Mathematics for Machine Learning (AI 512)

Amit Chattopadhyay

IIIT-Bangalore

Module 3

Syllabus

Topics

M3: Markov Chains, Stationary Property and Random Walk
M3:. Page Rank Algorithm
M3. Markov Chain Monte Carlo (MCMC) Sampling
M4. Multivariate Distributions, K-Means
M4. EM Algorithm
M4. EM Algorithm for GMM.

Reference Books

1. Foundations of Data Science. By John Hopcroft, Ravindran Kannan.
2. Finite Markov Chains and Algorithmic Applications. By OLLE HÄGGSTRÖM.
3. Introduction to Probability Models. By S.M. Ross.
4. Pattern Recognition and Machine Learning. By Christopher M. Bishop.
5. Mathematics for Machine Learning. By Marc Peter Deisenroth et al.

Markov Chain

Introduction: Stochastic Process

Definition

A stochastic process indexed by an index set $\mathbb{I}$, is a collection of random variables $\{X_t : t \in \mathbb{I}\}$ on a probability space (Ω, Δ, P).

Ω : event space/sample space
Δ : σ-algebra
P : probability function

$P: \Delta \to \mathbb{R}$

Examples:

1. X_n : The stock price of a commodity at the n-th week
2. X_t : Number of accidents in a city till time t

Introduction: Stochastic Process

- **State space:** Spectrum of X_n - $\{s_1, s_2, \ldots, s_k\} \subseteq \mathbb{R}$
- **Stage:** Index set $\mathbb{I}$
- **Chain:** Both the index set and state space are discrete

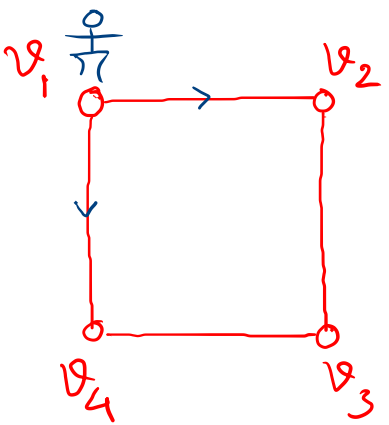
Example 1

A "random walker" in a small town

A random walker in a small town with four streets and four street corners v_1, v_2, v_3 and v_4.

He flips a fair coin and moves forward (clockwise) if 'H' occurs and backward (or counter-clockwise) if 'T' occurs.

X_n: r.v. denoting the index of the street corner at which the walker stands at time n.



$$P(X_0 = 1) = 1 \; ; \quad P(X_0 = 2) = \cdots = P(X_0 = 4) = 0$$

$$P(X_1 = 1) = 0$$

$$P(X_1 = 2) = \frac{1}{2} \; , \; P(X_1 = 3) = 0 \; , \; P(X_1 = 4) = \frac{1}{2}$$

$$P(X_2 = 1 \mid X_0 = i_0, X_1 = 2) = \frac{1}{2}$$

$$P(X_2 = 3 \mid X_0 = i_0, X_1 = 2) = \frac{1}{2}$$

Observations

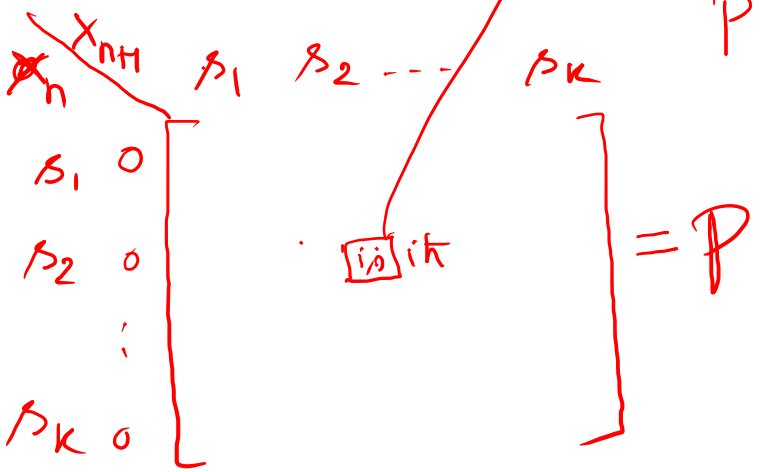
Markov Property

- **Memoryless property:** The coin flip at time $n+1$ is independent of all previous coin flips and hence independent of $X_0, X_1, \ldots, X_n$. Therefore, conditional distribution of X_{n+1} given $(X_0, X_1, \ldots, X_n)$ depends only on X_n.
- **Time homogeneity:** $P(X_{n+1} = j | X_n = i)$ is the same for all n

