DSBDA Unit 2 : Mathematical foundation of Big Data.

Probability - quantifies the likelihood of an event occuring, expressed as value between 0 and 1.

Sample space (S) - set of all porsible outcomes of a random experiment.

Event (E) - Subset of sample space.

Fundamental Rules of Probability :-

Probability of Event (A) - $P(A) = \frac{n(A)}{n(S)}$ → no. of favorable outcomes; → total no. of outcomes in S

Probability of Union of Two events - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$

Conditional Probability - The probability of event B given that event A has occured

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

Joint Probability - (Multiplication Rule)

For independent events A & B

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

For dependent events A & B

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

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

$\therefore P(A \cap B) = P(B|A) \cdot P(A)$

Random Variables - numerical representation of an outcome from a random event.

Discrete Random Variable :- takes countable values (finite or infinite). It has Probability Mass Function (PMF).

For the DRV 'X', the PMF is ... $P(X = x) = P_x$

probability that X takes value x

The sum of all probabilities must be 1 ie $\sum P(x = x) = 1$

Continuous Random Variable :- takes uncountable values. It has a Probability Density function (PDF)

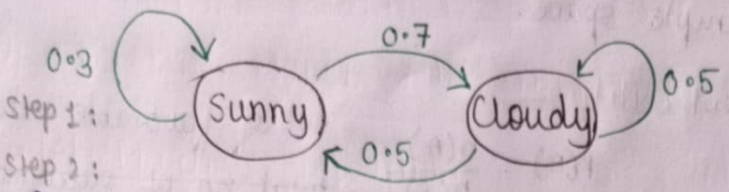
For CRV 'X', the PDF is ... $P(a < X \leq b) = \int_a^b f(x) \cdot dx$

Markov chains - is a mathematical system that jumps from one state to another where next state depends only on current state (not the past !)

- It is a chain of events that is memoryless.

$P(\text{weather}_t \mid \text{weather}_{t-1})$

Eg:



Step 1:

Step 2:

∴ P (Sunny | Sunny) = 0.3

P (Cloudy | Sunny) = 0.7

P (Cloudy | Cloudy) = 0.5

P (Sunny | Cloudy) = 0.5

	To Sunny	To Cloudy	
From Sunny	0.3	0.7	= 1
From Cloudy	0.5	0.5	= 1

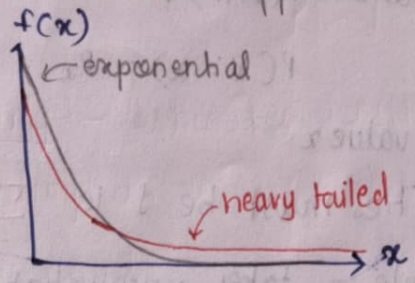
Step 3:

∴ State Transition probabilities = $\begin{bmatrix} 0.3 & 0.7 \\ 0.5 & 0.5 \end{bmatrix}$

Tail bounds - gives idea about how fast the probability of extreme values (tails) decreases as we move away from mean [ie how soon distribution tends to 0.)

Heavy tailed distributions :- decay slowly ie it goes to zero slowly / slower than one with exponential tails ie more bulk under the curve of probability distribution.
Tends to have more outliers with very high values.

Thin tailed distributions :- opposite of heavy tailed



Markov's chain & Random Walks. (A special type of Markov chain)

Sequence of steps, where each step is choosen randomly.
The direction doesn't depend on past, you just move based on some probability.

eg. Tossing a coin. Heads → move right (+1)
Tails → move left (-1)
} mean = 0

Pair-wise independence :-

A set of random variables $X_1, X_2, X_3, \ldots, X_n$ is pairwise independent if for all $i \neq j$, $P(X_i = a, X_j = b) = P(X_i = a) \cdot P(X_j = b)$ for any values of a & b.

↳ only holds for 2 not 3 or more

- If you know the value of one variable, it doesn't tell you anything about the value of another.
- But if you look at three or more variables together, they might be dependent.
- Eg. Rolling 2 Dice.

  let, X_1 be whether 1st die is odd.
  X_2 be whether 2nd die is even } pair-wise independent ie knowing one doesn't affect other.

  X_3 is whether sum of dice is odd

Then X_1, X_2, X_3 together might not be fully independent!.

