Mathematical Foundation of Big Data

Krushna N -

• Probability

Probability is the study of random phenomena.

It is measures of unpredictibility for the Particular event.

Range of Probability - 0 to 1.

"If there are two events X & Y Suppose X is one of the Possible event & Y is a Impossible event then Probability of X i.e. P(X) = 1 & P(Y) = 0 "

Lets consider the event E then Probability of event E is denoted by:

$$P(E) = \frac{\text{number of times E can happen}}{\text{Total number of sample space}}$$

$$\boxed{P(E) = \frac{n(E)}{n(S)}}$$

where

$$\boxed{P(\bar{E}) = 1 - \frac{n(E)}{n(S)}}$$ event of non-occurance.

- <u>Random variable</u>

Random variable is a set of Possible values from random experiment

In other words

Consider a function whose domain is the set of Possible out comes & whose range is the subset of set of reals such functions are known as Random Variable

Two types → Discreate RV
→ Contiuous RV.

- Discreate Random variable

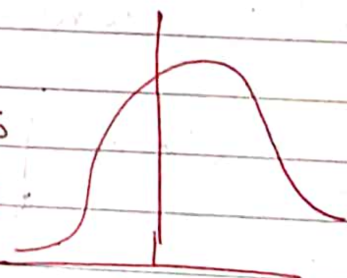
  - finite values
  - Range of Domain.
  - eg.

If we toss a coin what is Probability of event the top face is head

$$X(x) = \begin{cases} 0, & \text{if } x \text{ is tail} \\ 1, & \text{if } x \text{ is head.} \end{cases}$$

- Contenous Random varible
  - Infinite value

eg the probability of Point x is Zero what are the Probability is ~~avilable in an~~ middle age People in india lying between 40kg & 150kg.



- Conditional Probability

It is a Probability of an event occurring given that another event has already occured

denoted by P(A|B)

Reads as Probability of A given B

here event B is already occures

Range of Condital Probability is 0-1

Ex:-

Consider a Pack of 52 fair Card what is the Probability.

Let event A

card drawn is king

Event B

card is Black.

what is Probability of P(A|B).

$$P(A|B) = \frac{n(A \cap B)}{n(B)} = \frac{2}{26} = \frac{1}{13}$$

- Pairwise Independence

The events $A_1, A_2 \ldots A_n$ are said to be Pairwise independent if and only if

$$P(A_i \cap A_j) = P(A_i) \cdot P(A_j),$$
$$i \neq j = 1, 2, \ldots n$$

$\therefore$

$$P(A_1 \cap A_2) = P(A_1) \cdot P(A_2)$$
$$P(A_2 \cap A_3) = P(A_2) \cdot P(A_3)$$
$$P(A_3 \cap A_4) = P(A_3) \cdot P(A_4)$$

The events $A_1, A_2, A_3 \ldots A_n$ are said to be Mutually indepent iff

$$P(A_1 \cap A_2 \cap A_3 \ldots \cap A_n) = P(A_1) \cdot P(A_2) \cdot P(A_3) \ldots P(A$$

- Independence & Exccusiveness

$P(A \cap B) = P(A) \cdot P(B)$	$P(A \cap B) = 0$

If A & B are independent then they can't be execlusive & vice versa

$$\boxed{P(A \cap B) = 1 - P(\bar{A}) \cdot P(\bar{B}) = P(A) \cdot P(B)}$$

- Numericals

(1) In a fair Die

A = { outcome is greater than 3)

B = { out come is even}

find $P(A)$, $P(B)$, $P(A|B)$, $P(\bar{A} \cap B)$

→ $P(A) = \frac{3}{6} = \frac{1}{2}$

$P(B) = \frac{1}{2}$

$P(A \cap B) = \{4, 6\} = \frac{2}{6} = \frac{1}{3}$

$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{1/3}{1/2} = \frac{\cancel{1}}{\cancel{3}}\ \frac{2}{3}$

