

UNIT – II MATHEMATICAL FOUNDATION OF BIG DATA

- Probability theory, Tail bounds with applications, Markov chains and random walks, Pair wise independence and universal hashing, Approximate counting, Approximate median, The streaming models, Flajolet Martin Distance sampling, Bloom filters, Local search and testing connectivity, Enforce test techniques, Random walks and testing, Boolean functions, BLR test for linearity.

Some Definitions

- An **experiment** is an act for which the outcome is uncertain.
 - Examples rolling a die, tossing a coin, surveying a group of people on their favorite soft drink, etc...
- A **sample space** S for an experiment is the set of all possible outcomes of the experiment such that each outcome corresponds to exactly one element in S .
 - The elements of S are called sample points.
 - If there is a finite number of sample points, that number is denoted $n(S)$, and S is said to be a finite sample space.
- Any subset E of a sample space for an experiment is called an **event** for that experiment

Probability



- Probability is a measure of how likely an event is to occur.
- For example –
 - Today there is a 60% chance of rain.
 - The odds of winning the lottery are a million to one.
 - What are some examples you can think of?

Probability Word Problem:

- Lawrence is the captain of his track team. The team is deciding on a color and all eight members wrote their choice down on equal size cards. If Lawrence picks one card at random, what is the probability that he will pick blue?

Number of blues = 3

Total cards = 8

$\frac{3}{8}$ or 0.375 or 37.5%

blue

yellow

red

green

blue

blue

black

black



- Donald is rolling a number cube labeled 1 to 6. What is the probability of the following?

a.) an odd number

odd numbers – 1, 3, 5

total numbers – 1, 2, 3, 4, 5, 6

$$\mathbf{3/6 = 1/2 = 0.5 = 50\%}$$

b.) a number greater than 5

numbers greater – 6

total numbers – 1, 2, 3, 4, 5, 6

$$\mathbf{1/6 = 0.166 = 16.6\%}$$

Solve

- A pair of fair dice is tossed. Determine the probability that
- a) at least one of the dice shows a 6 and
- b) the sum of the two numbers is 5. Round answers to three decimal places.

$$P(\text{at least one of the dice shows a 6}) =$$

$$\frac{n(\text{at least one of the dice shows a 6})}{n(\text{rolls of a pair of fair dice})} =$$

$$\frac{11}{36} \approx .306$$

$$P(\text{the sum of the two numbers is 5}) =$$

$$\frac{n(\text{the sum of the two numbers is 5})}{n(\text{rolls of a pair of fair dice})} =$$

$$\frac{4}{36} = \frac{1}{9} \approx .111$$

What's the probability that you draw 2 aces when you draw two cards from the deck?

$$P(\text{draw ace on first draw}) = \frac{\text{\# of aces in the deck}}{\text{\# of cards in the deck}} = \frac{4}{52}$$

$$P(\text{draw an ace on second draw too}) = \frac{\text{\# of aces in the deck}}{\text{\# of cards in the deck}} = \frac{3}{51}$$

$$\therefore P(\text{draw ace AND ace}) = \frac{4}{52} \times \frac{3}{51}$$

Solve

- From a group of 10 women and 5 men, 2 people are selected at random to form a committee. Find the probability that a) only men are selected and b) exactly 1 man and 1 woman is selected. Round answers to three decimal places.

To find this probability we need to know $n(E) = n(\text{only men are selected})$, which is the number of committees that contain 2 men and $n(S) = n(\text{2 person committees})$, which is the total number of 2 person committees.

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

$${}_5C_2 = \frac{5!}{(5-2)!2!}$$

$$= \frac{5!}{3!2!}$$

$$= \frac{5 \cdot 4 \cdot 3!}{3!2!}$$

$$= \frac{5 \cdot 4}{2 \cdot 1}$$

$$= 10$$

$$n(E) = n(\text{only men are selected}) = 10.$$

- **$n(S) = n(2 \text{ person committees})$** there are 10 women and 5 men for a total of 15 people, taken 2 at a time.

$${}_n C_r = \frac{n!}{(n-r)!r!}$$

$${}_{15} C_2 = \frac{15!}{(15-2)!2!}$$

$$= \frac{15!}{13!2!}$$

$$= \frac{15 \cdot 14 \cdot 13!}{13!2!}$$

$$= \frac{15 \cdot 14}{2 \cdot 1}$$

$$= 105$$

$$n(S) = n(2 \text{ person committees}) = 105.$$

$$P(\text{only men are selected}) =$$

$$\frac{n(\text{only men are selected})}{n(2 \text{ person committees})} =$$

$$\frac{10}{105} = \frac{2}{21} \approx .095$$

Conditional Probability

- $P(B|A)$ means "Event B **given** Event A"
- In other words, event A has already happened, now what is the chance of event B?
- $P(B|A)$ is also called the "Conditional Probability" of B given A.

