# Probability Course

Lecture 6

Created by : Israa Abdelghany
LinkedIn : www.linkedin.com/in/israa-abdelghany-4872b0222
GitHub : https://github.com/IsraaAbdelghany9

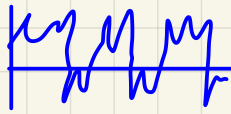
# Session 6

we will discuss:-

- Markou process / Markou chain
- Entropy (Information Entropy)

## Random Process (stochastic process)

↳ vs Random variable

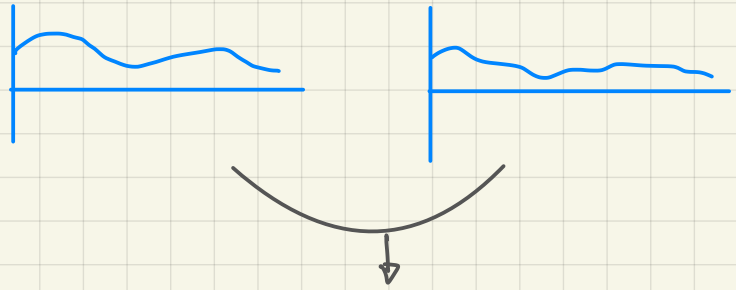
Process ⇒ changes with time, not only value, signal 

variable ⇒ number / value.

$x(\xi, t)$

Random var.

- Example of process is tempreture in different places ⇒ Rand. var
- measuring the tempreture over the day (time) ⇒ Random process



ensemble of different/ possible Realizations of process

Images: Random process "spatial domain"

space

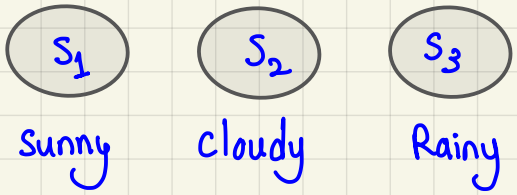
Markov Random process :- special type of Random processes

↓

Discrete → Continuous will not be discussed.

يستخدم في حساب التوقع الاقتصاد والبورصة و signal processing

depend on divide my process into some possible states.



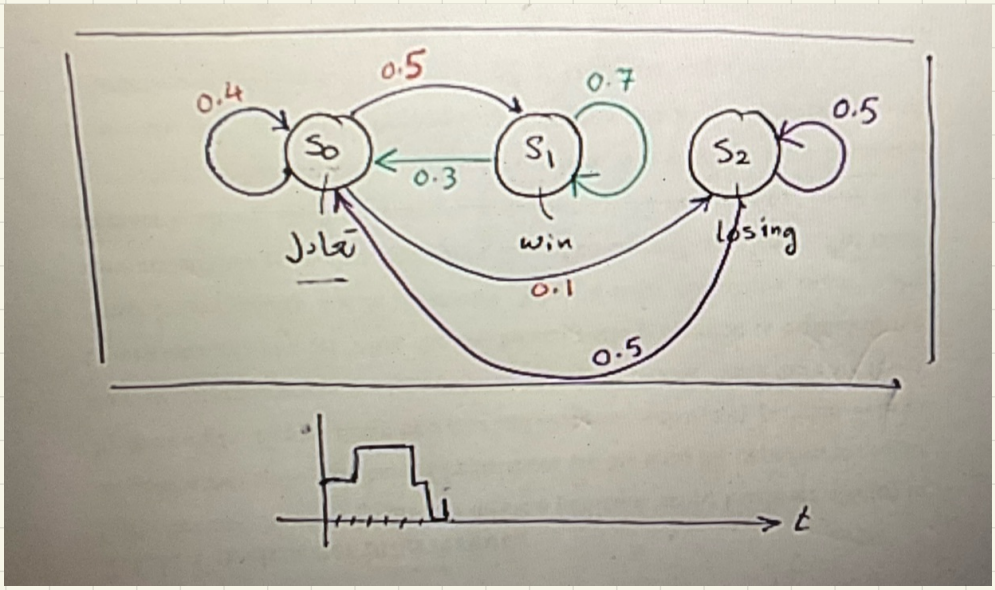

but note that not all processes that can be divided into states is a markov Random process

