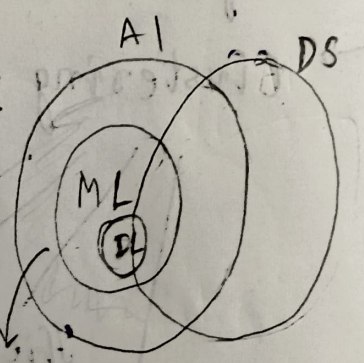
→ Assignment 2 — 15 M
— 10 M

→ 1 or 2 surprise tests : 10 M

80% attendance is compulsory.

AI ML DS

AI - A mechanism that can do everything that a human does.



Less data - ML

Huge data, hidden characteristics - DL

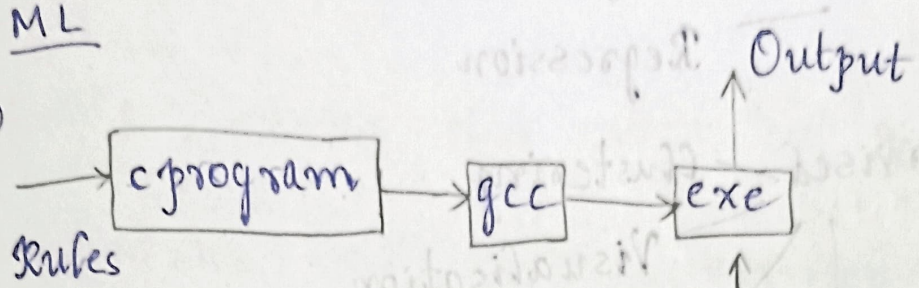
used for generating end product. (IOT sensors, etc)

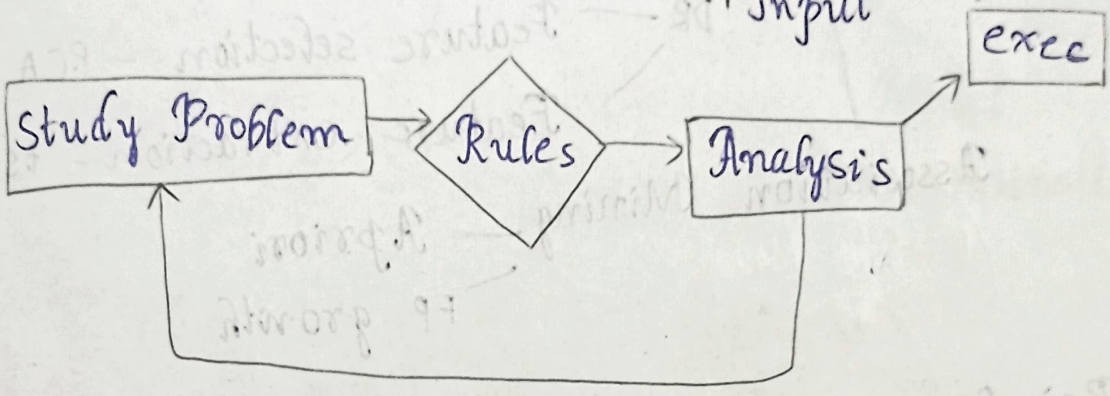
29/7/25

DS — Visualisation
— Missing
— ML

ML

①





Does this involve any learning? No

→ Here, in C we will set our rules

→ In ML, machine sets its own rules.

ML — Human Supervision — Supervised
— Unsupervised
— Semi-supervised
— RL

— Learning Style — Batch
— Online Learning

— Data Points comparison — Model based learning
— Instance based learning

Online Learning - Model keeps updated on real-time data

Batch Learning - We need to give the data in 1 go

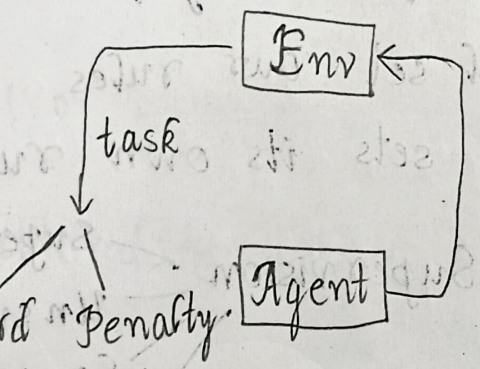
30/7/23

Once the training is completed, we cannot give any data again.

Supervised — Classific^n
— Regression

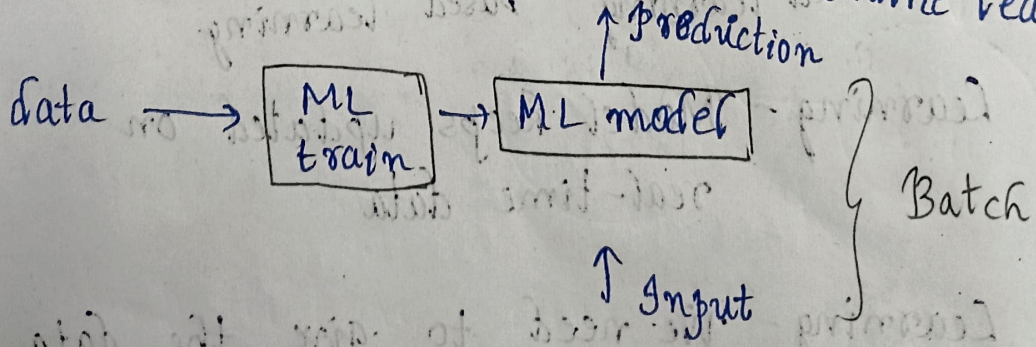
Unsupervised — Clustering
— Visualisation
— DR — Feature selection - PCA
— Feature extraction - tSNE
— Association Mining — Apriori
— FP growth

Reinforcement (No Human)

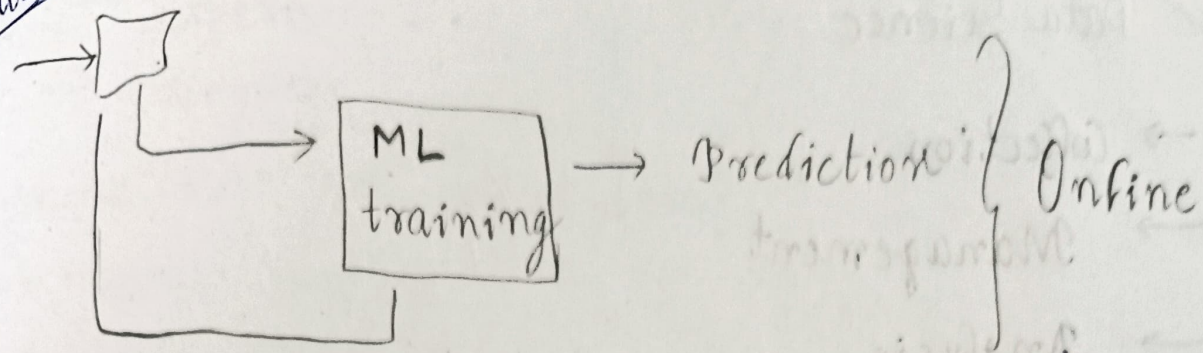


Policy → Loss
↓

② Based on Learning style — Batch
— Online Learning



30/2/26



Online Learning also called "out of core L"

③ Based on Data points comparison
- Instance (small dataset)
- Model (Discriminative)

Challenges behind ML

→ Data biased → Not represent

→ Complex data → Low training accuracy

Underfitting → low (training + testing) accuracy.

→ Overfitting
- ↳ Reduce model complexity
- ↳ Set threshold value
- ↳ Work on data set
- ↳ Give important data/feature

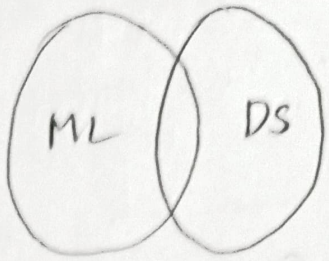
* Revise Pandas, Matplotlib

11/8/25 Data Science

→ Collection

→ Management

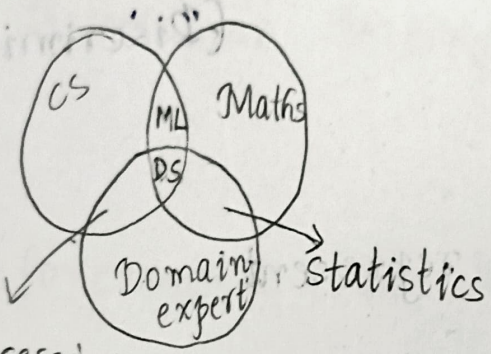
→ Analysis



Domain experts collect the data.

Management - Store the data - cs experts

Analysis - Maths & statistics

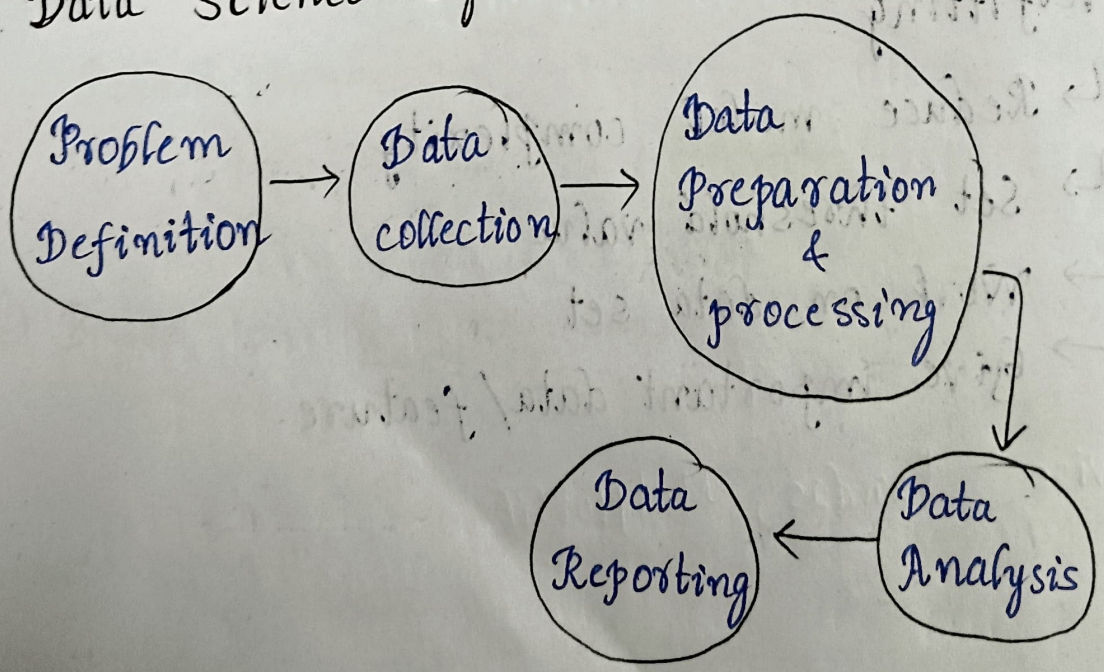


Preprocessing & Visualisation

Descriptive Method — Mean, Median, Mode

Probabilistic method — Gaussian distrib^n

Data Science Cycle



(i) We should have clear-cut objective } (1)