"Probability Of" "Given"

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

Event A *Event B*

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

$P(B|A)$ \rightarrow
 $P(A|B)$ \leftarrow

- A pair of number cubes is rolled . What is the probability that both numbers are odd if their sum is 6?
 - Let A be the event “Both numbers are odd.”
 - Let B be the event “The sum of the numbers is 6.”
 - You need to find the probability of A given B .
 - That is, you need to find $P(A|B)$.

$$P(A \cap B) = \frac{\text{number of outcomes in } A \text{ and } B}{\text{number of outcomes in sample space}} = \frac{3}{36}$$

$$P(B) = \frac{\text{number of outcomes in } B}{\text{number of outcomes in sample space}} = \frac{5}{36}$$

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

$$P(A|B) = \frac{3}{5}$$

- If dice are rolled .What is the probability that the sum of faces will not exceed 7 ? Given that at least one face should show 4 .

- Let A be the event that sum will not exceed 7

- Let B be the even that one face is 4

- To calculate $P(A|B)$

- $$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{11/36}{6/36} = 11/6$$

- The probability that a contractor will get a plumbing contract is $\frac{2}{3}$ and the probability that he will not get electric contract is $\frac{5}{9}$. If probability of getting at least one contract is $\frac{4}{5}$. What is the probability that he will get both the contracts?

- Let A be the event that contractor will get plumbing Contract
- Let B be the event that contractor will get Electric Contract

- $P(A) = 2/3 \quad P(\bar{B}) = \frac{5}{9}$

- $P(B) = \underline{1 - 5/9} = \underline{4/9}$

- $\underline{P(A \cup B)} = \underline{P(A)} + \underline{P(B)} - \underline{P(A \cap B)}$

- $\underline{P(A \cap B)} = \underline{P(A)} + \underline{P(B)} - \underline{P(A \cup B)} = 2/3 + 4/9 - 4/5 = \underline{14/45} \approx 0.311$

- Only 1 in 1000 people has rare disease. Given that true positive = 0.9 and false positive = 0.02. If randomly tested individual is positive, what is the probability that they have disease.

$A \rightarrow$ A person has disease.

$B \rightarrow$ positive tested.

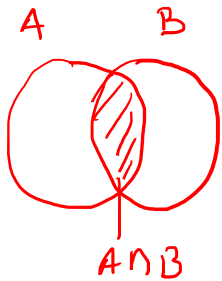
$$P(A) = 0.001$$

$$P(\bar{A}) = 0.999$$

$$P(B|A) = 0.9 \quad \checkmark$$

$$P(B|\bar{A}) = 0.02$$

$$P(A|B) = ?$$



$$P(A \cap B)$$

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

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \text{--- (1)}$$

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

$$= 0.9 * 0.001$$

$$= \underline{\underline{0.0009}} \quad \checkmark$$

$$P(B) = \underline{P(A \cap B)} + \underline{P(\bar{A} \cap B)}$$

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

$$\underline{P(\bar{A} \cap B)} = P(B|\bar{A}) * P(\bar{A})$$

$$= 0.02 * 0.999$$

$$= 0.01998$$

$$P(B) = 0.0009 + 0.01998 = 0.02087$$

$$\frac{0.0009}{0.02087} = \underline{\underline{0.043}}$$

Bayes' Theorem or Bayes' Rule

- Important Theorem associated with Conditional probability
- Allows you to find $P(A|B)$ from $P(B|A)$, i.e. to 'invert' conditional probabilities.

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

- If a single card is drawn from a standard deck of playing cards. If the card is the face card what is the probability that it is a king.

$$P(\text{King}) = \frac{4}{52} = 1/13 \quad \rightarrow \text{prior.}$$
$$P(\text{King} | \text{face}) = \frac{P(\text{face} | \text{King}) \cdot P(\text{King})}{P(\text{face})}$$

$$P(\text{face}) = \frac{12}{52} = 3/13$$

$$P(\text{face} | \text{King}) = 1$$
$$= \frac{1 * 1/13}{3/13}$$
$$= 1/3$$

- At a certain university, 4% of men are over 6 feet tall and 1% of women are over 6 feet tall. The total student population is divided in the ratio 3:2 in favour of women. If a student is selected at random from among all those over six feet tall, what is the probability that the student is a woman?

M = student is Male. F = student is Female.

T = Over 6 feet Tall.

$$P(M) = 2/5 \quad P(F) = 3/5$$

$$P(T|M) = 4/100 \quad P(T|F) = 1/100$$

$$P(F|T) = \frac{P(T|F) \cdot P(F) + P(T|M) \cdot P(M)}{P(T|F) \cdot P(F)}$$

$$\begin{aligned} & \frac{1}{100} \cdot \frac{3}{5} + \frac{4}{100} \cdot \frac{2}{5} \\ &= \frac{3}{11} \end{aligned}$$

- A factory production line is manufacturing bolts using three machines, A, B and C. Of the total output, machine A is responsible for 25%, machine B for 35% and machine C for the rest. It is known from previous experience with the machines that 5% of the output from machine A is defective, 4% from machine B and 2% from machine C. A bolt is chosen at random from the production line and found to be defective. What is the probability that it came from
 - (a) machine A (b) machine B (c) machine C?

