TCR INNOVATION

Machine Learning Course

Rishav Das
Lead Data Scientist- Wipro

Probability & Stats

Rishav Das
Lead Data Scientist- Wipro

Pre-Req

Probability & stats

Coin → Head (H)
Coin → Tail (T)

Toss - coin - one

$$\left.\begin{aligned} H &= \frac{1}{2} = 0.5 \\ T &= \frac{1}{2} = 0.5 \end{aligned}\right] \rightarrow 1$$

$$\left.\begin{aligned} HH &= \frac{1}{4} = 0.25 \\ HT &= 1/4 = 0.25 \\ TH &= 1/4 = 0.25 \\ TT &= \frac{1}{4} = 0.25 \end{aligned}\right] \text{Toss two coin's together}$$

$$\sum p(\text{coin}) = 1$$

Example of real world proabilistic approach

Casino / Lottery

i) Ticket price → 6 ⟶ 6 Cr

10 Cr → 60 cr
− 6
54 Cr

Probability will be more

500 ⟶ 1,00,000

Example of real world proabilistic approach

Question 1: A bag consists of 3 red balls, 5 blue balls, and 8 green balls. A ball is selected at random. Find the probability of

1. **Getting a red ball.**
2. **Getting a green ball.**
3. **Not getting a blue ball.**

Answer : Total number of the balls = 3 + 5 + 8 = 16.

1. Let R be the event of getting a red ball. The number of favorable outcome = 3. The required probability is $P(R) = 3/16$
2. Let G be the event of getting a green ball. The number of favorable outcome = 8. The required probability is $P(G) = 8/16 = 1/2$
3. Let B be the event of getting a blue ball. The number of favorable outcome = 5. The required probability of getting blue ball = $5/16$. The probability of not getting a blue ball = $1 - P(B) = 1 - 5/16 = 11/16$

Also, the event of not getting a blue ball is the same as getting a red or green ball. $P(B') = P(R) + P(G) = 3/16 + 8/16 = 11/16$.

Theorems - To be continued

Rule/Theorem

i) summation of all probability is 1

2) All the Probability scores are between 0 - 1

ex: $70\% = \frac{70}{100} = 0.7$

3) $P(A') = 1 - P(A)$

Probability → $\frac{\text{No of success/failure}}{\text{No of event}}$

Success of HH $= \frac{1}{4} = 0.25$

Failure of HH $= 1 - 0.25 = 0.75$

Permutation

- Permutation is the arrangement of items in which **order matters**
- Number of ways of **selection and arrangement of items** in which Order Matters

$$^{n}P_{r} = \frac{n!}{(n-r)!}$$

Combination

- Combination is the selection of items in which **order does not matters** .
- Number of ways of **selection of items** in which Order does not Matters

$$^{n}C_{r} = \frac{n!}{r!\,(n-r)!}$$

Examples:

There are seven boys and three girls on a school tennis team. The coach must select four people from this group to participate in the county championship?

a. *How many* four-person teams can be formed from the group of ten students?

$_{10}C_4 = 210$

b . *In how many ways* can two boys and two girls be chosen to participate in the county championship?

$_7C_2 \cdot {_3C_2} = 63$

c. *What is the probability* that two boys and two girls are chosen for the team?

$\frac{63}{210} \Rightarrow .3$

$\frac{_7C_2 \cdot {_3C_2}}{_{10}C_4}$

Event

Win/loose

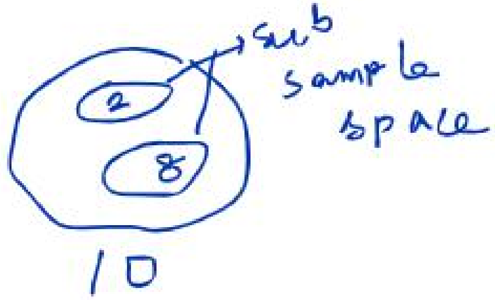
i> Independent

ii> dependent

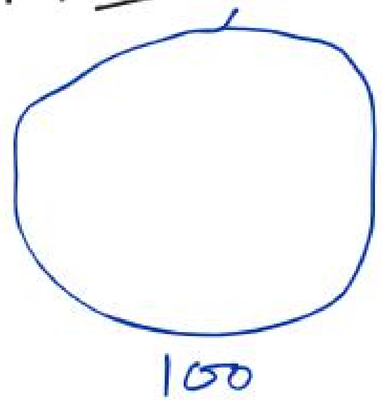
There is no relationship between Previous events and current events

There is a relationship between Previous event and current events

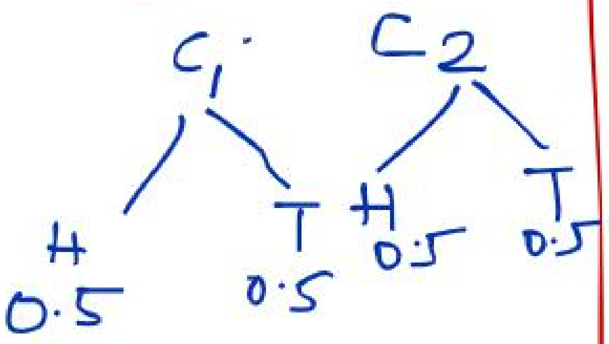
Sample space



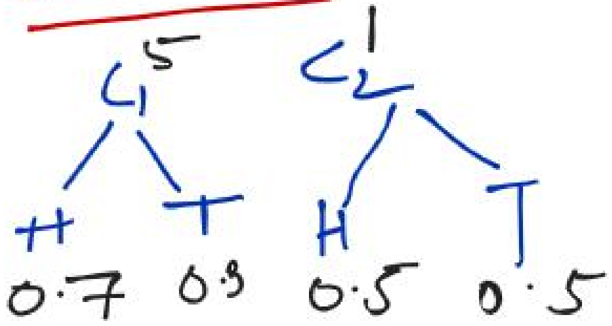
Population



Independent



dependent



Conditional probability

- To find the probability of the event **B** *given* the event **A**, we restrict our attention to the outcomes in **A**. We then find in what fraction of *those* outcomes **B** also occurred.

$$P(\mathbf{B}|\mathbf{A}) = \frac{P(\text{A and B})}{P(\text{A})}$$

- Note: $P(\mathbf{A})$ cannot equal 0, since we know that **A** has occurred.