1/8/26

→ Experts
→ Colleagues } ①

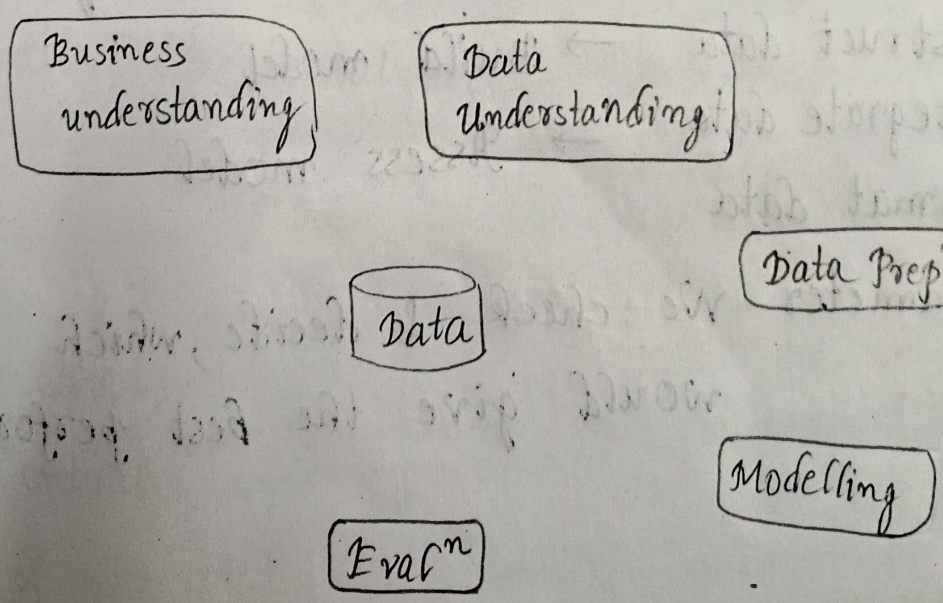
Where we can collect data?
→ Search engine
→ Web scraping
→ Surveys
→ Open ended interviews } ②

Missing values
Outliers } ③

→ statistical techniques
→ Probability
→ ANOVA
→ t-
p-test
→ Sample } ④

Data Science Market

① CRISP DM process ← (CRISP-DM Guide)



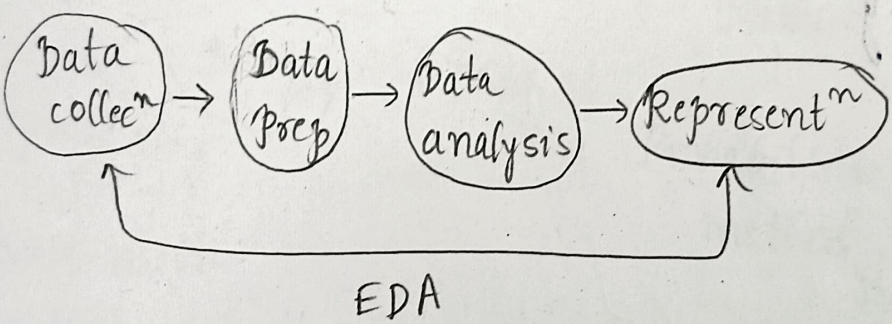
1/8/25

① Business Understanding

→ Determine business objectives
→ Assess situation
→ Determine data mining goals
→ Produce project plan

② Data Understanding

→ Collect initial data
→ Describe data
→ Explore data
→ Verify data quality



③ Data Prep^n

→ Select data
→ Clean data
→ Construct data
→ Integrate data
→ Format data

④ Modeling

→ Select modeling techniques
→ Generate test design
→ Build model
→ Assess model

Hyperparameter - We check & decide, which would give the best performance

Sales

389950, 230,500, 158000, 479,000, 639,000,

114,950, 55,00,000, 387000, 659,000, 529,000,

575,000, 488,800, 1095000

114950, 58000, 230500, 387000, 389950, 479000,

488800, 529000, 575000, 639000, 659000, 1095000,

6500000

$Q_2 = 488800$

$Q_1 = \frac{230500 + 387000}{2} = 308750$

$Q_3 = \frac{639000 + 659000}{2} = 649000$

$IQR = 649000 - 308750 = 340250$

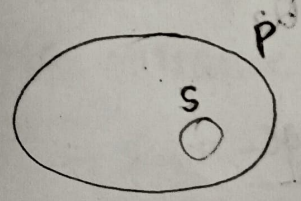
$(1.5)(IQR) = (1.5)(340250) = 510375$

Lower bound $= Q_1 - (1.5)(IQR) = -$

(Written by me)

Why n-1? (in sample variance)

In the case of skewed data, experiment^n have shown that, having n-1 gives better results. That is approximately equal to popu^n variance σ^2.



With the help of sample variance, we should estimate the popu^n variance

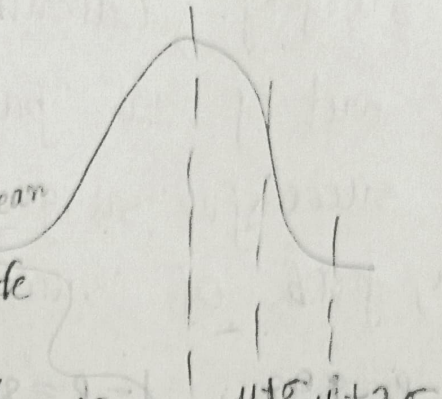
$s^2 \approx \sigma^2$

Mam: Degree of freedom concept

6/8/25

Z-score

$$\text{z-score} = \frac{x-\mu}{\sigma}$$



z can be
- −ve → Below mean, left side
- 0 → eq
- +ve → greater than mean μ

Suppose the x value of house price is 270000, Mean is 350000, std is 40000, then calculate the z-score

$$z = \frac{270000 - 350000}{40000} = -\frac{\cancel{8000}^{4}}{\cancel{40000}_{5}} = \cancel{\frac{4}{5}} -\frac{1}{5}$$

Probability Distribution

Binomial Distribution
↳ two success, failure
→ Independent trails, prob should not be changed
→ x → Discrete value

Discrete — Binomial, Poisson
Continuous — Normal, Gaussian

$$\boxed{P(x) = \binom{n}{x} p^x (1-p)^{n-x}}$$

A sample of 20 patients is selected for

6/8/26 surgery, Calculate the prob that 18 out of 20 patients will be having the successful surgery. From the past data, prob of success is 92%.

Sol):- $p = 92$, $1-p = 8$

$$P(x) = {}^{20}C_{18} \left(\frac{92}{100}\right)^{18} \left(\frac{8}{100}\right)^{2}$$

$$= \frac{20 \times 19}{1 \times 2} \left(\frac{92}{100}\right)^{18} \left(\frac{8}{100}\right)^{2}$$

$$= 190 \left(\frac{92}{100}\right)^{18} \left(\frac{8}{100}\right)^{2}$$

$=$

Poisson Distribution

Each n every thing will be dependent on whole dataset

μ, σ

$$P(x) = \frac{\mu^{x} e^{-\mu}}{x!}$$

Q) From the past data, a traffic engineer demands the avg. no. of vehicles entering a biking garrage, a 7 per 10 min period. Calculate the prob that the no of the vehicles entering in a parking is 9 in certain 10 min period.

6/8/20

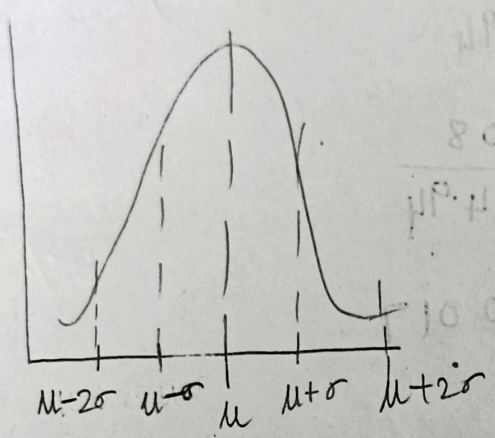
~~Ans~~ $e = 2.718$ Ans 0.101

$$P(x) = \frac{(7)^9 (2.718)^{-7}}{9!}$$

$=$

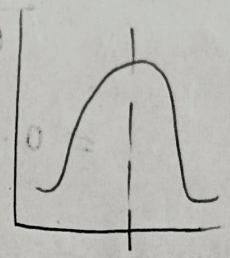
PMF - Probability Mass fn

Normal Distribution



$\mu = 0$

$\sigma = 1$



Mean
Median
Mode } overlap

$x \sim N(x, y)$

Mean Std

$$P(x) = \frac{1}{\sqrt{2\pi} \cdot \sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Each n every normal distribution will be standardized by z-score.

Q) At a software company, the mean of employe's salary is 60000 with a std dev 7500, Calculate the prob that a random employee earns more than 68000.

6/8/23

$$P(68000) = \frac{1}{\sqrt{2\pi}\cdot \cancel{(68000)}\,7500} e^{-\frac{(68000-60000)^2}{2(7500)^2}}$$

$$= \frac{1}{18794.94}(2.718)^{-\frac{64\times10^6}{5625\times10^4}}$$

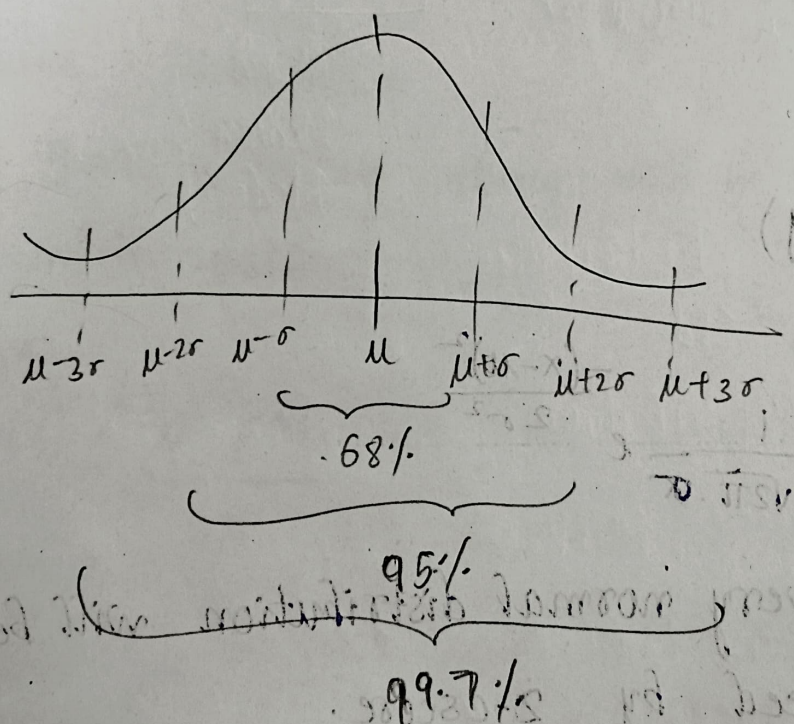
$$= \frac{1}{18794.94}(2.718)^{-\frac{6400}{5625}}$$

$$= \frac{1}{18794.94}(2.718)^{-1.137}$$

$$= \frac{0.3208}{18794.94}$$

$$= 0.000017$$

Ans: 14%



Q) Suppose an automated designer is interested in designing automatic ~~seats~~ seats to accomodate different heights of 95%. The heights of customers follow normal distribution with the mean value 68 &

6/8/25 Std 3. For wt range of heights, should the designer model the car seats

NEXT CLASS: CONFIDENCE LEVEL, ANOVA, HYPOTHESIS TESTING

Inferential statistics

We take some observ^n from sample & use it for popul^n

→ We find out error also

Confidence interval → Margin of error

Any statistic → Point of estimate

poe ± e

Confidence Interval

↑ confidence level

Lower bound of CI = poe − ME

Upper bound of CI = Poe + ME

A sample of 100 patients is taken. Determined at 64 year.