⇒ it has a characteristic :-

- memory less
  - next state don't depend on last state
- short term memory

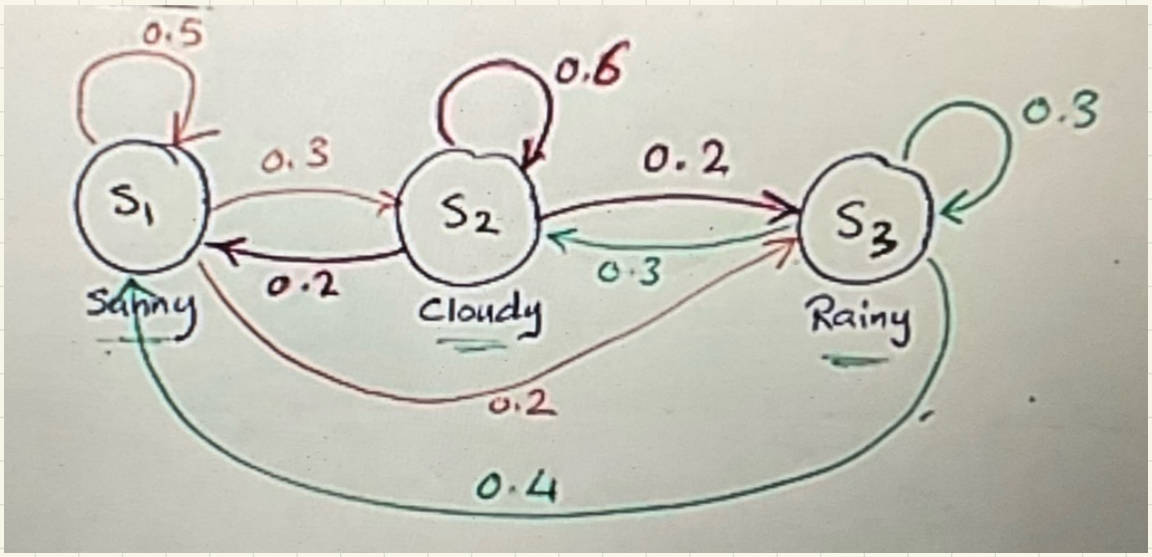
depend or has a value of prediction depend on the current state not affected by the future

Example :- match-




$$\begin{bmatrix} 0.4 & 0.5 & 0.1 \\ 0.3 & 0.7 & 0 \\ 0.5 & 0 & 0.5 \end{bmatrix}$$

Example 2 :-




Next state

| Current state | $S_1$ | $S_2$ | $S_3$ | |
|---|---|---|---|---|
| $S_1$ | 0.5 | 0.3 | 0.2 | = 1 |
| $S_2$ | 0.2 | 0.6 | 0.2 | = 1 |
| $S_3$ | 0.4 | 0.3 | 0.3 | = 1 |

← transition Probability matrix

Some books or websites transpose this matrix

predictions using markov chain:- $\vec{x}^{+} = A\vec{x}$ (A: transition matrix; $\vec{x}$: current state matrix; $\vec{x}^{+}$: next state matrix)

next state vector:- $\begin{bmatrix}1\\0\\0\end{bmatrix}$ ← initial state $S_0$ only (فقط)

$$x^{(1)} = \begin{bmatrix}0.4 & 0.5 & 0.1\\ 0.3 & 0.7 & 0\\ 0.5 & 0 & 0.5\end{bmatrix}$$

---

## Notes:-

Note that transition probability is like conditional probability


given $S_0$ the probability to get it again (فقط) is 0.4

$(S_0|S_0) = 0.4$

$(S_1|S_0) = 0.5$ and so on

⇐ markov assums the transition probabilities is constant along the process

$P(B|A) = P(BA)/P(A)$

$P(BA) = P(B|A)\,P(A)$

$P(B) = P(BA_1) + P(BA_2) + P(BA_3)$

$P(B) = P(B|A_1)P(A_1) + P(B|A_2)P(A_2) + P(B|A_3)P(A_3)$

$P(B) = \sum_i P(B|A_i)\,P(A_i)$

$P(S_j^+) = \sum_i P(S_j^+|S_i)\,P(S_i)$

$S_1^{(0)}$ $S_1^{(1)}$ $S_1^{(2)}$ ⋮

$$\begin{bmatrix}P_{0,0} & \cdots\\ \vdots & \end{bmatrix}\begin{bmatrix}S_0\\S_1\\\vdots\end{bmatrix}$$