Example:

	Number of times students visited tutoring			
	One or fewer times	Two to three times	Four or more times	Total
Full time student	12	25	8	45
Part time student	2	5	6	13
Total	14	30	14	58

P(part time | visited four or more times) = $\frac{6}{14} \approx \boxed{0.43}$

find

given

Random Variables & Experiments

Random experiments

HH → 2
HT → 1
TH → 1
TT → 0

Sample space | Real space

$P(H) = \{0, 1, 2\}$

0	1	2
1	HT TH	HH
$\frac{1}{3}$	$\frac{2}{3}$	$\frac{1}{3} = 0.33$
0.33	0.66	

Random Variable

- Discrete RV
  - finite
  - Infinite
  - whole no
  - ex:
    - no of marbles in a Jar
    - no of head/tail
- continous R.V
  - fraction
  - decimals
  - Inifite Values
  - ex: 10.00 — 10.30.
    → Speed of car

Discrete Random Variables

- Random variables (RVs) which may take on only a **countable** number of **distinct** values
  - E.g. the total number of tails X you get if you flip 100 coins
- X is a RV with arity *k* if it can take on exactly one value out of $\{x_1, \ldots, x_k\}$
  - E.g. the possible values that X can take on are 0, 1, 2, ..., 100

DRV - PMF

$P(HH) = \frac{1}{4}$

$P(HT, TH) = \frac{2}{4}$

$P(TT) = \frac{1}{4}$

$\underline{P_1 X_1}$

x_i	$P(x_i)$	Result
0	$0 \cdot \frac{1}{4}$	0
1	$1 \cdot \frac{\cancel{2}}{\cancel{4}}$	$\frac{1}{2} = 0.5$
2	$\cancel{2} \cdot \frac{1}{\cancel{4}\,2}$	$\frac{1}{2} = 0.5$
		1

Continuous Random Variables

Probability density function (pdf) instead of probability mass function (pmf)

A pdf is any function *f*(*x*) that describes the probability density in terms of the input variable *x*.

Probability of Continuous RV

- Properties of pdf
  - $f(x) \geq 0, \forall x$
  - $\int_{-\infty}^{+\infty} f(x) = 1$
- Actual probability can be obtained by taking the integral of pdf
  - E.g. the probability of X being between 0 and 1 is

$$P(0 \leq X \leq 1) = \int_{0}^{1} f(x)dx$$

PDF

computing PDF

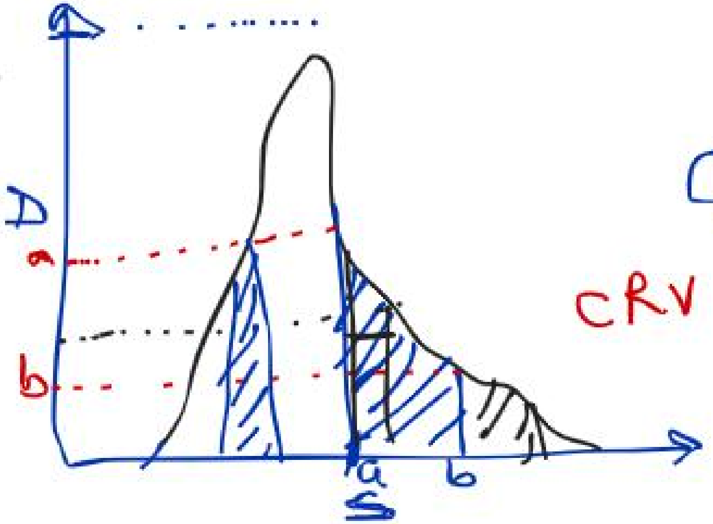
$$E(x) = \int_{-\infty}^{\infty} x \cdot f(x) \cdot dx$$

$$\int_{-\infty}^{1} x \cdot f(x) \cdot dx$$

$$\int_{1}^{\infty} x \cdot f(x) \cdot dx$$

$$\int_{10:00}^{10:30} x \cdot f(x)\, dx = 20$$

$$\int_{16:00:01}^{16:00:05} x f(x) \cdot dx$$



Bayes' theorem

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

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

ex

Bag-I: 4-W, 6-b — E_1

Bag II: 4-W, 3-B — E_2

1B — A

Find the Probability that it was Bag I

$P(E_1) = \frac{1}{2}$ $P(E_2) = \frac{1}{2}$

$P(A|E_1) = \frac{\cancel{6}}{\cancel{10}} = \frac{3}{5}$ $P(A|E_2) = \frac{3}{7}$

$$\underset{\text{Bag I}}{P(E_1}\,|\,\underset{\text{Black ball}}{A)} = \frac{P(A|E_1) \times P(E_1)}{\underbrace{P(E_1)\ P(A|E_1)}_{\text{Picking the ball in } E_1} + \underbrace{P(E_2)\ P(A|E_2)}_{\text{Picking the ball in } E_2}} = \frac{\frac{3}{5} \times \frac{1}{2}}{\frac{1}{2}\frac{3}{5} + \frac{1}{2}\frac{3}{7}} = \frac{7}{12} = 58\%$$

Joint Probability Distribution

	HHH	HHT	HTH	HTT	THH	THT	TTH	TTT
$E_x \rightarrow$	0	0	0	0	1	1	1	1
$E_y \rightarrow$	3	2	2	1	2	1	1	0

$X = \{0, 1\}$ $\quad$ $y = \{0, 1, 2, 3\}$

	y_0	y_1	y_2	y_3
x_0	ϕ	$\frac{1}{8}$	$\frac{2}{8}$	$1/8$
x_1	$\frac{1}{8}$	$\frac{2}{8}$	$\frac{1}{8}$	ϕ