Universal Hashing :-

- A hash function takes input and maps it to a fixed size output. (hashing) We want it to be random enough to avoid too many collisions (when 2 i/p maps to same o/p)
- Universal Hashing - picking hash function randomly from a family of hash functions, so that its pairwise independent & avoids worst-case attacks.
- Defⁿ: A family of hash functions H is considered universal if for any 2 distinct keys x & y, the number of hash function $h \in H$ where $h(x) = h(y)$ is atmost $\frac{|H|}{m}$, where m is the number of slots in hash table

Approximate Counting :- technique that allows to count a large number of events using a very small amount of computer memory.

- Invented by Robert Morris in 1977.
- This uses probabistic techniques to increment the counter.
- $v(n) = \left\lfloor \frac{\log\left(1 + \frac{n}{a}\right)}{\log\left(1 + \frac{1}{a}\right)} \right\rfloor$ minimized more errors.

eg. Flajolet Martin Algo

Approximate Median :- the median of medians.

- used to find central point for exact selection algorithm.

Steps → Divide list into sublist of len 5
→ sort each sublist & find median directly.
→ use median of medians algo to find median of all medians.
→ And this median is your pivot (central point).

DATA STREAMING MODELS AND STATISTICAL METHODS

Flajolet Martin's algorithm - is a probabilistic method used to estimate the number of distinct elements in a dataset.

- Developed by Phillippe Flajolet & Nigel Martin in 1984, it is particularly useful for handling large datasets where storing all unique elements is unpractical.
- Algorithm :-

① Hash function h(x), converts each element received from stream, into a number (hash value).

② This algo. converts hash value into binary number.

③ Then algo. counts the no. of trailing zeros in binary number & tracks the max number it sees as 'R'.

④ Then algo. estimates the no. of distinct elements passed in stream as 2^R.

$$R = \max(r(\text{binary value}))$$

x	h(x)	Binary	count trailing 0	R	2^R

Distance Sampling :- estimates density / population. Involves ~~detecting~~ recording distances of detected objects from randomly placed lines or points.

Detection function - models the probability of detecting an object based on its distance from observer.

(TB) Random Projections :- technique to reduce dimensionality (while preserving distances between data points) using random linear transformations.

Bloom filters - space-efficient probabilistic data structure used to determine whether an element is member of a set.

- It allows FP (but never FN)
- Advantages: Fast membership checking
  use less memory
  No false Negatives

  Disadvantages: Cannot Remove elements

- Explaination

Bit Array → A fixed sized array of bits (0 or 1), initially set to 0.

Hash functions → A set of 'k' independent hash functions that map element to positions in the bit array.

Co-relation Analysis – Measures strength & direction of a relationship between 2 variables. It ranges from -1 to +1.

$$r = \frac{\Sigma (x - \bar{x})(y - \bar{y})}{\sqrt{\Sigma (x-\bar{x})^2 \cdot \Sigma (y-\bar{y})^2}}$$

$x, y \rightarrow$ data points
$\bar{x}, \bar{y} \rightarrow$ mean of x & y

Analysis of Variation (ANOVA) – It is used to compare 3 or more groups to see if their means are significantly different.
Instead of correlation, it checks variation among groups.

$$F\text{-statistic} = \frac{\text{Between-grp variance}}{\text{Within-grp variance}}$$ (TB)

Mode – value that appears most frequently in a dataset.

eg. Dataset = [4, 1, 2, 4, 3, 4, 5]
Mode = 4

Variance – how spread out the data is

$$\sigma^2 = \frac{\Sigma (x - \bar{x})^2}{n-1}$$

$x \rightarrow$ data
$\bar{x} \rightarrow$ mean
$n \rightarrow$ no. of values

Standard Deviation

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{\Sigma (x-\bar{x})^2}{n-1}}$$

PYQ

Expectation – Expected value of a random variable is the long-term average outcome if an experiment repeated many times.

E of discrete random variable X is $E[X] = \sum x_i P(x_i)$

E of continuous random variable X is $E[X] = \int_{-\infty}^{\infty} x \cdot f(x) \cdot dx$

Expectation is linear. Used in: finance, ML, gaming, stats, risk analysis