$P(\bar{A} \cap B) = \{ \frac{1}{6}$

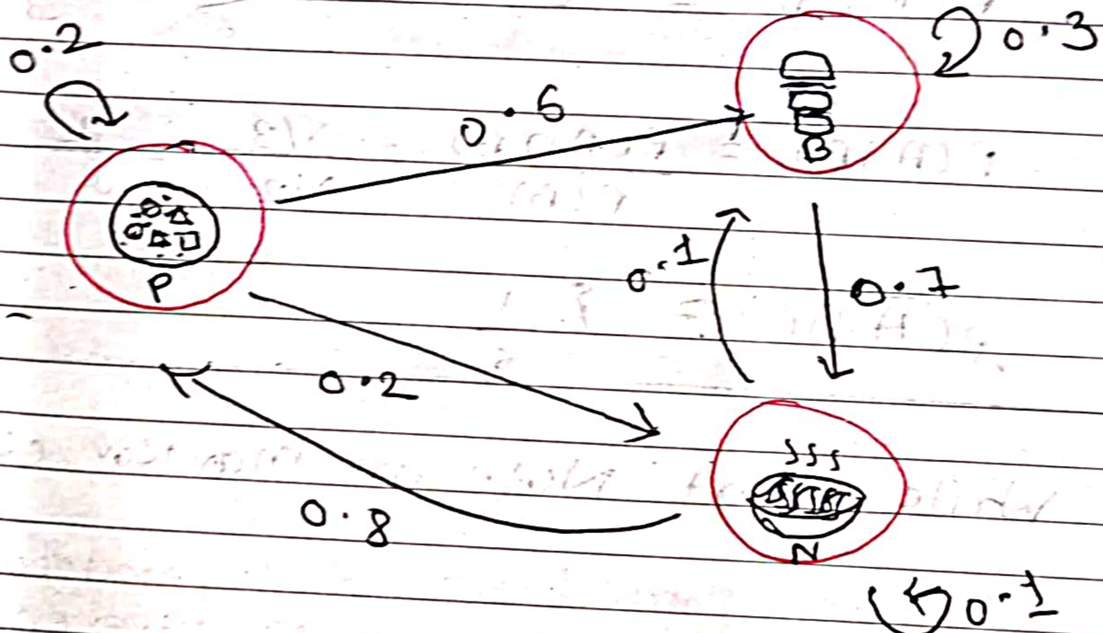
- Write Short Note on markov chain

→ Markov chain is a mathematical model that describes the sequence of events where the probability of each event depends only on State of Previous event and not on any event occur before that.

eg

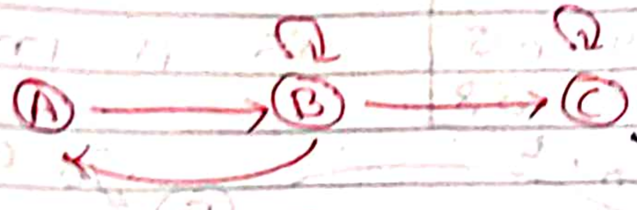
Let's consider a world of 3 food item only where you will be served with only one item each day based on what is served on previous day. the three items are Pizza, burger & Noodles. & their probabiliy are

	Pizza	Burger	Noodles
Pizza	0.2	0.6	0.2
Burger	0	0.3	0.7
Noodles	0.8	0.1	0.1



$P \to B \to B \to N \to P \to P$.

- **Absorbing :** If chain has one absorbing element



Absorbing element

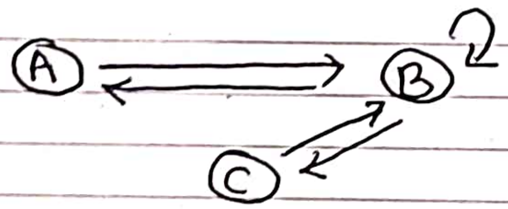
- **Regular chain :** If some power of matrix has only positive elements.

$$P = \begin{bmatrix} 0.25 & 0.50 \\ 0.75 & 0.50 \end{bmatrix} \qquad Q = \begin{bmatrix} 1 & 1 \\ 0 & 1 \end{bmatrix}$$

$$P^2 = \begin{bmatrix} 0.438 & 0.375 \\ 0.562 & 0.625 \end{bmatrix} \qquad Q^2 = \begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$$

Regular ✓ | Not regular ✗

- **Irreducible :** means every state is accisible from every other state.