Central Limit Theorem

ex $P = \{1, 2, 5, 3, 4, 6\}$

$S_1 = \{1, 1, 3, 6\}$

$\frac{1+1+3+6}{4} = \frac{11}{4}$

$\Rightarrow 2\ 7$

$S_2 = \{3, 4, 5, 6\}$

$\frac{3+4+5+6}{4}$

$\Rightarrow \frac{18}{4} = 4\ 4$

$S_3 = \{1, 1, 6, 6\}$

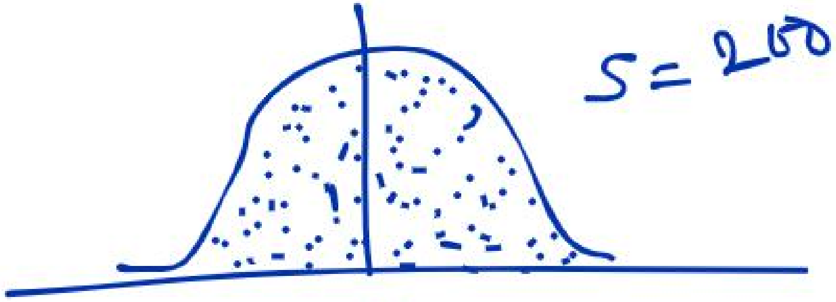
$\frac{1+1+6+6}{4}$

$= \frac{14}{4} = 3\ 5$

$S_1 = 2\ 7$

$S_2 = 4\ 4$

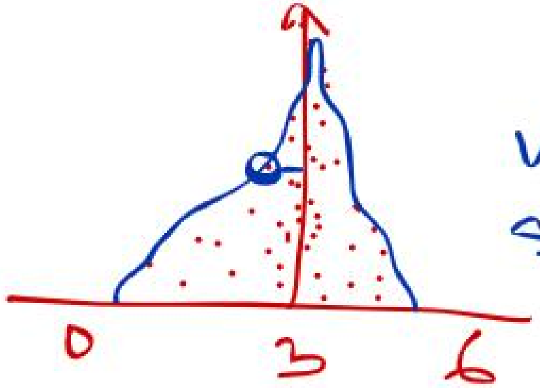
$S_3 = 3\ 5$



3

$S = 200$

Probability distribution



Mean $\rightarrow \mu = n \cdot p$

Varience $= \sigma^2 = npq$

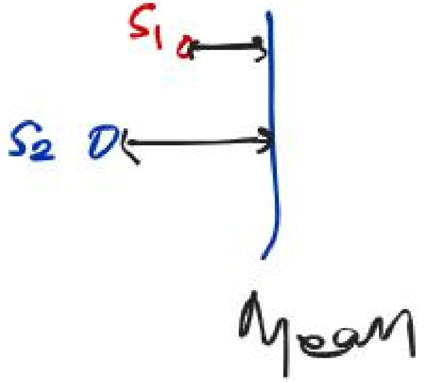
SD $= \sqrt{\sigma^2} = \sqrt{npq}$

$\hookrightarrow \sigma$

n = no of trials

P = success [Probability]

q = failures



ex:

$n = 10$

$H = \frac{1}{2}$ (s)

$T = \frac{1}{2}$ (f)

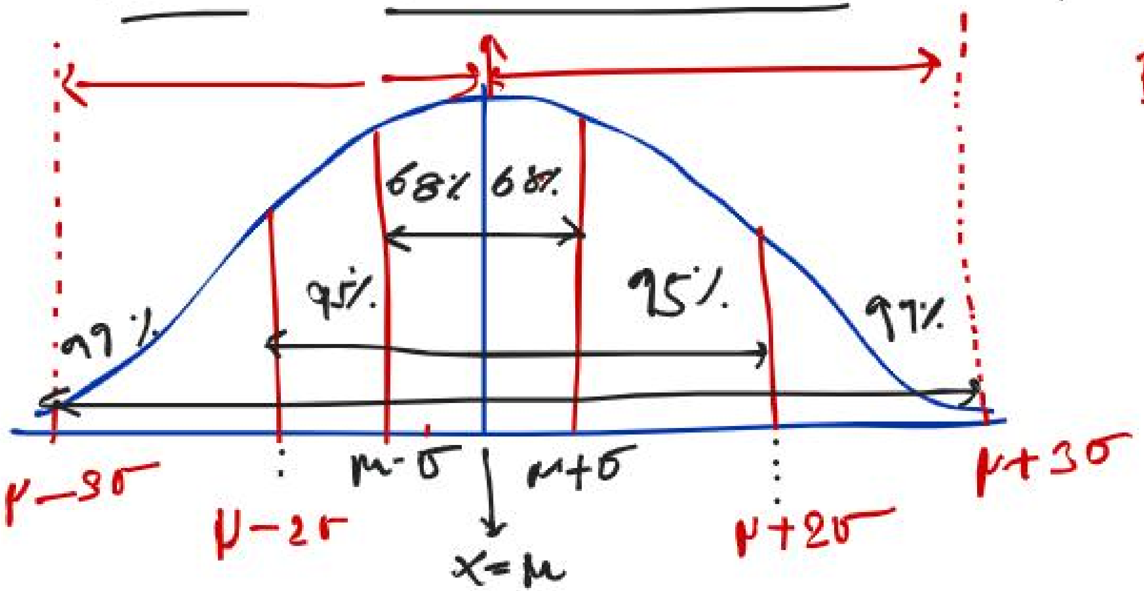
mean $= n \times p = \cancel{10}^{5} \times \frac{1}{\cancel{2}} = 5$

$\sigma^2 = n \times p \times q = \cancel{10}^{5} \times \frac{1}{\cancel{2}} \times \frac{1}{2}$