= 64 − 4 → 60 to 68

= 64 + 4

Mean is going to lie btw CI 60–68

Mean → Assumption, μ → popul^n mean

Proportion

μ_x → sample mean

$\mu_x = \mu$

$S_x = \frac{\sigma}{\sqrt{n}}$ n → sample size

Central Limit Theorem → When we are

13/8/25 going to aggregate the samples, it will form normal distribution.

We might get skewed graph.

2 cond^{ns}: ① We don't know dataset distrib^n

We know mean, std

If sample size > 30 ! for a m shape

② std is not known

CL → Mean — σ known → z-test (z-score)

— σ Not known → t-test

σ is known

Confidence level for mean $= z_c \frac{\sigma}{\sqrt{n}}$

Lower B of CI → Mean − E_c

Upper B of CI → Mean + E_c

Q) Sample size = 50, POE = Popul^n mean = 85000

Std = 9000. Distrib^n isn't known. Calculate the mean & std for sample & comment the shape of distrib^n for sample.

Sol)

50
85000
90000
Sample

} Popul^n

Sample std dev $S_x = \frac{\sigma}{\sqrt{x}}$

$= \frac{9000}{\sqrt{50}}$

$= \frac{9000}{5\sqrt{2}}$

13/8/25 * According to CLT, if sample size > 30, then it follows normal distribution.

Find CI for 95% confidence, $z = 1.96$

error of margin $CL = 1.96 \times \frac{9000}{\sqrt{50}} = \frac{196}{100} \times \frac{9000}{5\sqrt{2}} = 98\sqrt{2} \times 18$

$= 1764\sqrt{2}$

$\cancel{= 730.2}$

$= 2494.66$

Q) Suppose sample of 50 students is taken, in that 50 students, the mean of time spent on HW is 12.5 hrs. The popⁿ std devⁿ is 6.3 hrs. Create a forecasted CI using CL 90% & 95% and provide a conclusion regarding these confidence level that which interval is wider.

z-score for 90% = 1.645, 95% = 1.960

Sol):- SS = 50 ⇒ Normal distribⁿ

Mean = 12.5, $\sigma = 6.3$

CI Mean (90%) $= (1.645) \times \frac{6.3}{\sqrt{50}} = 1.4656$

CI Mean (95%) $= (1.960) \times \frac{6.3}{\sqrt{50}} = 1.7462$

(90%) Lower B = 12.5 ± 1.465

(95%) Upper B = 12.5 ± 1.746 → Wider

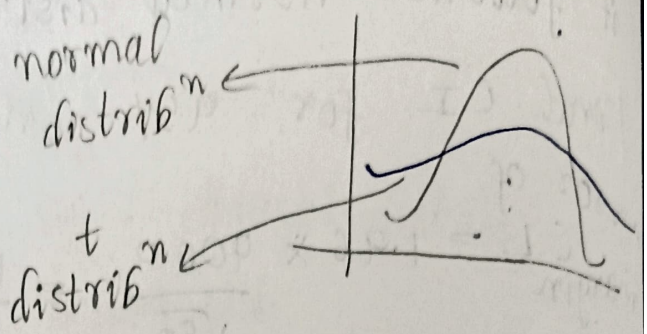
→ Random Sample p → If σ is not known, find out S_x

→ 30 or ND

→ σ

13/8/25

$$E_c = t_c \frac{S_x}{\sqrt{n}}$$



degree of freedom $= n-1$

→ Check from table

DOF	98%	95%	...
1			
2			
3			
⋮			

Q) Suppose in a company 5000 employees r there. Administr^n decided to take sample of 16 employee to find mean salary. The sample data indicate, sample mean of salary is 15.8 lakhs & the std is 3.2 lakh. Calculate the 99% ~~CL~~, CI for salary.

$t_c = 2.947$

Sol):- $E_c = (2.947)\frac{(3.2)}{\sqrt{16}} = 2.3576$

$CI = 15.8 \pm 2.357$

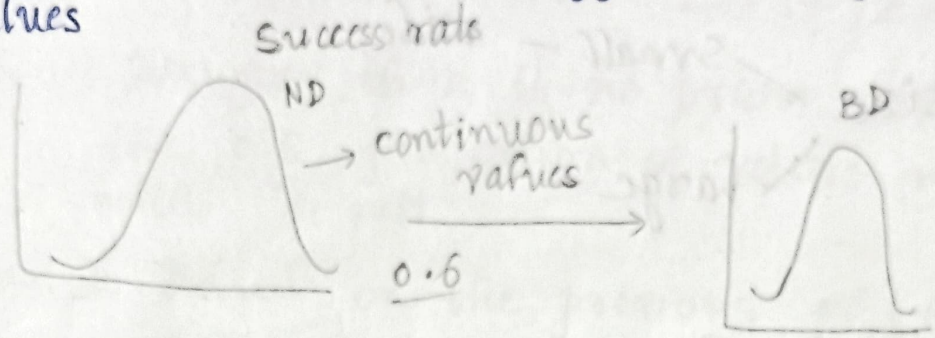
$= [\cancel{18.14}, 13.44, 18.16]$

13/8/25

Confidence Interval for Proportions

$$\hat{p} = \frac{x}{n}$$

Normal distrib^n is applicable for continuous values



$$\begin{bmatrix} np \\ n(1-p) \end{bmatrix} \rightarrow \text{at least } 5$$

If it is greater than 5, Normal distrib^n can represent binomial distrib^n also.

$$E_c = z_c \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Q) A medical researcher wants to see the change in the propor^n of smokers from 5 yrs ago & the " of adult " was 28%. A random sample of 1500 adults is taken & 360 were found smokers. Calculate the 95% confidence interval for the true pop^n portion of adults who smoke. $z_c = 1.960$

Sol:- $\hat{p} = \frac{360}{1500}$

$$E_c = (1.960)\sqrt{\frac{0.24(1-0.24)}{1500}}$$

$$= 0.0005$$

$\hat{p} \pm E_c$

$n\hat{p}$, $n(1-\hat{p})$ should be atleast 5

13/8/23

Q) How can we determine the sample size if we take ~~sam~~ small sample?

Sample size
- Small – wide
- Large –

Mean $E_c = z_c \frac{\sigma}{\sqrt{n}}$

$$n = \left(\frac{z_c \sigma}{E_c}\right)^2$$

~~$n = \left(\frac{z_c}{E_c}\right)^2 (1-p)$~~

~~$n = z$~~

Proportion $E_c = z_c \sqrt{\frac{(1-p)\cdot p}{n}}$

$$n = \left(\frac{z_c}{E_c}\right)^2 (1-\hat{p})\hat{p}$$

Q) A Benefits analyst ~~is~~ ~~[illegible]~~ interested in 90% CL for _____ for mean salary of employee. What should be the ~~[illegible]~~ Sample ~~[illegible]~~ size if margin of error 1000 is desired. σ(popu^n) = 8000

– (z_c = 1.645)

$$\Rightarrow n = \left(\frac{1.645 \times 8000}{1000}\right)^2 = \left(\frac{13160}{1000}\right)^2 = (13.16)^2$$

$= 173.18$

$\simeq 174$ / whole number will be taken

~~38/23 2Q) One candidate is smith is~~

Q) 1 candidate is smith is planning a survey to determine 95% C.I for the proportion of voters who planned to vote for smith. How many should be survived. $E_c = 3\%$, $z = 1.96$

Case-1: Assume there is no prior estimate for the 0.5 proportion of voters who vote for smith
↓ default

Case-2: Based on the previous elections, 42% ppl voted for smith

(I) $E_c = 3\%$, $z = 1.96$

95% CI

$$n = \left(\frac{1.96}{0.03}\right)^2 (1-0.96)(0.96)$$

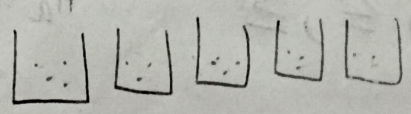
$= 1043$

Bootstrap sampling

Mean → 30, ND

Proportion → np, $n(1-p)$ → 8

Bootstrap → sample with replacement



$^6C_1 \times {}^4C_1$

↳ Sample without replacement

Wt is the advantage of bootstrap sampling?

(i) To increase popu^n size

(ii) Help us use valid^n

(iii) Helps remove bias in sample

19/8/

Stratified sampling:

If dataset contains 60% class A & 40% class B,

then sample should also contain "

① Population → repeatedly generate sample.

② $\bar{x}$, $\bar{s}_x$

③ Sort

④ 95%, $P_{2.5}$ $P_{92.5}$

⑤ 90%, P_5, P_{95}

Hypothesis Testing

We were defining statistical parameter & we were mapping it with pop^n "
(Previously)

Null H_0 → $=, \leq, \geq$

H_0 $\mu = 25\%$
H_a $\mu \neq 25\%$

Alternate H_a → $\neq, >, <$

After claiming, we need to apply some statiscal approach.

σ → z-test
→ t-test

Then, we'll set some level of significance

$Z \leq P$ (or) $Z > P$

19/8/25

(Predic^n)

Actual	Yes	No
YES	TP	FN
NO	FP	TN

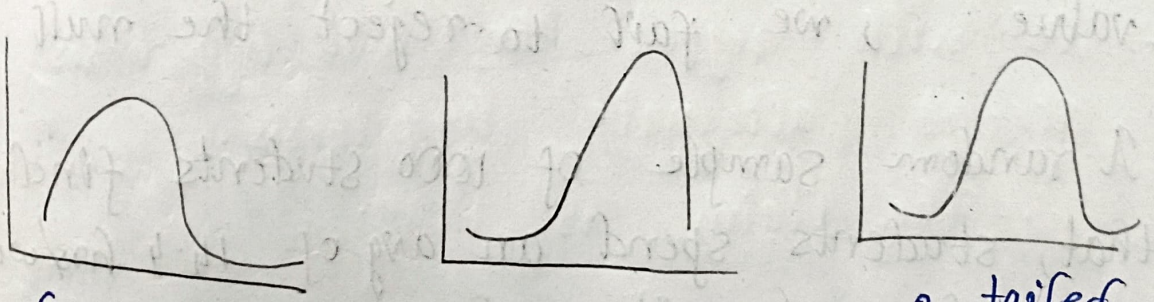
→ Type 2 error

↓ Type 1 error

If my null hypothesis is true & I'm failing to reject that one – Type 1

bcoz actually it is true but we r failing.

If my null hypothesis is false & we fail, – Type 2



left tailed right-tailed 2-tailed

$<$ $>$ $=$

σ known

mean $z = \dfrac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$

σ not known

mean $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$

propor^n In any sample, how many ← are there

$\hat{p} = \dfrac{x}{n}$ → $n\hat{p}$ & $n(\hat{p}-1)$ at least 5.