Bernoulli Distribution

- Named after Swiss Mathematician James Bernoulli
- Associated with Bernoulli Trial
- Bernoulli Trial
 - Experiment with 2 outcomes which are random and can be considered as success and failure

$$\underline{f(x; p)} = P(\underline{X = x}) = \begin{cases} p & \text{if } x = 1 \\ q = 1 - p & \text{if } x = 0 \end{cases}$$

success

failure

Binomial Distribution

- Discovered by James Bernoulli
- It is the discrete probability distribution of number of successes in a sequence of n independent Bernoulli trials with constant probability p in all n trials.

* • $f(x; n, p) = P(X = x) = \binom{n}{x} p^x q^{n-x}, x = 0, 1, 2, \dots, n$

$X \sim B(n, p)$

- If X is a binomial random variable with parameters n and p i.e. $X \sim B(n, p)$ then X has mean np and variance npq

- i.e. $\mu = np$ and $\sigma^2 = npq$

$q = 1 - p$

- Find formula for the probability distribution of the number of heads when a coin is tossed 4 times.

$$P(X=x) = {}^4C_x \cdot \left(\frac{1}{2}\right)^x \cdot \left(\frac{1}{2}\right)^{4-x} \quad \begin{array}{l} n=4 \\ p=1/2 \\ q=1/2 \end{array}$$

$$x = 0, 1, 2, 3, \underline{4}$$

$$x = 0$$

$$x = 3$$

- If X is binomially distributed RV with $E(X) = 2$ and $\text{Var}(X) = 4/3$ find $P(X=5)$.

$$E(X) = \underline{\underline{\mu}} = 2 \quad \text{var}(X) = 4/3 \quad P(X=5)$$

$$p = 1/3$$

$$n = 6$$

$$q = 2/3$$

$$\mu = np$$

$$\text{var}(X) = 4/3 = npq$$

$$P(X=5) = \frac{4}{243}$$

$$P(X=5) = {}^n C_x p^x q^{n-x}$$

$$= {}^6 C_5 p^5 q^1$$

$$= {}^6 C_5 \left(\frac{1}{3}\right)^5 \cdot \frac{2}{3}$$

$$= \frac{4}{243} = 0.0165$$

$$\underline{\underline{np}} = 2$$

$$npq = 4/3$$

$$q = \frac{4/3}{2} = \frac{2}{3}$$

$$p = 1 - q = 1 - \frac{2}{3} = \frac{1}{3}$$

$$np = 2$$

$$n = \frac{2}{p} = \frac{2}{1/3} = 6$$

$$\text{Mean} = 12$$

$$\text{var} = 4$$

find the distribution.

$$n = 18$$

$$p = 2/3$$

$$q = 1/3$$

$$P(X = x) = {}^{18}C_x \left(\frac{2}{3}\right)^x \left(\frac{1}{3}\right)^{18-x}$$

- A binomial distribution with parameter $n = 5$ satisfies the property $8P(X=4) = P(X=2)$.

- Find p

$$P(X=x) = {}^5C_x \cdot p^x \cdot q^{5-x}$$

$$8 * {}^5C_4 p^4 \cdot q = {}^5C_2 \cdot p^2 \cdot q^3$$

$$p = \frac{1}{3}$$

✓

~~$$p = 1$$~~

Flajolet Martin Algorithm

- **Flajolet-Martin algorithm** approximates the number of unique objects in a stream or a database in one pass.
- The Flajolet-Martin algorithm uses the position of the rightmost set and unset bit to approximate the count-distinct in a given stream.
- If stream contains n unique elements

$$S = 1, 3, 2, 1, 2, 3, 4, 3, 1, 2, 3, 1$$

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

Assume $|b| = 5$

Poisson Distribution

- It is a discrete probability distribution.
- Discovered by Simon –Dennis Poisson
- It is limiting case of binomial distribution under following conditions
 - n ,the number of independent Bernoulli trials ,increases indefinitely i.e. $n \rightarrow \infty$
 - p , the constant probability of success in each Bernoulli trial ,decreases indefinitely i.e. $p \rightarrow 0$
 - $np = \mu$ which is the expected number of success ,remains constant.

Poisson distribution is given by :

$$\lim_{n \rightarrow \infty} B(x; n, p) = \frac{e^{-\mu} \mu^x}{x!}, x = 0, 1, 2, \dots$$

- Where:
- x = number of times and event occurs during the time period
- e (Euler's number = the base of natural logarithms) is approx. 2.72

- On average, the daily sales volume of a mobile at XYZ Electronics is 5
Calculate the probability of XYZ Electronics selling nine mobiles
today.