$= 2.5$

$\sigma = \sqrt{2.5} = 1.25$

Normal Distribution



Bell Shaped Curve

200

- Continous Probability distribution

$$\text{Pdf} = f(x) = \frac{1}{\sigma\sqrt{2\pi}} \cdot e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

$$\int_0^{200} x \cdot f(x) \cdot dx$$

Standard Normal Distribution

$$Z = \frac{X-\mu}{\sigma}$$

$$Z = X$$

Standard Normal Variance has $\mu = 0$ and $\sigma = 1$

Poisson Distribution

Discrete R.V

$$p(x) = \frac{e^{-m} m^x}{x!}$$

$$= \frac{e^{np} (np)^x}{x!}$$

$m = np$

should be very large/infinite

should be very less

ex: casino/Lottery system

→ no of participant → ∞

→ $p(win) \simeq 0.1\%$

150 cr

↳ 5 cr

6 cr → 3 people

$$= \frac{3}{5 cr}$$

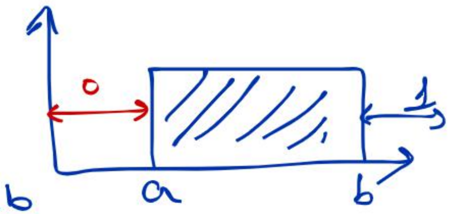
Difference between Variance and Standard Deviation	
Variance	**Standard Deviation**
It can simply be defined as the numerical value, which describes how variable the observations are.	It can simply be defined as the observations that get measured are measured through dispersion within a data set.
Variance is nothing but the average taken out of the squared deviations.	Standard Deviation is defined as the root of the mean square deviation
Variance is expressed in Squared units.	Standard deviation is expressed in the same units of the data available.
It is mathematically denoted as (σ^2)	It is mathematically denoted as (σ)
Variance is a perfect indicator of the individuals spread out in a group.	Standard deviation is the perfect indicator of the observations in a data set.

Uniform Distribution

In probability theory and statistics, the continuous uniform distribution or rectangular distribution is a family of symmetric probability distributions. The distribution describes an experiment where there is an arbitrary outcome that lies between certain bounds. The bounds are defined by the parameters, a and b, which are the minimum and maximum values.

Uniform Distribution

→ continuous RV



$$f(x) = \begin{cases} 0 & x < a \\ x-a/b-a & a \le x \le b \\ 1 & x \ge b \end{cases}$$

Multinomial Distribution

Multinomial distribution, in statistics, a generalization of the binomial distribution, which admits only two values (such as success and failure), to more than two values. Like the binomial distribution, the multinomial distribution is a distribution function for discrete processes in which fixed probabilities prevail for each independently generated value.

Multinomials distribution

Discrete RV | When events are Mutually exclusive

dice → 12 times | Find probability of each dice value occurring twice

1, 2, 3, 4, 5, 6

2 times

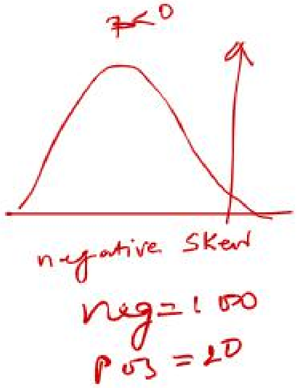
$$\Rightarrow \frac{n!}{n_1!\, n_2!\, n_3!\, n_4!\, n_5!\, n_6!} \cdot p_1^2 \cdot p_2^2 \cdot p_3^2 \cdot p_4^2 \cdot p_5^2 \cdot p_6^2$$

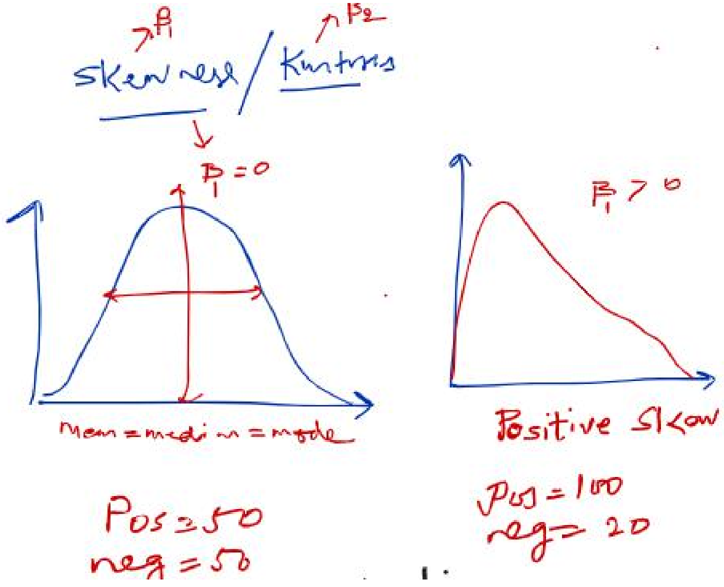
$$\Rightarrow \frac{12!}{2!\,2!\,2!\,2!\,2!\,2!} \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2 \cdot \left(\frac{1}{6}\right)^2$$

$$\Rightarrow 0.004$$

Skewness

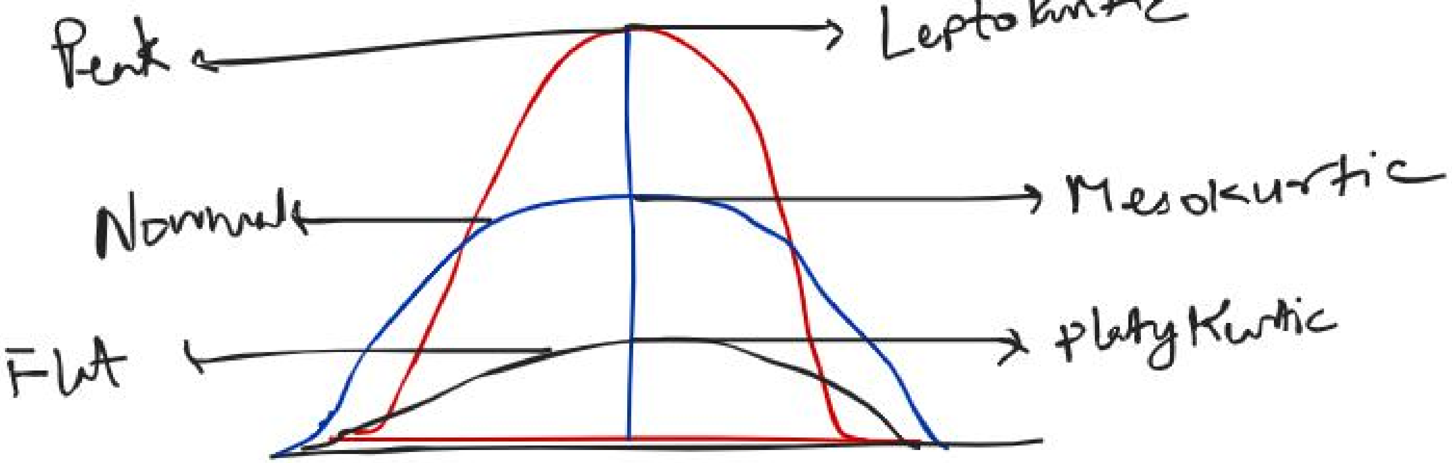
Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution. A normal distribution has a skew of zero, while a lognormal distribution, for example, would exhibit some degree of right-skew.