P_{ij}

$= P(X_1 = j | X_0 = i)$

$= P(X_2 = j | X_1 = i)$



Markov Chain

Definition

Let P be a $k \times k$ matrix with elements $\{P_{i,j:i,j=1,2,\ldots,k}\}$. A stochastic process $\{X_n : n = 0,1,2,\ldots\}$ with finite state space $S = \{s_1, s_2, \ldots, s_k\}$ is said to be a (homogeneous) Markov Chain if $\forall n$, $\forall i,j \in \{1,2,\ldots,k\}$ and $\forall i_0,\ldots,i_{n-1} \in \{1,2,\ldots,k\}$

$$P(X_{n+1} = s_j | X_0 = s_{i_0}, X_1 = s_{i_1}, \ldots, X_{n-1} = s_{i_{n-1}}, X_n = s_i) = P(X_{n+1} = s_j | X_n = s_i).$$

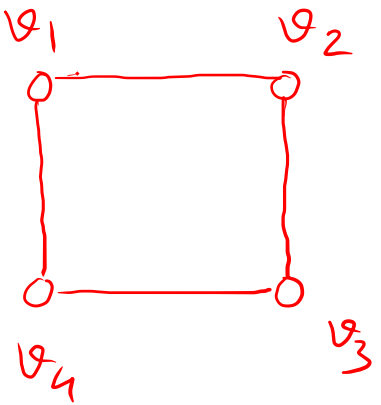
- Memoryless property
- $P(X_{n+1} = s_j | X_n = s_i)$ will be denoted by $P_{i,j}^{n,n+1}$ or $P_{i,j}$

  (Probability of X_{n+1} being in state s_j, given that X_n is in state s_i)

Transition Matrix

Transition matrix: $k \times k$ matrix whose (i,j)-th entry is the transition probability $P(X_{n+1} = s_j | X_n = s_i) = P_{i,j}$.

Ex 1: (Random Walker)



$$P = \begin{array}{c} \\ v_1 \\ v_2 \\ v_3 \\ v_4 \end{array} \begin{array}{c} \begin{array}{cccc} X_{n+1}\,v_1 & v_2 & v_3 & v_4 \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix} \end{array}$$

X_n

$$\rightarrow P(X_{n+1} = v_3 \mid X_n = v_2) = P_{2,3}$$

Transition Matrix: Properties

1. $P_{i,j} \geq 0$
2. $\sum_{j=1}^{n} P_{i,j} = 1$

Note: $(X_{n+1}=\beta_1), (X_{n+1}=\beta_2), \ldots, (X_{n+1}=\beta_K)$ are mutually exclusive & exhaustive set of events.

$$P(X_{n+1}=\beta_1 \mid X_n=\beta_i) + P(X_{n+1}=\beta_2 \mid X_n=\beta_i) + \cdots + P(X_{n+1}=\beta_K \mid X_n=\beta_i)$$

$$= \frac{P(\overset{A_1}{X_{n+1}=\beta_1}, X_n=\beta_i) + \cdots + P(\overset{A_K}{X_{n+1}=\beta_K}, X_n=\beta_i)}{P(X_n=\beta_i)}$$

$$= \frac{P(X_n=\beta_i)}{P(X_n=\beta_i)} \equiv P(X) = 1$$

(i) $A_1, A_2, \ldots, A_K$ are mutually exclusive & exhaustive set of events.

(ii) X is any event

$$P(X) = P(X \cap S) = P(XA_1 + XA_2 + \cdots + XA_K)$$

$$= P(XA_1) + P(XA_2) + \cdots + P(XA_K)$$

$$= P(A_1)P(X|A_1) + \cdots + P(A_K)P(X|A_K)$$

Distribution of X_n

For construction of a Markov Chain (MC) $\{X_n : n = 0, 1, 2, \ldots\}$ we require:

1. **Initial distribution:** of X_0
2. **Transition matrix:** consisting of transition probabilities $P(X_{n+1} = j | X_n = i) = P_{i,j}$

Distribution of X_n

$X_0 \xrightarrow{P} X_1 \xrightarrow{P} X_2 \xrightarrow{P} \cdots$

Notation: Distribution of X_n

$$\mu^{(n)} = (\mu_1^{(n)}, \mu_2^{(n)}, \ldots, \mu_k^{(n)})$$
$$= (P(X_n = s_1), P(X_n = s_2), \ldots, P(X_n = s_k))$$