- $\mu = 5$, since 5 mobiles are the daily sales average
- $x = 9$, because we want to solve for the probability of 9 mobiles being sold

- $e = 2.71828$

- Insert the values into the distribution formula: $P(x; \mu) = (e^{-\mu}) (\mu^x) / x!$

$$= (2.71828^{-5}) (5^9) / 9!$$

$$= (0.0067) (1953125) / (3262880)$$

$$= \underline{0.036}$$

3.6% is the probability of 9 mobiles will be sold today

Applications of Poisson Distribution

- The Poisson Distribution is a tool used in probability theory statistics to predict the amount of variation from a known average rate of occurrence, within a given time frame.
- Find or calculate
 - Load on the web servers ✓
 - Load on telecomm devices
 - No of passengers entering a railway station on a given day
 - No. of wrong telephone numbers dialed on a day

Geometric Distribution

- It is the probability distribution of the number X of independent Bernoulli trials performed until a success occurs, where Bernoulli trials have constant probability of success p .
- For Geometric distribution with parameter $p(0 < p < 1)$ if it takes on the values $1, 2, 3, \dots$. And its probability mass function is given by

$$f(x;p) = P(X=x) = \underline{\underline{q^{x-1} p}}, x = 1, 2, 3, \dots$$

Where $q=1-p$

7 1 ... 6 7
 Failure

- Products produced by a machine has a 3% defective rate. What is the probability that the first defective occurs in the fifth item inspected?

$$p = 0.03 \quad q = 1 - p = 0.97$$

$$P(X = 5) = \underbrace{P(\text{1st 4 non-defective})}_{p^4} \underbrace{P(\text{5th defective})}_p$$

$$= (0.97^4) (0.03)$$

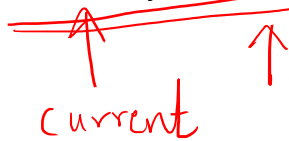
$$= 0.026558$$

Markov Chains

- Markov chains, named after Andrey Markov, are mathematical systems that hop from one "state" to another.
- ***Markov Chains are a class of Probabilistic Graphical Models (PGM) that***
- ***Used to represent dynamic processes i.e., a process which is not static but rather changes with time.***
- ***it concerns more about how the 'state' of a process changes with time.***

- A **Markov chain (MC)** is a *state machine* that has
 - a discrete number of states, q_1, q_2, \dots, q_n ,
 - the transitions between states are nondeterministic, i.e., there is a probability of transiting from a state q_i to another state q_j

$$P(S_t = q_j \mid S_{t-1} = q_i).$$


current

Markov Property

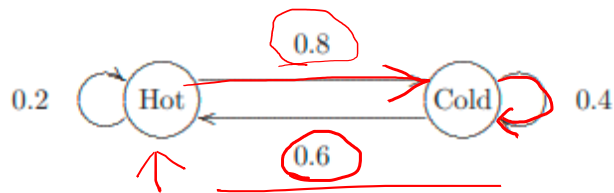
- Markov property states that, a state at time **$t+1$** is dependent only on the current state 't' and is independent of all previous states from **$t-1, t-2, \dots$** . In short, to know a future state, we just need to know the current state.

$$\mathbb{P}(X_{t+1} = s \mid X_t = s_t, X_{t-1} = s_{t-1}, \dots, X_0 = s_0) = \mathbb{P}(X_{t+1} = s \mid X_t = s_t),$$

for all $t = 1, 2, 3, \dots$ and for all states s_0, s_1, \dots, s_t, s .

- A Markov chain model can help answer the three basic problems/questions:
- Problem 1: What is the probability of a certain state sequence?
- Problem 2: What is the probability that the chain remains in a certain state for a period of time?
- Problem 3: What is the expected time that the chain will remain in a certain state?

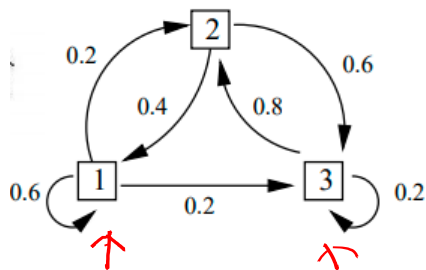
- Transition Diagram



- Transition Matrix

$$X_t \begin{Bmatrix} \text{Hot} \\ \text{Cold} \end{Bmatrix} \begin{pmatrix} \text{Hot} & \text{Cold} \\ \underline{\underline{0.2}} & \underline{\underline{0.8}} \\ \underline{\underline{0.6}} & \underline{\underline{0.4}} \end{pmatrix}$$

The diagram shows the transition matrix for the Markov chain. The rows represent the current state X_t (Hot, Cold) and the columns represent the next state X_{t+1} (Hot, Cold). The transition probabilities are: $P(\text{Hot} \rightarrow \text{Hot}) = 0.2$, $P(\text{Hot} \rightarrow \text{Cold}) = 0.8$, $P(\text{Cold} \rightarrow \text{Hot}) = 0.6$, and $P(\text{Cold} \rightarrow \text{Cold}) = 0.4$. Red boxes and underlines highlight the matrix structure and the specific probability values.