$z = \dfrac{\hat{p} - p}{\sqrt{p(1-p)/n}}$

p → claim propor^n of hypothesis

10/8/25

Q) An auto repair facility claims the avg time for oil-change is less than 20 mins.

atleast, atmost → null

less than, greater → alternate

Null $\mu \geq 20$

Alt. $\mu < 20$

Q) A medical researcher claims that the proportion of adults in the United States who r smokers is atmost 25%.

Null $\mu \leq 25\%$

Alt $\mu > 25\%$

* If $\hat{p}$ value is less than or equal to significance value, we reject
* If $\hat{p}$ value is greater " value, we fail to reject the null

Q) A random sample of 1000 students find that, students spend an avg of 14.4 hrs/week on social media. Use this sample data to test the claim that clg student spend atleast 15 hrs/week on social media. Level of Significance = 0.05, Popn of std devn = 3.25

Sol:- Null $\mu \geq 15$

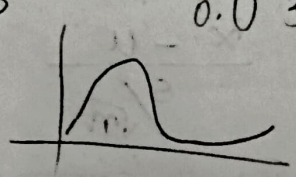
$\mu < 15$

$$Z = \frac{14.4 - 15}{\frac{3.25}{\sqrt{1000}}}$$

-1.83

0.05

$0.0324 < 0.05$



19/8/25 We're going to reject the null hypothesis.

→ The Claim will be on sample, that is going to be reflected in popuc^n.

Based on t-test

Q) A smart phone manufacturer claims that the mean battery life of the phone is 25 hrs. Consumer & collect the sample of 50 S.P. and determine the mean value 24.1 hr & std dev 4.1 hr. Use this sample data to test the claim made by smart phone manufacturer. Use the level of significance, 0.10.

Null $\mu = 25$

Alt $\mu \neq 25$

$$t = \frac{50 - \cancel{25}\ 24.1}{\frac{4.1}{\sqrt{50}}} = \frac{25.9}{}$$

$s/\sqrt{n} = -1.552$

↳ 0.0635

2 tails ⇒ double value

0.127

→ We r failing to reject the null hypothesis i.e, we r accepting.

0.127 > 0.1

So, wt the manufacturer is claiming is true

Based on Propor^n

Q) A Clg professor claims that the propor^n of students using the Chatgpt in their

19/8/25 assignment is less than 45%. Professor selects sample of 200 students & figure it out that 74 out of 200 use gpt. Use LOS = 0.05

Sol:-

Null $\mu \geq 0.45$ $\quad p = 0.45$

Alt $\mu < 0.45$ $\quad \hat{p} = 74$

$$z = \frac{74 - 0.45}{\sqrt{0.45(1-0.45)/200}} =$$

first, $n\hat{p} = 200 \times 74 \geq 5$

$n(\hat{p}-1) = 200(74-1) \geq 5$

$z = -2.274$

$z \hookrightarrow$ p-value $= 0.012 < 0.05$

→ We will reject the null hypothesis, accept alternate hypothesis.

∴ The claim is true.

Hypothesis Testing for two samples

μ_1, μ_2 – mean of sample 1 & sample 2

Null $\mu_1 = \mu_2$ $\quad \mu_1 \geq \mu_2$ $\quad \mu_1 \leq \mu_2$

Alt $\mu_1 \neq \mu_2$ $\quad \mu_1 < \mu_2$ $\quad \mu_1 > \mu_2$

Samples can be dependent or independent

Independent

→ Random & Indep^t

20/8/26

→ Sample should be of size 30 or follow normal distribun

→ sample variance or SD

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

Q) Suppose a SE is claiming that the avg salary of women engineer is less than male. Two samples are taken, sample 1 = 45, sample 2 = 50, LOS = 0.05

	WE	ME
~~Sim~~		
Sample mean	72700	76900
SD	9800	10700
Size	45	50

WE < ME ⇒ Null : $\mu_1 \geq \mu_2$

μ_1 μ_2 At $\mu_1 < \mu_2$

If there is no difference, then there is no change in claim. So, there is = always in null hypothesis.

$\therefore \mu_1 = \mu_2 \Rightarrow (\mu_1 - \mu_2) = 0$

$t =$ (

$= -1.997 \rightarrow 0.025 < 0.05$

20/8/25

∴ The Claim is true

Dependent samples:

difference (d) - before data - after data

$$\bar{d} = \frac{\Sigma d}{n}$$

$$t = \frac{\bar{d} - \mu_d}{s_d/\sqrt{n}}$$

→ when drug is taken, what is the bp & when drug is not taken, what is bp?

μ_d, k constant

Null $\mu_d = 0$ or k, $\mu_d \geq k$

A $\mu_d \neq 0$ or k, $\mu_d < k$

$\mu_d \leq k$

$\mu_d > k$

Q) Cholesterol lowering medicine.

Claim - Medic^n helps to lower cholestrol level., LOS = 0.05

Null $\mu_d <= 0$ Claim - cholestrol is lower

Alt $\mu_d > 0$ ∴ before - after > 0

But if claim is cholestrol is higher then before - after < 0

20/8/05

Patient	BM	AM
1	218	210
2	232	241
3	259	223
4	266	244
5	248	227
6	298	273
7	263	252
8	281	276
9	290	281
10	271	259

$\bar{d} = \frac{\Sigma d}{n} = 13.9 \rightarrow$ On avg

$S_d = 12.414$

$$t = \frac{13.9 - 0}{\frac{12.414}{\sqrt{10}}} = 3.541$$

$\hookrightarrow 0.003$

$0.003 < 0.05$

Two proportion

$x_1 \quad x_2$

$n_1 \quad n_2$

$$\hat{p}_1 = \frac{x_1}{n_1}, \quad \hat{p}_2 = \frac{x_2}{n_2}$$

$$\bar{p} = \frac{x_1 + x_2}{n_1 + n_2}$$

$n_1\bar{p}, n_2\bar{p}, n_1(1-\bar{p}), n_2(1-\bar{p})$ should be atleast 5

$$z = \frac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

→ Based on our hypothesis

Q) Investigⁿ to address lung cancer.
People who took drug in 211 was cancer free among 250, 172 among 230
LOS = 0.01,
Claim: Proportion of ~~subjects~~ candidates $p_1 > p_2$ taking cancer-free drug are more cancerfree than the candidates ~~to~~ who r not taking drug.

Sol:- $n_1 \cdot \bar{p} =$

$$\hat{p}_1 = \frac{211}{250}, \quad \hat{p}_2 = \frac{172}{230}$$

$$\bar{p} = \frac{211+172}{250+230} = 0.798$$

$$n_1 \bar{p} = 250 \times 0.798 = 199.5$$

$z = 2.6$

$0.004 < 0.01$

ANOVA test — random Independent.

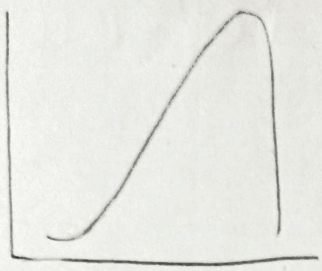
Analysis btw the variance —

→ z, t → one sample

→ t → two sample < D, I

22/8/25

F → distribution



mean → prop

① $\bar{x}_1, \bar{x}_2, \bar{x}_3$ & s_1^2, s_2^2, s_3^2

② Grand mean $\bar{\bar{x}}$

③ $SSB = n_1(\bar{x}_1 - \bar{\bar{x}})^2 + n_2(x_2 - \bar{\bar{x}})^2 + n_3(\bar{x}_3 - \bar{\bar{x}})^2$

④ $SSW = (n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + (n_3 - 1)s_3^2$

⑤ df_n = no of groups − 1

df_d = total ss − no of group

⑥ $MSB = \frac{SSB}{df_n}$ ⑦ $MSW = \frac{SSW}{df_d}$

⑦ $F = \frac{MSB}{MSW}$

<u>two degree</u>

→ no. of group

→ total no samples − no. of g

A	B	C
5	5	5
$\bar{x}_1$	$\bar{x}_2$	$\bar{x}_3$
s_1^2	s_2^2	s_3^2

F < level of

P-value < level of sig

22/8/23

S.NO	A	B	C	D
1	8	12	18	13
2	10	11	12	9
3	12	9	16	12
4	8	14	6	16
5	7	4	8	15

Claim: Height of pop^n > Mean value of pop^n ×

Mean(A) $= \frac{45}{5} = 9$ Var(A) $= \frac{16}{4} = 4$ Std(A) $= 2$

Mean(B) $= \frac{50}{5} = 10$ Var(B) $= \frac{58}{4} = 14.5$ Std(B) $= 3.8$

Mean(C) $= \frac{60}{5} = 12$ Var(C) $= \frac{104}{4} = 26$ Std(C) $= 5.099$

Mean(D) $= \frac{65}{5} = 13$ Var(D) $= \frac{30}{4} = 7.5$ Std(D) $= 2.73$

② Grand mean $\bar{\bar{x}} = \frac{9+10+12+13}{4} = 11$

③ $SSB = 5(9-11)^2 + 5(10-11)^2 + 5(12-11)^2 + 5(13-11)^2$

$= 20 + 5 + 5 + 20$

$= 50$

④ $SSW = (5-1)4 + (5-1)14.5 + (5-1)26 + (5-1)7.56$

$= 208$

⑤ $df_n = 3$

$df_d = 20 - 3 = 17$

Ans

⑥ $MSB = \frac{50}{3}$

$= 16.66$

⑦ $MSW = \frac{208}{17}$

$= 12.23$

Corelation Analysis → Regression
~~Represent~~

Whether the features are going to have impact on output or not.

+ve → ↑ indep ↑ dep

0 → X

-ve → ↑ inde ↓ dep

Pearson correlation $= \frac{n \Sigma xy - (\Sigma x)(\Sigma y)}{\sqrt{n \Sigma x^2 - (\Sigma x)^2} \sqrt{x \Sigma y^2 - (\Sigma y)^2}}$

r

value → strength

sign → direction

ρ → Popur^n P.C Coefficient

Linear relationship ⇒ $|r| > \frac{2}{\sqrt{n}}$

Q) Suppose the chief financial officer is investigating the correl^n between rest of price

In a sample data size, 10 points r taken

Now, Based on this analysis can chief financial

2218hrs officer conclude that the correlⁿ is significant b/w the stock price & unemployment rate with the LOS = 0.06.

Soln: $r = -0.68$

$n = 10$

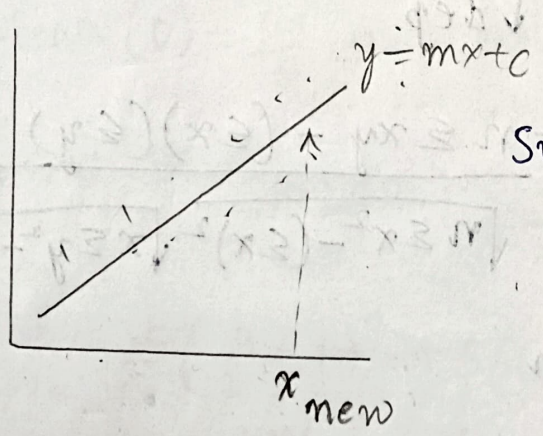
$|-0.68| > \frac{2}{\sqrt{10}}$

$0.68 > 0.63$

Yes, the relⁿship is there.