Kurtosis

kurtosis is a statistical measure that is used to describe distribution. Whereas skewness differentiates extreme values in one versus the other tail, kurtosis measures extreme values in either tail. Distributions with large kurtosis exhibit tail data exceeding the tails of the normal distribution (e.g., five or more standard deviations from the mean). Distributions with low kurtosis exhibit tail data that are generally less extreme than the tails of the normal distribution.



Lines

Lines

slope

Every Straight lines $y = mx + C \longrightarrow mx + C$

constant

where m = 0 and x = 6, C= 1

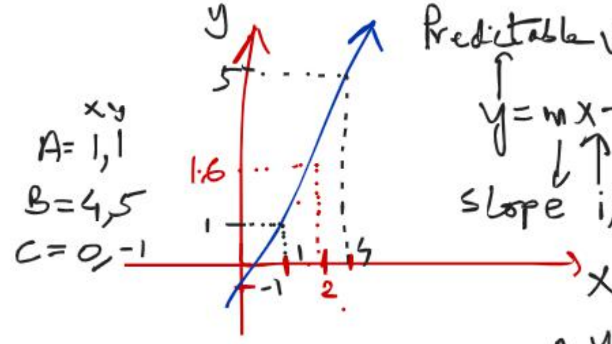
then 0x6 +1 = 0+1 = 1
hence the constant helps to remove the nullifcation from the straight line

thus we can say that when:
mx = 0

y = C, hence C is referred as y intercept

m(slope) = $\dfrac{\text{change in y value}}{\text{change in x value}}$

x y
A= 1,1
B=4,5
C=0,-1



Predictable value

$y = mx + C \longrightarrow$ Intercept

slope i/p

$m = \dfrac{5-1}{4-1} = \dfrac{4}{3} = 1.3$

$m = \dfrac{\Delta y}{\Delta x} = \dfrac{dy}{dx}$

$y = mx + C$
$= 1.3 \cdot x + (-1)$
$y = 1.3x - 1$
$= 1.3(2) - 1$
$= 2.6 - 1 = 1.6$

$m = 1$

$y = 1 \cdot x + C = x + c$
$y = 0 \cdot x + C = C$

Introduction to Linear Regression

- The Pearson correlation measures the degree to which a set of data points form a straight line relationship.
- **Regression** is a statistical procedure that determines the equation for the straight line that best fits a specific set of data.

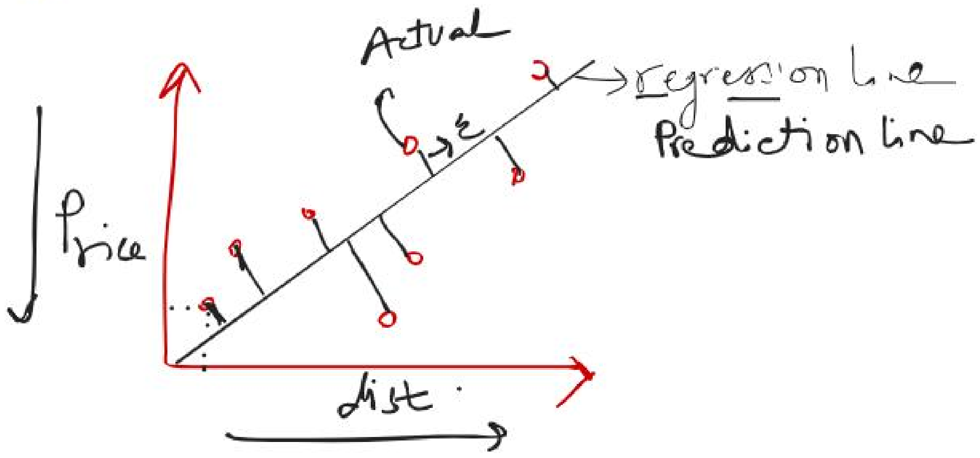
Introduction to Linear Regression (cont.)

- Any straight line can be represented by an equation of the form $Y = bX + a$, where b and a are constants.
- The value of b is called the slope constant and determines the direction and degree to which the line is tilted.
- The value of a is called the Y-intercept and determines the point where the line crosses the Y-axis.

Regression - Linear & MultiLinear

Dependent Variable · Population Y intercept · Population Slope Coefficient · Independent Variable · Random Error term

$$Y_i = \underbrace{\beta_0 + \beta_1 X_i}_{\text{Linear component}} + \underbrace{\varepsilon_i}_{\text{Random Error component}}$$