$$\left\{ \begin{array}{l} 1 \rightarrow 1 \rightarrow 3 = \\ 1 \rightarrow 2 \rightarrow 3 = \end{array} \right\}$$

$$0.6 * 0.2 \rightarrow$$

• For the given Transition diagram find

• Transition Matrix

• $P(X_2=3 | X_0=1)$

$=$

$X_2 \leftarrow X_1 \leftarrow X_0$

P'

$$\begin{matrix} & \begin{matrix} 1 & 2 & 3 \end{matrix} \\ \begin{matrix} 1 \\ 2 \\ 3 \end{matrix} & \begin{bmatrix} 0.6 & 0.2 & 0.2 \\ 0.4 & 0 & 0.6 \\ 0 & 0.8 & 0.2 \end{bmatrix} \end{matrix}$$

$P(X_2=2 | X_0=2)$

P^2

P^3

$$P^{(m+1)} = P^{(1)} \cdot P^{(m)}$$

$X_3=3$

$$P^2 = \begin{bmatrix} 0.44 & 0.28 & 0.28 \\ 0.24 & 0.56 & 0.2 \\ 0.32 & 0.16 & 0.52 \end{bmatrix}$$

- A computer System can operate in two different modes. Every hour it remains in the same mode or switches to a different mode according to following transition matrix.

$$\begin{matrix} 0.4 & 0.6 \\ 0.6 & 0.4 \end{matrix}$$

- Compute 2-step transition Probability Matrix
- If the system is in Mode 1 at 5:30 pm what is the Probability that it will be in Mode 1 at 8:30pm on the same day?

$$1) P^2 = \begin{matrix} 0.52 & 0.48 \\ 0.48 & 0.52 \end{matrix}$$

M

5:30 — 6:30
6:30 — 7:30
7:30 — 8:30

$$P^3 = \begin{matrix} 0.504 & 0.496 \\ 0.496 & 0.504 \end{matrix}$$

$$P_{11}^{(3)} = 0.496$$

- The pattern of sunny and rainy days on the planet Rainbow is observed. Every sunny day is followed by another sunny day with probability of 0.8. Every rainy day is followed by another rainy day with probability of 0.6.

1) Today is sunny. What is the chance of rain the day after tomorrow?

2) Compute the probability that April 1 next year is rainy day on Rainbow planet.

$P =$

p^2 —

$$p_{12}^2 = \underline{\underline{0.28}}$$

Stationary Distribution of Markov chain

- Let $\{X_n, n \geq 0\}$ be a Markov Chain with transition probability matrix P . If there exists a probability vector such π that

$$\pi P = \pi$$

Then π is called as Stationary Distribution or steady state distribution of Markov Chain

$$\pi = [\pi_1 \ \pi_2 \ \pi_3]$$

$$n \rightarrow \infty \quad P^n$$

$$\pi_1 + \pi_2 + \pi_3 = 1$$

$$P = \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} \rightarrow 1$$

non zero \rightarrow regular

state 1 - sunny
2 - rainy

$$\pi P = \pi$$

$$[\pi_1, \pi_2] \begin{bmatrix} 0.8 & 0.2 \\ 0.4 & 0.6 \end{bmatrix} = [\pi_1, \pi_2]$$

$$0.8\pi_1 + 0.4\pi_2 = \pi_1$$

$$\pi_1 + \pi_2 = 1$$

$$0.2\pi_1 + 0.6\pi_2 = \pi_2$$

$$\underline{\pi_1 = 2/3} \quad \underline{\pi_2 = 1/3}$$

The probability April 1 next year is rainy = $1/3$.

- A computer device can be either in a busy mode (state 1) processing a task, or in an idle mode (state 2), when there are no tasks to process. Being in a busy mode, it can finish a task and enter an idle mode any minute with the probability 0.2. Thus, with the probability 0.8 it stays another minute in a busy mode. Being in an idle mode, it receives a new task any minute with the probability 0.1 and enters a busy mode. Thus, it stays another minute in an idle mode with the probability 0.9. The initial state is idle. Let X_n be the state of the device after n minutes.

- (a) Find the distribution of X_2
- (b) Find the steady-state distribution of X_n .

$$p_0 = (0 \ 1) \quad P = \begin{matrix} & \begin{matrix} 1 & 2 \end{matrix} \\ \begin{matrix} B \\ I \end{matrix} & \begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{bmatrix} \end{matrix} \quad \begin{cases} \pi P = \pi \\ \pi_1 + \pi_2 = 1 \end{cases}$$

$$\underline{(0 \ 1)} \cdot P^2 = (\underline{0.17} \ \underline{0.83}) \quad \downarrow$$

$$P\{X_n = \text{busy}\} = \frac{1}{3}$$

$$P\{X_n = \text{Idle}\} = \frac{2}{3}$$

Solution:

(a) The one-step transition matrix P is given by,

$$\begin{aligned} P &= \begin{bmatrix} p_{11} & p_{12} \\ p_{21} & p_{22} \end{bmatrix} \\ &= \begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{bmatrix} \end{aligned}$$