They're going to be -vely correlated

Linear Regression



Supervised ML — Classificⁿ (Discrete) CM, P, R, A

— Regression (Continuous) ↳ Loss/cost

$\hat{y} = w_1 f_1 + w_2 f_2 + w_3 f_3 + w_4 f_4 + c$

w_1 f_1	w_2 f_2	w_3 f_3	w_4 f_4	y

$c = w_0 f_0$

$f_0 = 1$

$\hat{y} = w_0 f_0 + w_1 f_1 + \dots + w_n f_n$

$\hat{y} = WF$ — $\hat{y} = WX$

→ $y = ax + b$

→ Closed Normal form $\frac{\partial \text{Loss}}{\partial w}$

→ Gradient descent → η

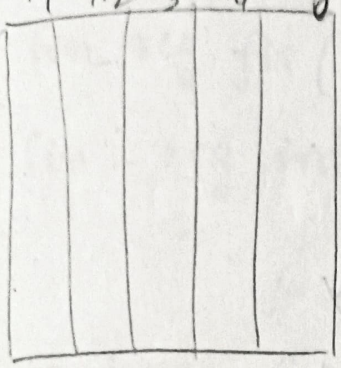
20/8/25

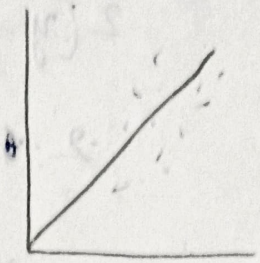
Linear Regression

$y = mx + c$

$w_1 \; w_2 \; w_3 \; w_4$

$x_1 \; x_2 \; x_3 \; x_4 \quad y$





Loss/ Cost/ Error/ Bias

We will minimize

$y - \hat{y}$

$\hat{y} = w_1 x_1 + w_2 x_2 + w_3 x_3 + w_4 x_4 + c$

$\downarrow$

$w_0 x_0$

where $x_0 = 1$

$MAE = \frac{1}{n} |y - \hat{y}|$

$MSE = \frac{1}{n} (y - \hat{y})^2$

$\hat{y} = WX$

$W = [w_0 \ldots w_n]$

$RMSE = \sqrt{\frac{1}{n} (y - \hat{y})^2}$

$X = [x_0 \ldots x_n]$

where $x_0 = 1$

Cross entropy

$\hookrightarrow \hat{y}^{y} (1 - \hat{y})(1 - y)$

LR — Closed normal form

— Gradient descent

$\hat{y} = WX$

$Loss = \frac{1}{n} (y - \hat{y})^2$

$\frac{\partial Loss}{\partial w} = \frac{1}{n} [2 \times (y - \hat{y})] \times \frac{\partial}{\partial w} (y - \hat{y})$

$\Rightarrow 2X (y - \hat{y}) * \frac{\partial}{\partial w} (y - Xw)$

$\Rightarrow$ Equate it to 0.

26/8/25

$$2(y-\hat{y}) * \frac{\partial}{\partial w}(y - Xw) = 0$$

$$2(y - Xw) * (-X) = 0$$

$$2 \cdot X^T(y - Xw) = 0$$

$$X^T(y - Xw) = 0$$

$$X^Ty - X^TXw = 0$$

$$X^TXw = X^Ty$$

$$w = (X^TX)^{-1}X^Ty$$

Code

```
X = 2* np.random.rand(m,1)
y = 4 + 3*X + np.random.randn(m,1)
```

normal distributed values

```
from sklearn.preprocessing
import add_dummy_feature
X_b =
```

a.T → a^T

```
theta_best = np.linalg.inv(X_b.T.dot(X_b)).
theta_best
```

dot (X_b.T).dot(y)

```
X_new = np.array([[0], [2]])
print (X_new)
X_new_b = add_dummy_feature(X_new)
y_predict = X_new_b.dot (theta_best)
y_predict
```

26/8/25

Using SKlearn:

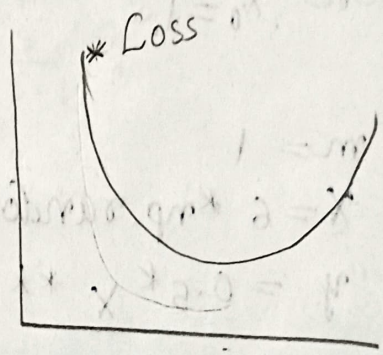
```
from linear_model import LinearRegression
lin_reg = LinearRegression()
lin_reg.fit(X, y)
lin_reg.intercept_, lin_reg.coef_
```

$\downarrow$ ω_0 $\qquad$ $\downarrow$ ω_1

Gradient descent

$$\omega_{i+1} = \omega_i - \eta \frac{\partial}{\partial \omega} \text{Loss}$$

$\downarrow$ learning rate

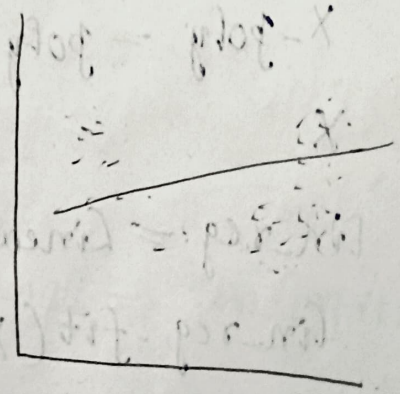


In the starting, take some random values for ω.

Curve should be convex curve

GD
- Batch G-D - Whole dataset is given in 1go
- Stochastic G-D - Any 1 random instance
- Mini-batch

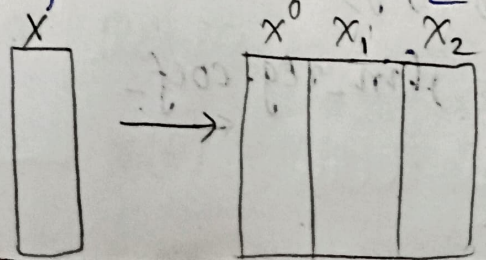
Underfitting ←



Polynomial Regression

x, $d = 2$

x^0, x^1, x^2 $\qquad$ $\omega_0\ \omega_1\ \omega_2$



x, y ; degree - 2

$1, x, y, xy, x^2, y^2$

$\omega_1, \omega_2, \omega_3$ ...

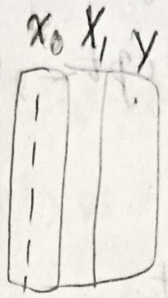
26/8/23 degree - d , num of features - n

→ How many new features will come?

$$^{n+d}C_d = \frac{(n+d)!}{n! \times d!}$$

$y = WX$

$= w_0 + w_1 x$ ⇒



$w_0 x_0 + w_1 x$

where $x_0 = 1$

Code

```
m = 1
X = 6 *np.random.rand(m,1) -3
y = 0.5* X **2 + X + 2 + np.random.randn(m,1)
```

for finding polynomial features:

```
from
        import PolynomialFeatures
poly_features = PolynomialFeatures
                (degree=2, include_bias=False)
X_poly = poly_features.fit_transform(X)
X

lin_reg = LinearRegression()
lin_reg.fit(X_poly, y)
lin_reg.intercept_, lin_reg.coef_
```

09/8/25

Regularization

$$\text{Cost} = \frac{1}{n}(y-\hat{y})^2 + \underline{\quad}$$

Minimize

Regularizⁿ — Ridge (L2 Norm)
— Lasso (L1 Norm)
— Elastic Net

1) Ridge $J(\theta) = MSE + \alpha \sum_{i=1}^{m} \theta_i^2$

$$\text{Cost} = \frac{1}{n}(y-\hat{y})^2 + \alpha \sum_{i=1}^{n} \theta_i^2$$

$$= \frac{1}{n}(y-\hat{y})^2 + \alpha \cdot \theta^2$$

$$\frac{\partial \text{Cost}}{\partial \theta} = \frac{1}{n} \times 2 \times (y-\hat{y}) \times \frac{\partial}{\partial \theta}(y-\hat{y}) + \alpha \times 2\theta$$

$$= \frac{2}{n}(Y - X\theta) \times \frac{\partial}{\partial \theta}(Y - X\theta) + 2\alpha\theta$$

$$= \frac{2}{n}(Y - X\theta) \times (-X) + 2\alpha\theta = 0$$

$$= -XY + X^T X\theta + \alpha\theta = 0$$

$$= \theta = \frac{XY}{X^T X + \alpha}$$

$\alpha \rightarrow$ large $\Rightarrow \theta \rightarrow$ very small $\Rightarrow$ Underfitting

2) Lasso Regularizⁿ

$$J(\theta) = MSE + \alpha \sum_{i=1}^{n} |\theta_i|$$

29/8/23

$$\text{cost} = \frac{1}{n}[y-\hat{y}]^2 + \alpha\theta$$

$$\frac{\partial}{\partial\theta}\text{cost} = \frac{1}{n} \times 2\,[y-\hat{y}] \times \frac{\partial}{\partial\theta}[y-\hat{y}] + \alpha$$

$$= \frac{2}{n}[y - X\theta] \times \frac{\partial}{\partial\theta}[y - X\theta] + \alpha$$

$$= \frac{2}{n}[y - X\theta] \times (-X) + \alpha = 0$$

$$-\frac{2}{n}XY + \frac{2}{n}X^T X\theta + \alpha = 0$$

$$\theta = \frac{\frac{2}{n}XY - \alpha}{\frac{2}{n}X^T X}$$

3) Elastic Net

$$J(\theta) = MSE + r\alpha \underbrace{\sum_{r=1}^{m} \theta_i}_{L_1} + \frac{1-r}{2}\alpha \underbrace{\sum_{r=1}^{n} \theta_i^2}_{L2}$$

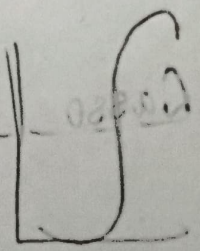
Lass Ridge

We r adding something to MSE, so that our model will work in real-time scenario

Logistic Regression

→ Classification

Apply sigmoid fn 0-1 $\Rightarrow \frac{1}{1+e^{-t}}$



0.5 < N Negative Class

> P Positive Class

Softmax — multi-class classific^n

$$\frac{e^{zx}}{\sum_{i=1}^{n} e_i^{zx}}$$

Classification

→ Binary — Yes — ✓, ✗
— No — ✓, ✗

Predicted

Actual	Yes	No
Yes	TP	FN → Type II
No	FP	TN

↓ Type I

Precision → $\frac{TP}{TP+FP}$

Recall (Completeness) → $\frac{TP}{TP+FN}$

Accuracy → $\frac{TP+TN}{TP+TN+FP+FN}$

We need a mid-way of Precision & Recall → F1-score → $\frac{2 \times Precis^n \times Recall}{Precis^n + Recall}$

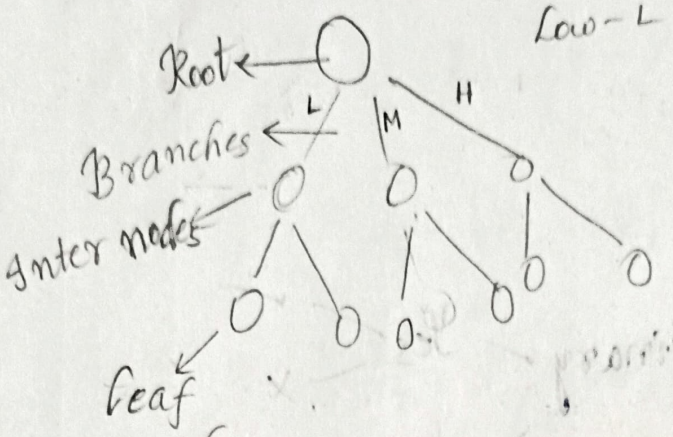
Underfitting → Training ✗ Testing ✗

Overfitting → Training ✓ Testing ✗

$$\text{Loss}_{error} = Bias^2 + Variance + Irreducible\ Noise$$

2/9/23 Decision tree

Low-L Medium-M High-H



2/3/B

* On leaf node, we'll be having final class

→ It can work on n-dimensional dataset
→ Prior info is not required
→ Domain-expert "

} Advantages

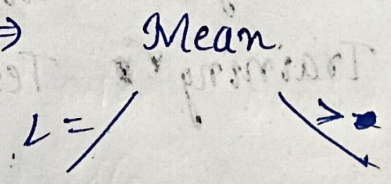
→ Overfitting - Because it's a rule-based algorithm
No learning

} Disadvantage

How are we going to split & Attribute selection?