ex: House

1 Loc
2 - Sqft
3 - Medical
4 - School
5 - Market
6 - Airport
7 - Railway
8 - Quality

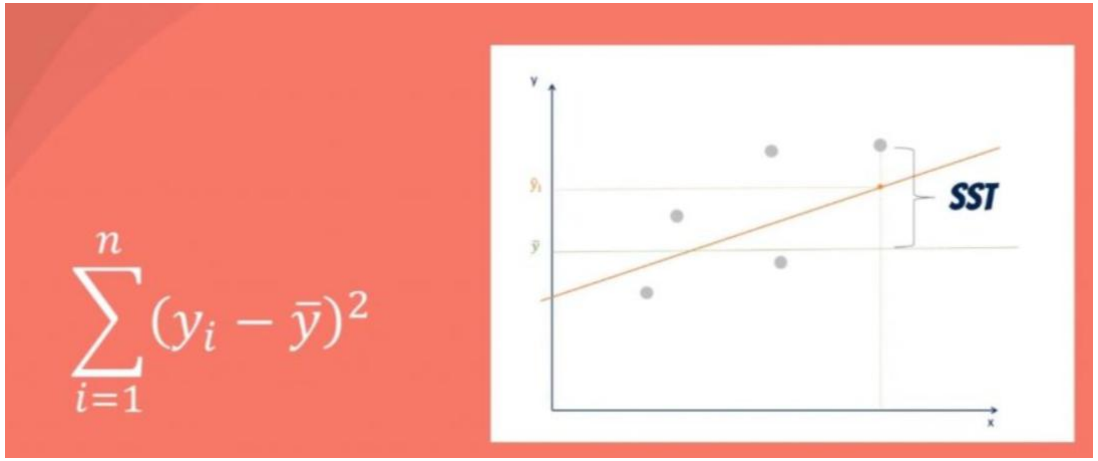
Independent

Price → dependent

Error Calculation

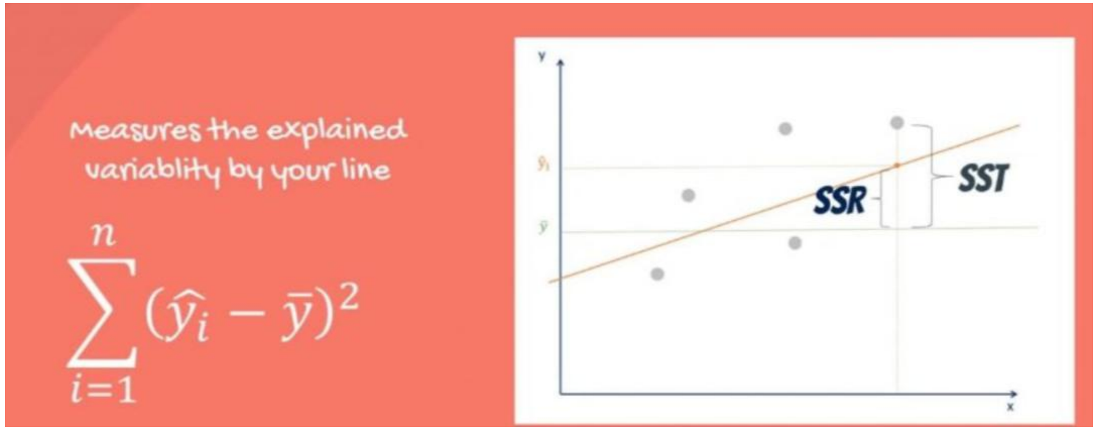
Total Sum of Square

The total sum of squares , denoted TSS, is the squared differences between the observed dependent variable and its mean



Sum of Square Error - Regression

The second term is the sum of squares due to regression, or SSR. It is the sum of the differences between the predicted value and the mean of the dependent variable. Think of it as a measure that describes how well our line fits the data.



Error Calculation

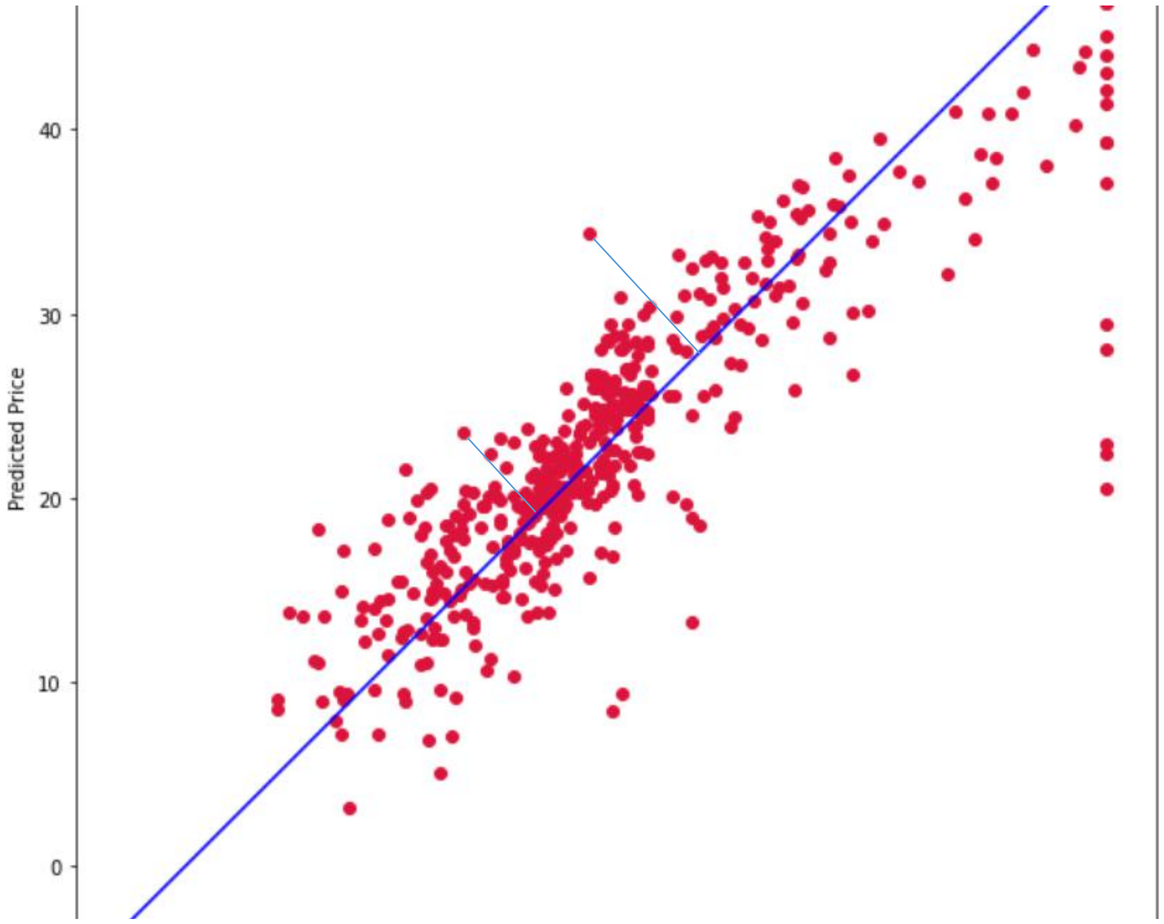
Sum of Squared Error - Prediction

$$= \sum \left(y_i - \hat{y} \right)^2$$

$\hat{y}$ = Predicted Value

y = Actual Value

The last term is the sum of squares error, or SSE. The error is the difference between the observed value and the predicted value.