The distribution of X_2 is given by,

$$p^{(2)} = p^{(0)} . P^2$$

where $p^{(0)}$ is given as,

$$p^{(0)} = \begin{bmatrix} 0 & 1 \end{bmatrix}$$

We can find P^2 as,

$$P^2 = \begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{bmatrix} = \begin{bmatrix} 0.66 & 0.34 \\ 0.17 & 0.83 \end{bmatrix}$$

Therefore, we obtain the distribution of X_2 as,

$$p^{(2)} = \begin{bmatrix} 0 & 1 \end{bmatrix} \begin{bmatrix} 0.66 & 0.34 \\ 0.17 & 0.83 \end{bmatrix} = \begin{bmatrix} 0.17 & 0.83 \end{bmatrix}$$

We can find the steady-state distribution Π using the relation,

$$\Pi P = \Pi$$

Substituting the values, we get

$$\begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix} \begin{bmatrix} 0.8 & 0.2 \\ 0.1 & 0.9 \end{bmatrix} = \begin{bmatrix} \pi_1 & \pi_2 \end{bmatrix}$$

Equating the values, we get the equation,

$$0.8\pi_1 + 0.1\pi_2 = \pi_1$$

which is equivalent to

$$0.1\pi_2 = 0.2\pi_1$$

Also, we know that,

$$\pi_1 + \pi_2 = 1$$

Therefore, solving these two equations, we get,

$$0.1\pi_2 = 0.2(1 - \pi_2)$$

$$0.3\pi_2 = 0.2$$

Therefore the steady state probabilities are

$$\pi_2 = \frac{2}{3}$$

$$P = \begin{bmatrix} 3/4 & 1/4 & 0 \\ 1/4 & 1/2 & 1/4 \\ 0 & 3/4 & 1/2 \end{bmatrix}$$

state 0, 1, 2

$$p^{(0)} = \begin{bmatrix} \frac{1}{3} & \frac{1}{3} & \underline{\underline{\frac{1}{3}}} \end{bmatrix}$$

$$P(x_3=1, x_2=2, x_1=1, x_0=2) = ?$$

$$\underline{\underline{P(x_1=1 | x_0=2)}} = \underline{\underline{P(x_0=2)}} * \underline{\underline{P(x_1=1 | x_0=2)}}$$

$$= \frac{1}{3} * \frac{3}{4} = \frac{1}{4}$$

$$\underline{\underline{P(x_2=2, x_1=1, x_0=2)}} = \underline{\underline{P(x_1=1, x_0=2)}} * \underline{\underline{P(x_2=2 | x_1=1, x_0=2)}}$$

$$= \frac{1}{4} * \frac{1}{4} = \frac{1}{16}$$

Q4) a) A petrol station owner is considering the effect on his business (Superpet) of a new petrol station (Global) which has opened just down the road. Currently (of the total market, shared between Superpet and Global) Superpet has 80% of the market and Global has 20%. Analysis over the last week has indicated the following probabilities for customers switching the station they stop at each week: [6]

		To	
		Superpet	Global
From	Superpet	0.75	0.25
	Global	0.55	0.45

- What will be the expected market share for Superpet and Global after another two weeks have past?
- ✓ • What would be the long-run prediction for the expected market share for Superpet and Global?

$$\pi = [\pi_1, \pi_2]$$

$$\pi \pi^T = \pi$$

$$\pi_1 + \pi_2 = 1$$

Q3) a) Assume that a man's profession can be classified as professional, skilled labourer or unskilled labourer. Assume that of the sons of professional men, 80 percent are professional, 10 percent are skilled labourers, and 10 percent are unskilled labourers. In the case of sons of skilled labourers, 60 percent are skilled labourers, 20 percent are professional and 20 percent are unskilled. Finally, in the case of unskilled labourers, 50 percent of the sons are unskilled labourers, and 25 percent each are in the other two categories. Assume that every man has at least one son, and form a Markov chain by following the profession of a randomly chosen son of a given family through several generations. Set up the matrix of transition probabilities. Find the probability that a randomly chosen grandson of an unskilled labourer is a professional man.

[6] //

3 P S U

P 0.8 0.1 0.1

 0.2 0.6 0.2

 0.25 0.25 0.5

U

$$P = \begin{bmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.25 & 0.25 & 0.5 \end{bmatrix}$$

- company selling orange juice (call it brand A) has 20% of the market share and wishes to increase their market share via a marketing campaign. They estimated that the marketing campaign has the effect that:
- Someone using brand A will stay with brand A with 90% probability
- Someone NOT using brand A will switch to brand A with 70% probability
- What is initial state matrix ✓
- Write transition probability Matrix
- Calculate Probability that some one uses brand A after 1 week.

Handwritten notes and calculations:

Initial state matrix $P_0 = 0.2 \quad 0.8$

Transition probability Matrix $P = \begin{bmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{bmatrix}$

Calculation for probability after 1 week:

$$\begin{bmatrix} 0.2 & 0.8 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.7 & 0.3 \end{bmatrix} = \begin{bmatrix} 0.74 & 0.26 \end{bmatrix}$$

The result 0.74 is underlined.