Discrete ⇒ L/M/H or Yes/No

Continuous ⇒ Mean



[value <= Mean]

Attribute selection — Entropy or Info^n. Gain → ID3

Gain Ratio → C4.5

Index ~~Root~~ → CART

Entropy (or) IG

→ Randomness → Measure that quantifies the avg. amount of inform^n or uncertainity.

$$IG\,E(Class) = \cancel{-\sum}$$

$$= -\sum P_i \log_2 P_i$$

→ how the class is going to be distributed

0 0 1 1 2 0 3 4

$\frac{3}{8}$ $\frac{2}{8}$ $\frac{1}{8}$

$$IG_A(Class) = \sum_{i=1}^{N} \frac{|D_i|}{|D|} \times IG(D_i)$$

Root

D_1 D_2

$$\Delta Gain_A(Class) = IG(Class) - IG_A(Class)$$

→ When all instances are saying, this is the class

→ If all instances & rows are fully utilized

Basic criteria for stopping DT:

① All instances are agreeing for 1 class

② All attributes are utilized

③ Subset that we are generating does not have any instance.

8/9/23

S.N	Age	Income	Student	Credit Rating	Buy-camp
1.	Y	H	N	F	N
2.	Y	H	N	E	N
3.	M	H	N	F	Y
4.	S	M	N	F	Y
5.	S	L	Y	F	Y
6.	S	L	Y	E	N
7.	M	L	Y	E	Y
8.	Y	M	N	F	N
9.	Y	L	Y	F	Y
10.	S	M	Y	F	Y
11.	Y	M	Y	E	Y
12.	M	M	N	E	Y
13.	M	H	Y	F	Y
14	S	M	N	E	N
		0.029	0.151	0.048	

Y-Youth, M-Middle Age, S-Senior Citizen

$$IG(Class) = -\sum_{i=1}^{n} P_i \log_2 P_i$$

$$IG_{Att}(Class) = \sum_{i=1}^{N} \frac{|D_i|}{|D|} IG(D_i)$$

3/9/23

$\Delta G_{(A)} = IG(\text{Class}) - IG_{Att}(\text{Class})$

$Y = 9/14$ $\quad IG(\text{Buy-Comp}) = -\frac{9}{14}\log_2\frac{9}{14}$

$N = 5/14$ $\quad -\frac{5}{14}\log_2\frac{5}{14}$

$= -\frac{9}{14}(-0.637) - \frac{5}{14}(-1.485)$

$= 0.4095 + 0.5303$

$\cancel{= 1.449}$

$= 0.939$

≈ 0.94

$\log_a b = \frac{\log b}{\log a}$

$\log_2 \frac{9}{14} = \frac{\log \frac{9}{14}}{\log 2}$

$IG.(\text{Class})_{Age} = \frac{5}{14} IG(D_1) + \frac{4}{14} IG(D_2) + \frac{5}{14} IG(D_3)$

Y 5 < 2 Y, 3 N

M 4 < 4 Y, 0 N

S 5 < 3 Y, 2 N

$= \frac{5}{14}\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}\right]$

$+ \frac{4}{14}\left[-\frac{4}{4}\log_2\frac{4}{4} - 0\right]$

$+ \frac{5}{14}\left[-\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5}\right]$

$= \frac{5}{14}\left[-\frac{2}{5}(-1.321) - \frac{3}{5}(-0.736)\right]$

$+ \frac{4}{14}\left[-1(0) - 0\right] + \frac{5}{14}\left[-\frac{3}{5}(-0.736) - \frac{2}{5}(-1.321)\right]$

$= \frac{5}{14}\left[0.5284 + 0.441\right] + \frac{5}{14}\left[0.4416 + 0.5284\right]$

$= \frac{5}{14} \times 0.9694 + \frac{5}{14} \times 0.9694 = 0.692$

≈ 0.694

30/25

$\Delta G = 0.94 - 0.694$

$= 0.246$

~~$IG(\text{Student}) =$~~ $IG_{student}(Class) = \frac{7}{14} IG(D_1) + \frac{7}{14} IG(D_2)$

Age

N - 7

Y - 7

Y — 6, 1

N — 3, 4

$= \frac{7}{14}\left[-\frac{6}{7}\log_2\frac{6}{7} - \frac{1}{7}\log_2\frac{1}{7}\right] + \frac{7}{14}\left[-\frac{3}{7}\log_2\frac{3}{7} - \frac{4}{7}\log_2\frac{4}{7}\right]$

$= \frac{7}{14}\left[0.191 + 0.401\right] + \frac{7}{14}\left[0.524 + 0.461\right]$

$= \frac{7}{14} \times 0.592 + \frac{7}{14} \times 0.985$

$= 0.296 + 0.493 = 0.789$

$\Delta G_{(Student)} = 0.94 - 0.789 = 0.151$

$IG_{Income}(Class) = \frac{4}{14} IG(D_1) + \frac{6}{14} IG(D_2) + \frac{4}{14} IG(D_3)$

L — 3, 1

M — 4, 2

H — 2, 2

$= \frac{4}{14}\left[-\frac{3}{4}\log_2\frac{3}{4} - \frac{1}{4}\log_2\frac{1}{4}\right] + \frac{6}{14}\left[-\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}\right] + \frac{4}{14}\left[-\frac{2}{4}\log_2\frac{2}{4} - \frac{2}{4}\log_2\frac{2}{4}\right]$

$= \frac{4}{14}\left[0.311 + 0.5\right] + \frac{6}{14}\left[0.390 + 0.528\right] + \frac{4}{14}\left[1\right]$

$= \frac{4}{14}(0.811) + \frac{6}{14}(0.918) + 0.286$

$= 0.232 + 0.393 + 0.286 = 0.911$

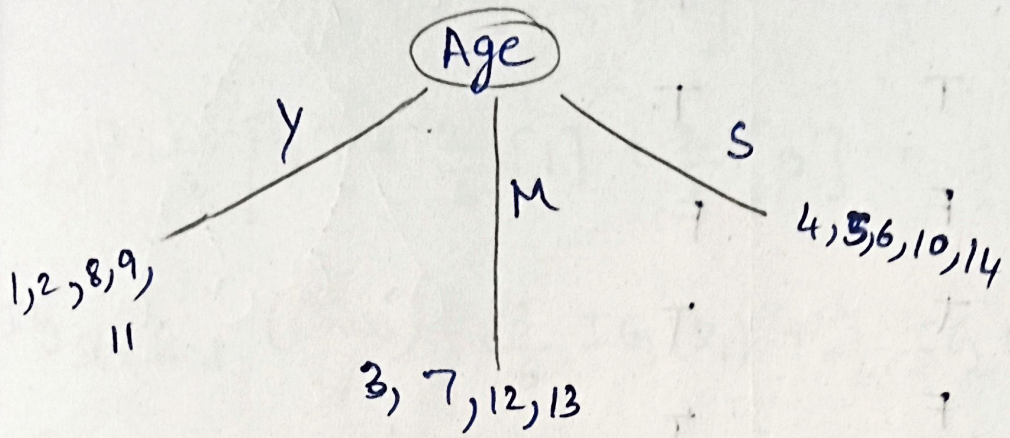
$\Delta G_{Income} = 0.94 - 0.911 = 0.029$

Similarly, $\Delta G_{Credit\ Rating} = 0.048$

13/9/25

* Age has highest - 0.246

So, it will be root node



$$IG(D_1) = -\frac{2}{5}\log_2\frac{2}{5} - \frac{3}{5}\log_2\frac{3}{5}$$

$$IG_{income}(D_1) = \frac{2}{5}IG(D_1^{H}) + \frac{2}{5}IG(D_1^{M}) + \frac{1}{5}IG(D_1^{L})$$

$IG_{student}(D_1)$

$IG_{CR}(D_1)$

$$= +\frac{2}{5}\left[-\frac{2}{5}\log_2\frac{2}{5} - \frac{2}{5}\log_2\frac{2}{5}\right.$$

H < 2, 2, 0

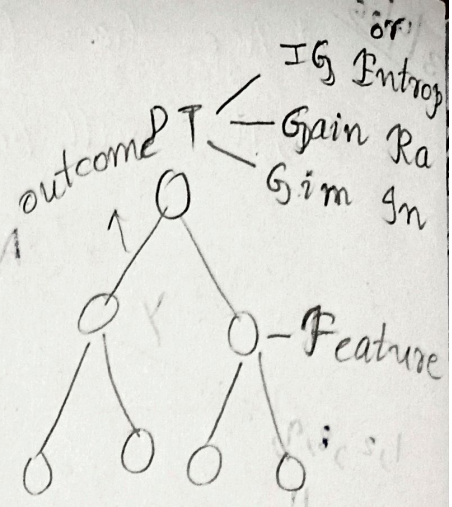
M < 2, 1, 3

L < 1, 1, 2

9/9/23

	GPA	Studied	Passed
1	L	F	F
2	L	T	T
3	M	F	F
4	M	T	T
5	H	F	T
6	H	T	T

Added by me, not feature



$$IG(Class) = -\sum_{i=1}^{n} P_i \log_2 P_i$$

$$IG_A(Class) = \sum_{i=1}^{n} \frac{|D_i|}{|D|} IG\left(\frac{D_i}{D}\right)$$

$$\Delta G = IG(Class) - IG_A(Class)$$

Stopping Criteria for DT:→ All instances belong to same class

→ No features are left

→ No instances are left

$$IG(Class) = -\frac{4}{6}\log_2\frac{4}{6} - \frac{2}{6}\log_2\frac{2}{6}$$

$$= 0.390 + 0.528$$

$$= 0.918 \approx 0.92$$

$$IG_{GPA}(Class) = \frac{2}{6}\left[IG(D_1)\right] + \frac{2}{6}IG(D_2) + \frac{2}{6}IG(D_3)$$