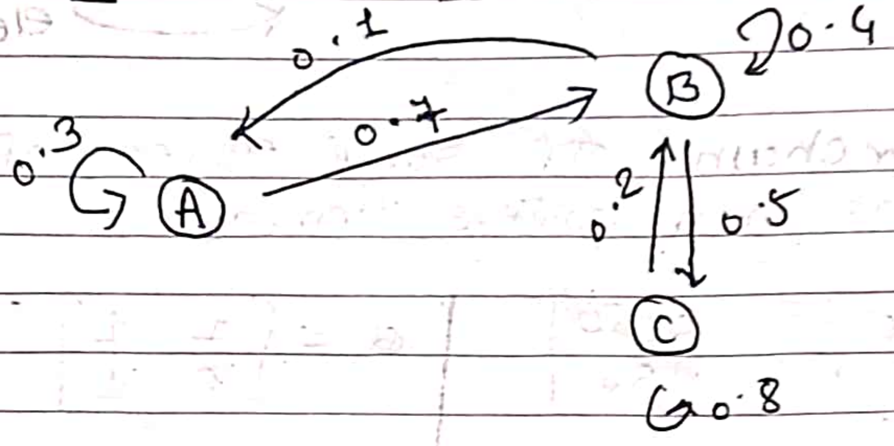


-

- Numericals on markov

① $$\begin{array}{c|ccc} & A & B & C \\ \hline A & 0.3 & 0.7 & 0 \\ B & 0.1 & 0.4 & 0.5 \\ C & 0 & 0.2 & 0.8 \end{array}$$ is it irreducible

→



every state can access from any state
it is irrducible.

②

- Tail and Bound

→ In Probabistic analysis we often need to Bound the Probability that a random Variable deviate far from its mean. These ~~varietous~~ variteous formula for Purpos are called as Tail Bound.

• Flajolet martin Algorithm

The Flajolet martin algorithm can better solve the problem of estimating the number of indepen

• used to count distint element in a given stream

• It's an approximate algorithm to count distict element.

Time complexity = $O(n)$

Space complexity = $O(\log m)$

where n is total number of object & m is number of unique objects.

eg. Input stream $X = 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1$

Hash function $= 6x + 1 \bmod 5$

Step 1 find values of Hash funⁿ

$\therefore h(1) = 6(1) + 1 \bmod 5$

$= 7 \bmod 5$

$= 2.$

h(1) = 2

h(3) = 6(3) mod 5

= 18+1 mod 5

= 3+1

= 4.

h(3) = 4

h(2) = 13 mod 5 = 3

h(4) = 0.

Step 2: Binary equivalent for Hash function

h(1) = 2 = 011

h(3) = 4 = 100

h(2) = 3 = 101

h(1) = 2 = 011

h(2) = 3 = 101

h(3) = 4 = 100

h(4) = 0 = 000

h(3) = 4 = 100

h(1) = 2 = 011

h(2) = 3 = 101

h(3) = 4 = 100

h(1) = 2 = 011

Step 3 count the trailing 0's

h(1) = 010 = 1   h(4) = 0 = 000 = 0

h(3) = 100 = 2   h(3) = 4 = 100 = 2

h(2) = 011 = 0   h(1) = 2 = 011 = 0

h(1) = 010 = 1   h(2) = 3 = 011 = 0

h(2) = 011 = 0   h(3) = 4 = 100 = 2

h(3) = 100 = 2   h(1) = 2 = 011 = 0

Step 4:

Write the value of maximum number of trailing 0's

Value of $r = 2$

The distinct values $= 2^r$

$= 2^2$

$= \underline{4}$

The 4 distinct elements are

1, 2, 3, 4

Blooms Filter

Blooms filter is a space efficient probabistic data structure that used to test wheather an element is member of set or Not.

eg: Cheacking avibility for the username is set membership problem.
wheather set belongs to list of register username or Not.

- Result can be false +ve
  means if algo tells us that name is taken but it accuty not

  Less memory & less accurate.

- Blooms filter of a fixed size can represent a set with an oribitary large number of element

- Adding an element never fails

- Deleting is not possible.

False -ve (Not Possible)

Telling you username dosen't exist even if it exists

False +ve (Possible)

Telling you username exist even it dosen't.

Based on bit vecter of size m & K independent & uniformly distubuted hash fuction

Advantages

- use constant space regardless number of element inserted
- No false -ve so you can trust when it say item does not exist.
- adding element never fails
- It does not store actual elemen

Disadvantages

- can return False +ve
- cannot delete element
- cannot retrive inserted element

- Types of co-relations

① Positive & negative
② Simple & multiple
③ Partial & total
④ Linear & non-linear.