Matrix multiplication is shown as $\begin{bmatrix} 0.2 & 0.8 \end{bmatrix} * \begin{bmatrix} P \\ P \end{bmatrix}$ with an arrow pointing to the transition matrix P .

Labels A and B are written above the columns of the transition matrix P .

The final result is labeled P^2 .

Bloom Filter

- A Bloom filter is a data structure
- invented in 1970s by Burton Bloom
- It is designed to tell you, rapidly and memory-efficiently, whether an element is present in a set.
- The price paid for this efficiency is that a Bloom filter is a probabilistic data structure:
 - It tells us that the element either definitely is not in the set or may be in the set.
- Applications
 - checking availability of username is set membership problem
 - including tracking which articles you've read, detecting malicious web sites, and improving the performance of caches.

Properties of Bloom Filters

- **It uses hash k functions**
- Unlike a standard hash table, a Bloom filter of a fixed size can represent a set with an arbitrarily large number of elements.
- Adding an element never fails.
- The only possible errors are **false positives**: a search for a nonexistent element can give an incorrect answer.
- Bloom filters never generate **false negative** result, i.e., telling you that a username doesn't exist when it actually exists.
- Deleting elements from filter is not possible because, if we delete a single element by clearing bits at indices generated by k hash functions, it might cause deletion of few other elements.

A Bloom filter has two main components:

- A bit array $A[0..m-1]$ with all slots initially set to 0; and
- k independent hash functions h_1, h_2, \dots, h_k , each mapping keys uniformly randomly onto a range $[0, m-1]$
- Insert
 - To insert an element in blooms filter we calculate k hash functions of element.
 - For each resulting hash set the corresponding bit to 1
- Lookup /Search
 - computes k hash functions on x ,
 - The first time one of the corresponding slots of A equal to 0, the lookup reports the item as Not Present, otherwise it reports the item as Present

k



0	1	2	3	4	5	6	7	8	9
1		1		1	1				1

set to zero

x

k = 3

$$\begin{cases} h_1(x) = 0 \\ h_2(x) = 4 \\ h_3(x) = 5 \leftarrow \end{cases}$$

Y

$$h_1(Y) = 2$$

$$h_2(Y) = 5 \leftarrow$$

$$h_3(Y) = 9$$

Y → hash Y → present

may be.