Predicted Price
0
10
20
30
40

Correlation Analysis - Using Rank

Correlation formula

CRIM	Price	d
2 – R_1	40 R_5	4
10 – R_2	35 R_4	2
20 – R_4	25 – R_2	–2
50 – R_5	20 – R_1	–4
15 – R_3	30 R_3	0
		0

$$\rho = 1 - \frac{6 \sum d^2}{n(n^2-1)}$$

$$\rho = 1 - \frac{6 \cdot (0)^2}{5(5^2-1)}$$

$$= \frac{1}{120} = 0.00$$

Neg		Pos	
CRIM	Price	Age	Price
↑	↓	↑	↑
↓	↑	↓	↓

Normalization

Normalization

$$X_{normal} = \frac{X - min}{max - min} = \frac{65 - 2.90}{100 - 2.90} = \frac{62.1}{97}$$

$$= 0.63$$

$\downarrow$

Range = 0 to 1

Standardization

Standarization

$$X_{st} = \frac{X - \mu}{\sigma} = \frac{65 - 68.57}{28} = -0.12$$

$\mu = 0$

$\sigma = 1$

Logistic Regression

Logistic Regression

$$y = \beta_0 + \beta_1 x + \xi$$

$$P = \frac{1}{1+e^{-y}} \rightarrow \frac{1}{1+e^{-2}} = 0.88 \rightarrow 1$$

$$\frac{1}{1+e^{-(-2)}} = 0.11 \rightarrow 0$$

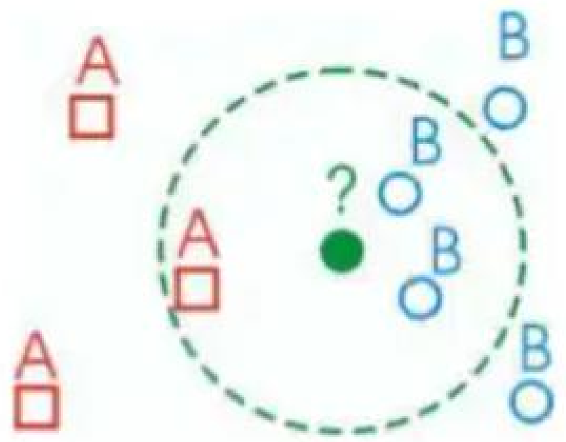
Range: 0 – 1

What is KNN?

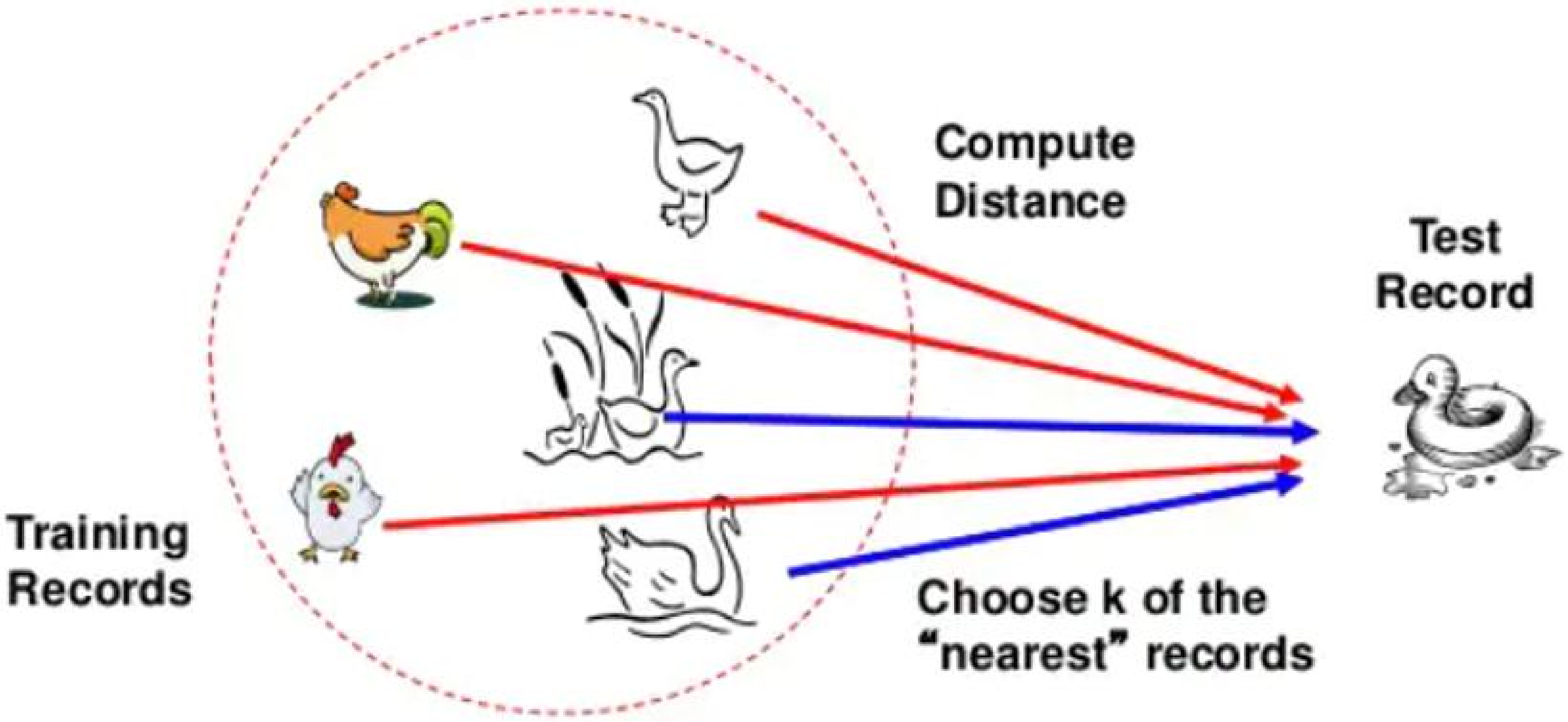
- A powerful classification algorithm used in pattern recognition.
- K nearest neighbors stores all available cases and classifies new cases based on a *similarity measure*(e.g **distance function**)
- One of the top data mining algorithms used today.
- A non-parametric lazy learning algorithm (An Instance-based Learning method).