$P_{0,0} \equiv P(S_0^+/S_0)$

$P_{0,1} \equiv P(S_1^+/S_0)$

$P_{i,j} \equiv P(S_j^+/S_i)$

$$P(B) = \sum_i P(B/A_i)\,P(A_i)$$

$$P(S_j^+) = \sum_i P(S_j^+/S_i)\,P(S_i)$$

$S_1^{(0)}$
$S_1^{(1)}$
$S_1^{(2)}$

$$P(S_j^+/S_0)\,P(S_0) + P(S_j^+/S_1)\,P(S_1) + P(S_j^+/S_2)\,P(S_2)$$

$$P_{0,0} \equiv P(S_0^+/S_0)$$
$$P_{0,1} \equiv P(S_1^+/S_0)$$
$$P_{i,j} \equiv P(S_j^+/S_i)$$

notations of Markov transition Probabilities.

$$\equiv P_{0,j}\,P(S_0) + P_{1,j}\,P(S_1) + P_{2,j}\,P(S_2) = P(S_j^+)$$

[P_o ] [P(S_1) P(S_2) ⋮]

[ ] [ ]

(4)

$$P = \begin{bmatrix} P_{0,0} & P_{0,1} & P_{0,2} \\ P_{1,0} & P_{1,1} & P_{1,2} \\ P_{2,0} & P_{2,1} & P_{2,2} \end{bmatrix}$$

~~$P(S_0^+) = P_{0,0} \cdot P(S_0) + P_{1,0} \cdot P(S_1) + P_{0,2}$~~

$$P(S_0^+) = P_{0,0} \times P(S_0) + P_{1,0} \times P(S_1) + P_{2\times 0}\,P(S_2)$$

$$\vec{X}^+ = [P(S_0)\ P(S_1)\ P(S_2)] \begin{bmatrix} P_{0,0} & P_{0,1} & P_{0,2} \\ P_{1,0} & P_{1,1} & P_{1,2} \\ P_{2,0} & P_{2,1} & P_{2,2} \end{bmatrix}$$

$$[P(S_0^+)\ P(S_1^+)\ P(S_2^+)] =$$

next state "column number"



$$P = \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.4 & 0.3 & 0.3 \end{bmatrix}$$

current state "row number"

$$P = [\leftarrow P_{i,j} \rightarrow]$$

Initially

$$\vec{X}_0 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\vec{X}^{(1)^T} = [1 \quad 0 \quad 0] \begin{bmatrix} 0.5 & 0.3 & 0.2 \\ 0.2 & 0.6 & 0.2 \\ 0.4 & 0.3 & 0.3 \end{bmatrix}$$

$$\vec{X}^{(1)^T} = [0.5 \quad 0.3 \quad 0.2]$$

$$\vec{X}^2 = \vec{X}^{(1)^T} [P] = [0.5 \quad 0.3 \quad 0.2] \begin{bmatrix} & & \\ & & \\ & & \end{bmatrix}$$

$$[x \quad x \quad x]$$

# "steady state Concept"

→ if I stayed doing this expermint for a long time will it be const or it will stay differ

→ steady state is about it will be const in some point

< the probability will stay the same "vector" >

current step

$$P = \begin{bmatrix} p_{0,0} & p_{1,0} & p_{2,0} \\ p_{0,1} & p_{1,1} & p_{2,1} \\ p_{0,2} & p_{1,2} & p_{2,2} \end{bmatrix}$$

next step →

$$\| \quad \| \quad \|$$
$$1 \quad 1 \quad 1$$

- This is the transpose of the transition matrix

$$x^{(1)} = \begin{bmatrix} \quad \end{bmatrix} = \begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} \quad \end{bmatrix} \begin{bmatrix} \quad \end{bmatrix} \rightarrow x^{(0)}$$

$$x^{(n)} = P^n \vec{x}^{(0)}$$

↳ $P \times P \times P \cdots P$, n times

if P is diagonalizable.