Z

$$h_1(Z) = 0$$

$$h_2(Z) = 9$$

$$h_3(Z) = 4$$

present
false +ve //

specific bit

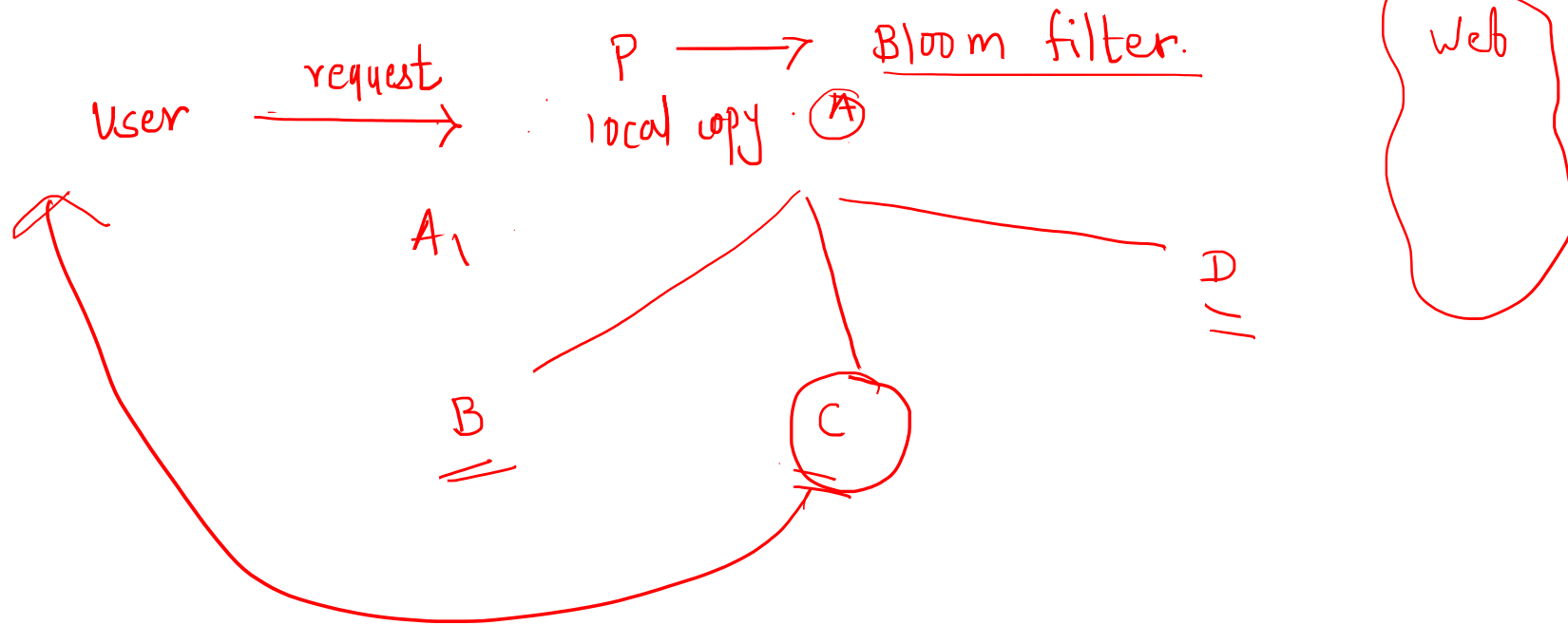
n →

m

$$\left(1 - e^{-\frac{kn}{m}}\right)^k$$

$$\left(1 - \frac{1}{m}\right)^{Kn}$$

Squid → Web proxy cache.

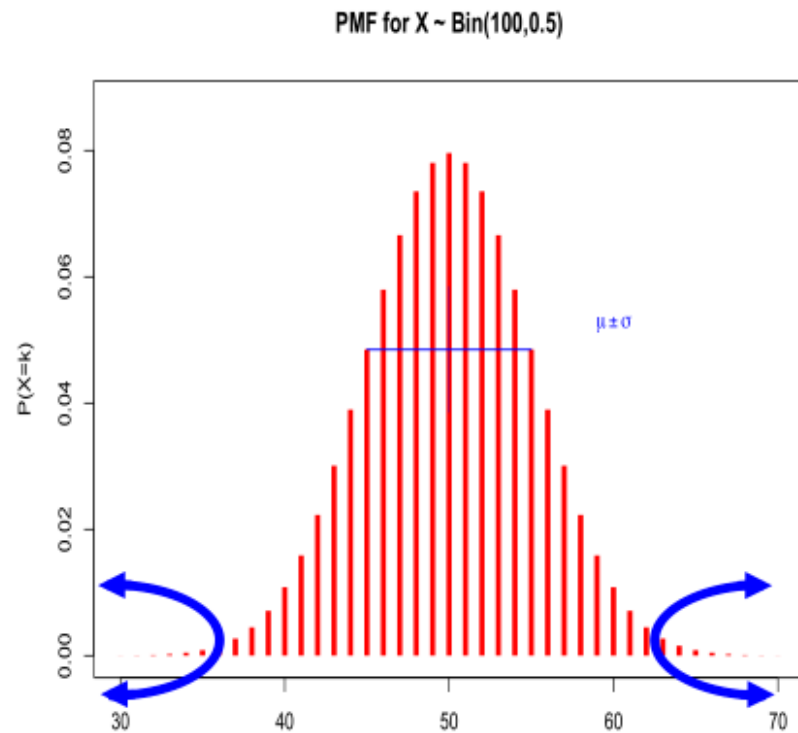


Bitcoin Mobile

Tail Bounds

Tail

- For a random variable X , the tails of X are the parts of the PMF that are “far” from its mean.



- Tail bounds – bound probabilities of extreme events
Important,
- e.g., for “risk management” applications
- Three (of many):
 - Markov: $P(X \geq k\mu) \leq 1/k$ (weak, but general; only need $X \geq 0$ and μ)
 - Chebyshev: $P(|X - \mu| \geq k\sigma) \leq 1/k^2$ (often stronger, but also need σ)
 - Chernoff: various forms, depending on underlying distribution;

- If we know the expected advertising cost is \$1500/day, what's the probability we go over budget? By a factor of 4?
- I only expect 10,000 homeowners to default on their mortgages. What's the probability that 1,000,000 homeowners default?

Markov's Inequality

- Theorem: If X is a non-negative random variable, then for every $\alpha > 0$, we have

$$\bullet P(X \geq \alpha) \leq \frac{E(X)}{\alpha}$$

- Example: if X = daily advertising expenses and $E[X] = 1500$

Then, by Markov's inequality,

$$P(X \geq 6000) \leq \frac{1500}{6000} = 0.25$$

Chebyshev's inequality

- Chebyshev's inequality gives a simple bound on the probability that a random variable deviates from its expected value by a certain amount.
- If we know more about a random variable, we can often use that to get better tail bounds.
- Suppose we also know the variance.
- If Y is an arbitrary random variable with $E[Y] = \mu$, then, for any $\alpha > 0$,

$$P(|Y - \mu| \geq \alpha) \leq \frac{\text{Var}(Y)}{\alpha^2}$$

Chernoff bounds

- There are many different forms of Chernoff bounds
- Let $X = \sum_{i=1}^n X_i$ where $X_i = 1$ with Probability p_i and $X_i = 0$ with Probability $1 - p_i$ and all X_i are independent .
- Let $\mu = E(X) = \sum_{i=1}^n p_i$
- Then

$$P(|X - \mu| \geq \delta\mu) \leq 2e^{-\mu\delta^2/3} \quad \text{for all } 0 < \delta < 1$$