KNN: Classification Approach

- An object (a new instance) is classified by a majority votes for its neighbor classes.
- The object is assigned to the most common class amongst its K nearest neighbors.(*measured by a distant function*)



Distance Measure



Distance Between Neighbors

- Calculate the distance between new example (E) and all examples in the training set.

- *Euclidean* distance between two examples.
  - $X = [x_1, x_2, x_3, .., x_n]$
  - $Y = [y_1, y_2, y_3, ..., y_n]$
  - The Euclidean distance between *X* and *Y* is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Distance Between Neighbors

- Calculate the distance between new example (E) and all examples in the training set.

- *Euclidean* distance between two examples.
  - $X = [x_1, x_2, x_3, .., x_n]$
  - $Y = [y_1, y_2, y_3, ..., y_n]$
  - The Euclidean distance between *X* and *Y* is defined as:

$$D(X,Y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

Decision Tree

Sample Dataset (was Tennis Played?)

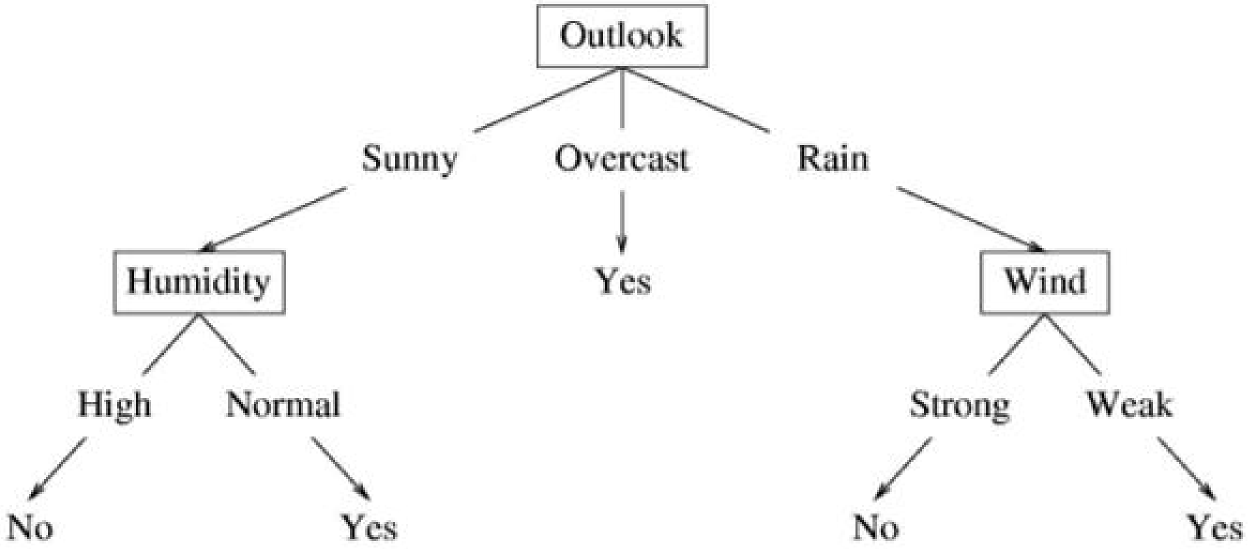
- Columns denote features $\mathrm{X_i}$
- Rows denote labeled instances $\langle \boldsymbol{x}_i, y_i \rangle$
- Class label denotes whether a tennis game was played

$\langle \boldsymbol{x}_i, y_i \rangle$

Predictors				Response
Outlook	**Temperature**	**Humidity**	**Wind**	**Class**
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Decision Tree

- A possible decision tree for the data:



- Each internal node: test one attribute X_i
- Each branch from a node: selects one value for X_i
- Each leaf node: predict Y

Decision Tree

Entropy

Entropy measures the impurity in the given dataset. In Physics and Mathematics, entropy is referred to as the randomness or uncertainty of a random variable X. In information theory, it refers to the impurity in a group of examples. **Information gain** is the decrease in entropy. Information gain computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values.

Entropy is represented by the following formula:-

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Here, **c** is the number of classes and **pi** is the probability associated with the ith class.

Decision Tree

Entropy

Entropy measures the impurity in the given dataset. In Physics and Mathematics, entropy is referred to as the randomness or uncertainty of a random variable X. In information theory, it refers to the impurity in a group of examples. **Information gain** is the decrease in entropy. Information gain computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values.

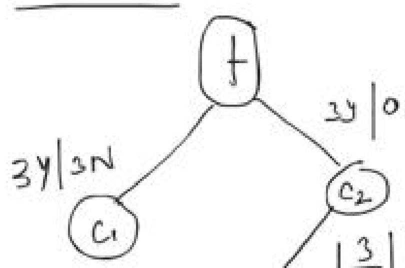
Entropy is represented by the following formula:-

$$Entropy = \sum_{i=1}^{C} -p_i * \log_2(p_i)$$

Here, **c** is the number of classes and **pi** is the probability associated with the ith class.

Rough Work - calculation

<u>Entropy</u>



$\frac{3}{3+0} = \frac{3}{3}$

$N = \frac{0}{3}$

$-\left(\frac{3}{3}\right)\log_2\left(\frac{3}{3}\right) - \frac{0}{3}\log_2\left(\frac{0}{3}\right)$

$-(1)\log_2(1) - \underbrace{0\log_2(0)}_{1}$

$= 0$

E-N: $-\left(\frac{\cancel{3}}{\cancel{6}}\right)\cdot\log\left(\frac{\cancel{3}}{\cancel{6}}\right) - \left(\frac{\cancel{3}}{\cancel{6}}\right)\log_2\left(\frac{\cancel{3}}{\cancel{6}}\right)$

$= -\frac{1}{2}\log\left(\frac{1}{2}\right) - \left(\frac{1}{2}\right)\cdot\log\left(\frac{1}{2}\right)$

$= 1$

GI: $1 - [(P_+)^2 + (P_-)^2]$

$1 - \left[\left(\frac{3}{6}\right)^2 + \left(\frac{3}{6}\right)^2\right] \Rightarrow 1 - [0.25 + 0.25]$

$1 - [0.5] = 0.5$