$$P \times P = \begin{bmatrix} \uparrow \\ v \\ \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_i \end{bmatrix} \begin{bmatrix} \uparrow \\ v \\ \downarrow \end{bmatrix}^{-1} \begin{bmatrix} \uparrow \\ v \\ \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_i \end{bmatrix} \begin{bmatrix} \uparrow \\ v \\ \downarrow \end{bmatrix}^{-1}$$

$$P^n \vec{x}^{(0)} = \begin{bmatrix} \uparrow \\ v \\ \downarrow \end{bmatrix} \begin{bmatrix} \lambda_1^n & & \\ & \ddots & \\ & & \lambda_i^n \end{bmatrix} \begin{bmatrix} \uparrow \\ v \\ \downarrow \end{bmatrix}^{-1} \quad x^{(0)} = \vec{x}^{(n)}$$

steady state? will the probability remains const. after period of time?

means $\vec{x}^{(n)} = \underbrace{[\quad]}_{P} \vec{x}^{(n-1)} = \vec{x}^{(n-1)}$

أوصل لقيم احتمالات مهما ضربت فيها الناتج هيثبت مثلاً الفوز 75% دائمًا أو 50% أو إلى آخره

$Px = \underset{\sim}{1}\, x \Rightarrow$ as it a transformation matrix

↝ if I did eigen decomposition

$\lambda_i = 1 \rightarrow \vec{v}_i$ is the steady state

# Information Entropy

Claude shannon considered as the father of modern Communication discussed the Entropy of Information

set1: 0 1 0 1 0 1 1 0 1 0 0 1 1

set2: 0 0 0 0 0 1 0 0 0 1 0 0 0 0 0 1

In our example set1 usually 1, 0 has same amount of information because 0 & 1 have same occurance probability but in set 2 1 has more because its probability is much less than Zero

$$I(x) \propto \frac{1}{f(P(x))}$$

$I(x)$: amount of information

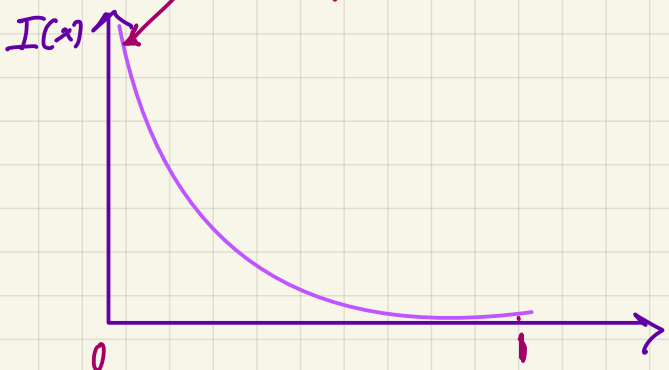
total $= I(x_1) + I(x_2)$ $(x_1, x_2)$

$$I(x) = \log_2 \frac{1}{P(x)} = \text{bits} = -\log_2 P(x)$$

$\log_2$: to let the output be in bits

$P(x)$: probability

it happens once so it has alot of info.



# Information theory

Information entropy - topic in Information theory

$P(H) = 0.5$

$P(T) = 0.5$

$I(H) = -\log_2(0.5)$

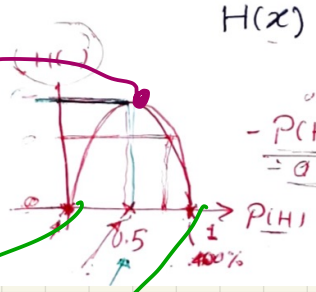
$I(T) = -\log_2(0.5)$

average information $= \sum_i I(x_i)\, P(x_i)$

Entropy $\sim H(x) = \sum_i (-\log_2 P(x_i))\, P(x_i)$

$H(x) = -\sum_i P(x_i) \log P(x_i)$

$-P(H) \log P(H) - P(T) \log P(T)$




0.5 is the avg.

each time on tail

each time on head

both information = zero because nothing New

if Coin was baised the value will appear in the curve

Lossless compression

Information gain

## Cross-Entropy loss function (Classification)

$$H(P,q) = -\sum_i P(x_i) \log q(x_i)$$

- $H(P,q)$: Cross Entropy
- $P(x_i)$: true Prob. / ground truth
- $q(x_i)$: prediction prob.

1

0.6

0.4

normalize error more than L1 norm & L2 norm

(mean square distance & mean absolute distance)