L < 1
M < 1
H < 2

$$= \frac{2}{6}\left[\frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right] + \frac{2}{6}\left[\frac{-1}{2}\log_2\frac{1}{2} - \frac{1}{2}\log_2\frac{1}{2}\right]$$

$$+ \frac{2}{6}\left[-\frac{2}{2}\log_2\frac{2}{2} - 0\right]$$

$$= \frac{2}{6}[1] + \frac{2}{6}[1] + \frac{2}{6}[0] = \frac{4}{6} = 0.667$$

$$IG_{studied}(Class) = \frac{3}{6} IG(D_1) + \frac{3}{6} IG(D_2)$$

T — 3, 0

F — 1, 2

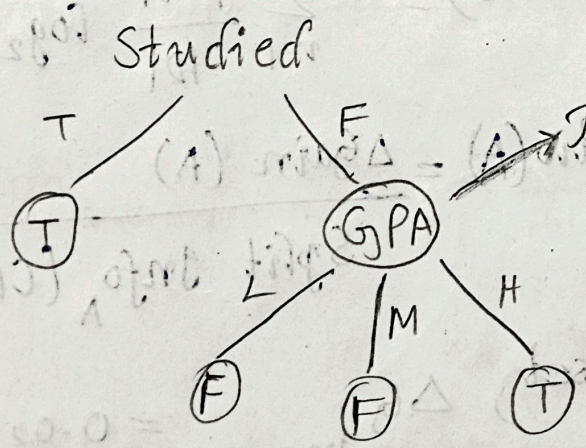
$$= \frac{3}{6}\left[\frac{-3}{3}\log_2\frac{3}{3} - 0\right] + \frac{3}{6}\left[\frac{-1}{3}\log_2\frac{1}{3} - \frac{2}{3}\log_2\frac{2}{3}\right]$$

$$= \frac{3}{6}[0] + \frac{3}{6}[0.528 + 0.390]$$

$$= \frac{3}{6}[0.918] = 0.459$$

$$\Delta G_{GPA} = 0.92 - 0.67 = 0.25$$

$$\Delta G_{studied} = 0.92 - 0.46 = 0.46 \checkmark$$



All instances doesn't match. So, branching will come here. Then we should calculate IG but since only 1 attribute is left, we take it.

→ Then all belongs to same class.

High Variance → Overfitting

(Training accuracy is high & testing accuracy is less)

Why?

9/9/25 Problems of Information Gain or Entropy method

$$IG_{S.No}(Class) = \frac{1}{6}\left[\frac{1}{1}\log_2\frac{1}{1}\right] + \frac{1}{6}\left[\frac{1}{1}\log_2 1\right]$$
$$= 0$$

$$\Delta G_{S.No} = 0.92 - 0$$
$$= 0.92$$

→ Uncertainity very low for these columns, then ΔG high.

Gain Ratio

$$\text{Split Info}_A(Class) = -\sum_{i=1} \frac{|D_i|}{|D|}\log_2\frac{|D_i|}{|D|}$$

$$\text{Gain Ratio}(A) = \frac{\Delta\text{Gain}(A)}{\text{Split Info}_A(Class)}$$

From 1st dataset, $\Delta G_{Income} = 0.029$

$$\text{Split Info}_{Income}(Class) = -\frac{4}{14}\log_2\frac{4}{14} - \frac{6}{14}\log_2\frac{6}{14}$$
$$= \frac{4}{14}\log_2\frac{4}{14}$$
$$= 0.516 + 0.524 + 0.516$$
$$= 1.556$$

$$\text{Gain Ratio}_{(Income)} = \frac{0.029}{1.556} = 0.019$$

a/a/25

2nd dataset:

$\Delta G_{GPA} = 0.26$

$$\text{Split Info}_{GPA}(\text{Class}) = -\frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6} - \frac{2}{6}\log_2\frac{2}{6}$$

$$= 0.528 \times 3 = 1.584$$

$$\text{Gain ratio (GPA)} = \frac{0.26}{1.584} = 0.164$$

$\Delta G_{studied} = 0.46$

$$\text{Split Info}_{studied}(\text{Class}) = -\frac{3}{6}\log_2\frac{3}{6} - \frac{3}{6}\log_2\frac{3}{6}$$

$$= 0.5 + 0.5 = 1$$

$$\text{Gain ratio (studied)} = \frac{0.46}{1} = 0.46$$

$\Delta G_{S.No} = 0.92$

$$\text{Split Info}_{S.No}(\text{Class}) = -\frac{1}{6}\log_2\frac{1}{6} \times 6$$

$$= 0.431 \times 6$$

$$= 2.586$$

$$\text{Gain ratio}_{S.No} = \frac{0.92}{2.586} = 0.358$$

Name of DT based on Gain Ratio = C45 wrong

Name of DT based on Inform^n Gain = ID3

Data mining & Techniques - Chapter 8

Decision Tree

10/9/25

Gini Index → CART

⇒ This will be used for classific^n & regression.

$$Gini(Class) = 1 - \sum_{i=1}^{n} P_i^2$$

$$Gini_A(Class) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

$$\Delta Gini_A(Class) = Gini(Class) - Gini_A(Class)$$

→ Gini index is working on the binary split

→ We're going to form the sets.

3 attributes:

{~~Ø~~} {L} {M} {H} {LM}

{LH} {MH} {~~LHM~~}

→ Whichever split has higher Δ value, that split will be considered.

* In the question, if it is mentioned 'multiway' split, then no need to do that process.

10/9/26

S.NO	Body-Temp	Birth	4 Legged	Hibernates	Class (Mammal)
1	Warm	Y	Y	Y	Y
2	Warm	Y	Y	N	Y
3	Warm	Y	N	Y	N
4	Warm	Y	N	N	N
5	Cold	N	Y	Y	N
6	Cold	N	Y	N	N
7	Cold	N	N	Y	N
8	Cold	N	N	N	N
9	Warm	N	N	N	N
10	Cold	Y	N	N	N

$$\text{Gini Index (Class)} = 1 - \sum_{i=1}^{n} P_i^2$$

$$= 1 - \left(\frac{2}{10}\right)^2 - \left(\frac{8}{10}\right)^2$$

$$= 0.32$$

$$\text{Gini}_{\text{Body Temp}}(\text{Class}) = \frac{5}{10}\,\text{Gini (Warm)} + \frac{5}{10}\,\text{Gini (Cold)}$$

W — 2, 3

C — 0, 5

$$= \frac{5}{10}\left[1 - \left(\frac{2}{5}\right)^2 - \left(\frac{3}{5}\right)^2\right] + \frac{5}{10}\left[1 - 0 - \left(\frac{5}{5}\right)^2\right]$$

$$= \frac{5}{10}\left[0.48\right] + \frac{5}{10}(0) = 0.24$$

$$\Delta\text{Gini}_{\text{Body temp}}(\text{Class}) = 0.32 - 0.24 = 0.08$$

$$\text{Gini}_{\text{Birth}}(\text{Class}) = \frac{5}{10}\,\text{Gini}(Y) + \frac{5}{10}\,\text{Gini}(N)$$

$$= 0.24$$

Y — 2, 3

N — 0, 5

$$\therefore \Delta\text{Gini}_{\text{Birth}} = 0.08$$

10/9/25

$$Gini_{4\ legged} = \frac{4}{10}\left[1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2\right]+\frac{6}{10}\left[1-\left(\frac{6}{6}\right)^2\right]$$

Y — 2, 2

N — 0, 6

$$= \frac{4}{10}\times 0.5 + 0$$

$$= 0.2$$

$$\Delta Gini_{4\ legged} = 0.12$$

$$Gini_{Hibernate} = \frac{4}{10}\left[1-\left(\frac{1}{4}\right)^2-\left(\frac{3}{4}\right)^2\right]+\frac{6}{10}\left[1-\left(\frac{1}{6}\right)^2-\left(\frac{5}{6}\right)^2\right]$$

Y — 1, 3

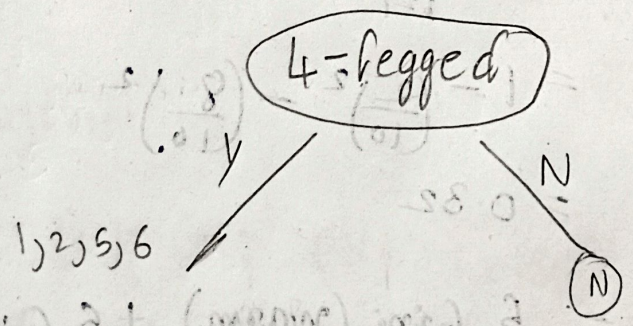
N — 1, 5

$$= \frac{4}{10}\times 0.375 + \frac{6}{10}[0.278]$$

$$= 0.15 + 0.167$$

$$= 0.317$$

$$\Delta G_{Hibernate} = 0.32 - 0.317 = 0.003$$



$$Gini\ (Class\ D_1) = 1-\left(\frac{2}{4}\right)^2-\left(\frac{2}{4}\right)^2 = 0.5$$

$$Gini_{BT}(D_1) = \frac{2}{4}\left[1-\left(\frac{2}{2}\right)^2-0\right]+\frac{2}{4}\left[1-\left(\frac{2}{2}\right)^2-0\right]$$

W — 2, 0

C — 0, 2

$$= 0$$

$$\Delta Gini_{BT} = 0.5 - 0 = 0.5$$

10/4/26

$$Gini_{Birth}(D_1) = \frac{2}{4}\left[1-\left(\frac{2}{2}\right)^2-0\right] + \frac{2}{4}\left[1-\left(\frac{2}{2}\right)^2-0\right]$$

$$= 0$$

Y < 2, 0

N < 0, 2

$$\Delta Gini_{Birth} = 0.5 - 0 = 0.5$$

$$Gini_{Hibernate}(D_1) = \frac{2}{4}\left[1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right] + \frac{2}{4}\left[1-\left(\frac{1}{2}\right)^2-\left(\frac{1}{2}\right)^2\right]$$

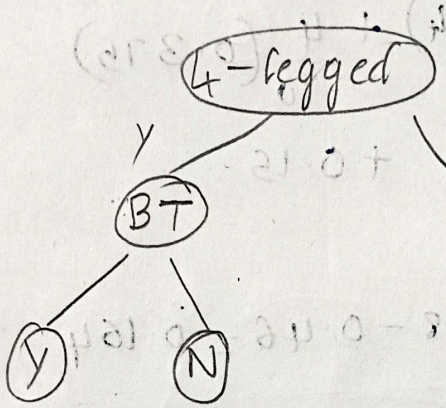
Y < 1, 1

N < 1, 1

$$= \frac{2}{4}(0.5) + \frac{2}{4}(0.5)$$

$$= 0.5$$

$$\Delta Gini_{Hibernate} = 0.5 - 0.5 = 0$$



Multiway Split :

Weather: Sunny, Winter, Rainy

Money: Rich, Poor

Decision: Cinema, Tennis, stay in, shopping

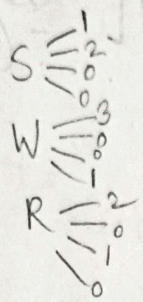
S.NO	Weather	Parents	Money	Decision
1	S	Y	R	C
2	S	N	R	T
3	W	Y	R	C
4	R	Y	P	C
5	R	N	R	S
6	R	Y	P	C
7	W	N	P	C
8	W	N	R	Sh
9	W	Y	R	C
10	S	N	R	T

10/9/25

$$\text{Gini}(\text{Class}) = 1 - \left[\left(\frac{6}{10}\right)^2 + \left(\frac{2}{10}\right)^2 + \left(\frac{1}{10}\right)^2 + \left(\frac{1}{10}\right)^2\right]$$

$$= 0.58$$

$$\text{Gini}_{\text{Weather}}(\text{Class}) = \frac{3}{10}\left[1 - \left(\frac{1}{3}\right)^2 - \left(\frac{2}{3}\right)^2\right]$$



$$+ \frac{4}{10}\left[1 - \left(\frac{3}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right]$$

$$+ \frac{3}{10}\left[1 - \left(\frac{2}{3}\right)^2 - \left(\frac{1}{3}\right)^2\right]$$

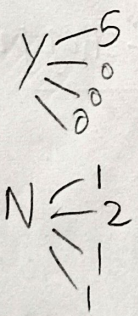
$$= \frac{6}{10}(0.444) + \frac{4}{10}(0.375)$$

$$= 0.266 + 0.15$$

$$= 0.416$$

$$\Delta\text{Gini}_{\text{Weather}} = 0.58 - 0.46 = 0.164$$

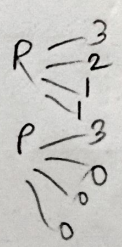
$$\text{Gini}_{\text{Parents}}(\text{Class}) = \frac{5}{10}\left[1 - \left(\frac{5}{5}\right)^2\right] + \frac{5}{10}\left[1 - \left(\frac{1}{5}\right)^2 - \left(\frac{2}{5}\right)^2 - \left(\frac{1}{5}\right)^2 - \left(\frac{1}{5}\right)^2\right]$$



$$= 0 + \frac{5}{10}\left[0.72\right] = 0.36$$

$$\Delta\text{Gini}_{\text{Parents}} = 0.58 - 0.36 = 0.22$$

$$\text{Gini}_{\text{Money}}(\text{Class}) = \frac{7}{10}\left[1 - \left(\frac{3}{7}\right)^2 - \left(\frac{2}{7}\right)^2 - \left(\frac{1}{7}\right)^2 - \left(\frac{1}{7}\right)^2\right]$$

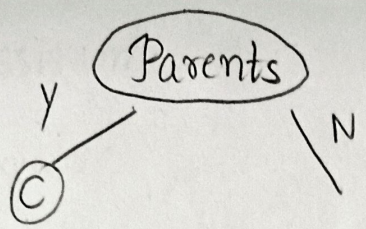


$$+ \frac{3}{10}\left[1 - \left(\frac{3}{3}\right)^2\right]$$

$$= \frac{7}{10}\left[0.694\right] + 0 = 0.486$$

$$\Delta\text{Gini}_{\text{Money}} = 0.58 - 0.486 = 0.094$$

10/10/25



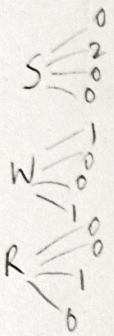
$$\text{Gini}(\text{Class}_P) = 1 - \left[\left(\frac{1}{5}\right)^2 + \left(\frac{2}{5}\right)^2 + \left(\frac{1}{5}\right)^2 + \left(\frac{1}{5}\right)^2\right]$$

$$= 1 - 0.28 = 0.72$$

$$\text{Gini}_{\text{Weather}}(\text{Class}P) = \frac{2}{5}\left[1 - \left(\frac{2}{2}\right)^2\right] + \frac{2}{5}\left[1 - \left(\frac{1}{2}\right)^2 - \left(\frac{1}{2}\right)^2\right]$$

$$+ \frac{1}{5}\left[1 - \left(\frac{1}{1}\right)^2\right]$$

$$= \frac{2}{5}(0.5) = 0.2$$

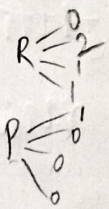


$$\Delta\text{Gini}_{\text{Weather}} = 0.72 - 0.2 = 0.52$$

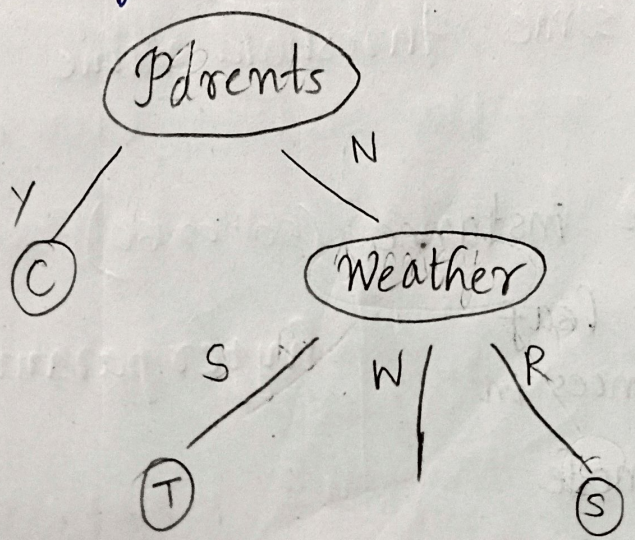
$$\text{Gini}_{\text{Money}}(\text{Class}\,P) = \frac{4}{5}\left[1 - \left(\frac{2}{4}\right)^2 - \left(\frac{1}{4}\right)^2 - \left(\frac{1}{4}\right)^2\right]$$

$$+ \frac{1}{5}\left[1 - \left(\frac{1}{1}\right)^2\right]$$

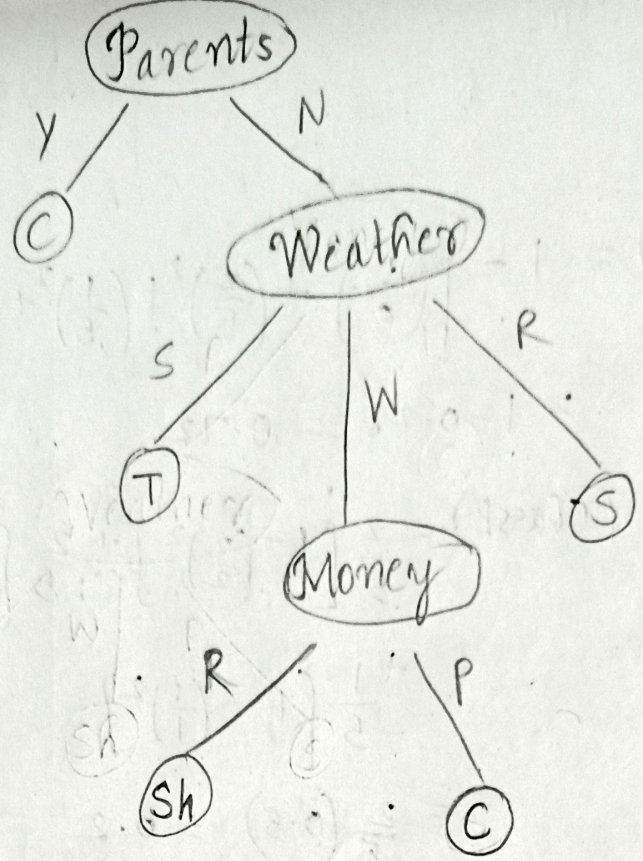
$$= \frac{4}{5}(0.625) = 0.5$$



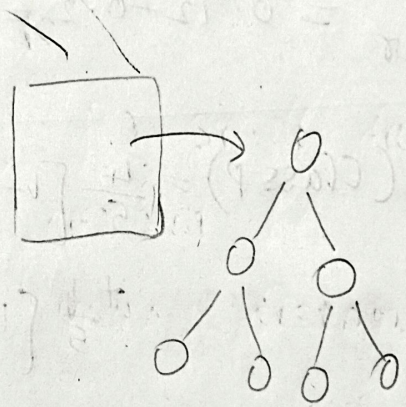
$$\Delta\text{Gini}_{\text{Money}} = 0.72 - 0.5 = 0.22$$



19/9/25



Tree Pruning



Overfitting
high variance

Tree Pruning
- Pre-Pruning
- Post-Pruning

→ We will set some threshold value

→ Max-depth
→ Min-sample instance (50 or 60)
→ Min no of leaf instances in
→ Max_leaf node
→ Max_features

hyperparameters

Assignmt: Take any dataset, set hyperparameter;

if value low-impact?
" " high-impact? } based on this what's the analysis?

Take the features 1 by 1.

→ We need to check if X_1 & X_2 are required for Y_2 or not. Test is there: chi-square test

Pre-Pruning → Thresholds
→ Dependency
→ Improvement (difference in delta values) ← for each split

Post → Whole decision tree should be constructed

→ Removal of a branch impacts the DT accuracy or not. For this, we can apply cross validation approach.

KNN

Predicⁿ based on Majority Voting → Class

Brightness	Saturation	Class
40	20	Red
50	50	Blue
60	90	Blue
10	25	Red
70	70	Blue
60	10	Red..
25	80	Blue

K=5

Brightness = 20

Saturⁿ = 35

Class = ?

12/9/23

Dist(new, X_i)	
② 25	$\sqrt{(20-40)^2+(35-20)^2}$
③ 33.54	$\sqrt{(20-50)^2+(35-50)^2}$ 3 Red, 2 Blue } Red
68.01	$\sqrt{(20-60)^2+(35-90)^2}$
① 14.1	$\sqrt{(20-10)^2+(35-25)^2}$
61.03	$\sqrt{(20-70)^2+(35-70)^2}$
⑤ 47.17	$\sqrt{(20-60)^2+(35-10)^2}$
④ 45.27	$\sqrt{(20-25)^2+(35-80)^2}$

KNN - Instance based

Limitation of KNN:

① Applicable only for small dataset.
RAM Capacity cannot be that much to calculate many distances.

② $x_1 \rightarrow 0.1$, $x_2 \rightarrow 100000$ → Varying ranges b/w features
Apply normalizⁿ first

② It is applicable only for numeric data
Issue of missing values, even if we fill it, it will effect the performance

KNN limitatⁿˢ
- Small
- Numeric
- Normalizⁿ is required

Training time is 0, testing time is high.

12/9/23

Naive Bayes classific^n

likelihood → Prior

$$P(C/X) = \frac{P(X/C) \cdot P(C)}{P(X)} \approx P(X/C) * P(C)$$

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

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

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

If 2 classes: $P\left(\frac{C_1}{X_{new}}\right)$ $P\left(\frac{C_2}{X_{new}}\right)$

Model-based learning.

It will tell the class properties

$$P(X/C) = \prod_{i=1}^{k} P(X_i/C)$$

↳ $X_1 \ldots X_n$

$$P\left(\frac{C_1}{X_{new}}\right) = \frac{P(X_{new}/C_1) \times P(C_1)}{P(X)}$$

$$P\left(\frac{C_2}{X_{new}}\right) = \frac{P(X_{new}/C_2) \times P(C_2)}{P(X)}$$

2 more terms:

① Generative classifier

② Discriminative classifier

→ Naive-Bayes is Generative classifier