$\mu^{(0)}$: Initial distribution

$\mu^{(0)} P^{n+1} = \cdots = \mu^{(n-1)} PP = \mu^{(n)} P = \mu^{(n+1)}$

$\mu^{(n-1)} P^2$

Theorem

For a MC $\{X_n : n = 0, 1, 2, \ldots\}$ with state space $\{s_1, s_2, \ldots, s_k\}$, initial distribution $\mu^{(0)}$ and transition matrix P, we have:

$$\boxed{\mu^{(n)} = \mu^{(0)} P^n}$$

for $n = 0, 1, 2, \ldots$.

$$\boxed{\mu = \mu P}$$

Stationary Distribution

$n=1$ To show $\boxed{\mu^{(1)} = \mu^{(0)} P}$

$$\mu_j^{(1)} = P(X_1 = s_j) = \sum_{i=1}^{K} P(X_0 = s_i, X_1 = s_j)$$

[Note: $(X_0 = s_1)$, $(X_0 = s_2)$ ---- $(X_0 = s_K)$ are m.e & exhaustive set of events]

$$= \sum_{i=1}^{K} P(X_0 = s_i) P(X_1 = s_j | X_0 = s_i)$$

$$\mu_j^{(1)} = \sum_{i=1}^{K} \mu_i^{(0)} P_{i,j} = \left(\mu^{(0)} P\right)_j$$ - j-th element of the row vector $\mu^{(0)} P$

$$\Rightarrow \boxed{\mu^{(1)} = \mu^{(0)} P}$$

Use induction to complete the proof.

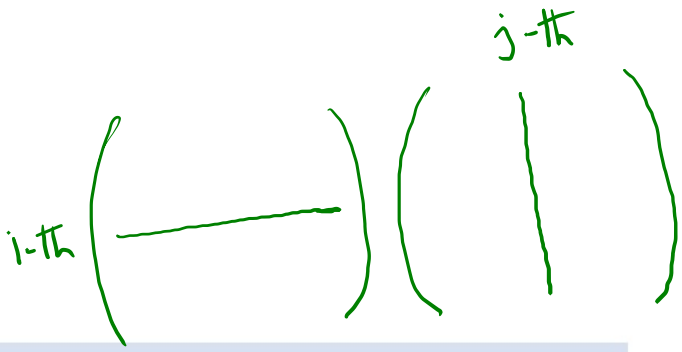
Interpretation of P^n

$P_{i,j}^{(n)} = n$-step transition probabilities

$$= P(X_n = j | X_0 = i)$$

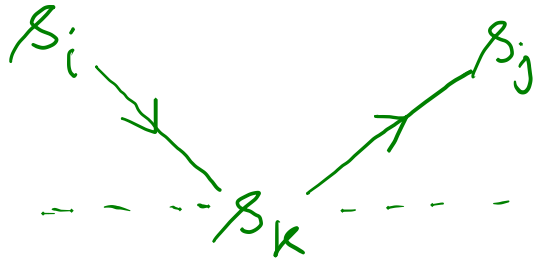
$$= P(X_{n+k} = j | X_k = i)$$

for $n \geq 0$, $i, j, k \geq 0$. In particular, $P_{i,j}^{(1)} = P_{i,j}$.



Chapman-Kolmogorov Equation

$$P_{i,j}^{(m+n)} = \sum_{k=0}^{\infty} P_{i,k}^{(n)} P_{k,j}^{(m)}, \quad \forall m, n \geq 0.$$



$$P_{i,j}^{(2)} = \sum_{k=0}^{k} P_{i,k}^{(1)} P_{k,j}^{(1)} = \sum_{k=0}^{k} P_{i,k} P_{k,j}$$

$$= (P\,P)_{(i,j)}$$

$$= P_{i,j}^2$$

$$P_{ij}^{(m+n)} = P\left(X_{n+m}=s_j \mid X_0=s_i\right)$$

$$= \sum_{k=0}^{\infty} P\left(X_{n+m}=s_j, X_n=s_k \mid X_0=s_i\right)$$

$\because \left\{ (X_n=s_k) : k=0,1,\ldots \right\}$ are m.e. & exhaustive set of events.

$$= \sum_{k=0}^{\infty} P\left(X_n=s_k \mid X_0=s_i\right) P\left(X_{n+m}=s_j \mid X_n=s_k, X_0=s_i\right)$$

$$= \sum_{k=0}^{\infty} P\left(X_n=s_k \mid X_0=s_i\right) P\left(X_{n+m}=s_j \mid X_n=s_k\right)$$

$$= \sum_{k=0}^{\infty} P_{ik}^{(n)} P_{kj}^{(m)}$$

Interpretation of P^n

Note

1. $P^{(n+m)} = P^{(n)} \cdot P^{(m)}$
2. $P^{(2)} = P^{(1+1)} = P^{(1)} \cdot P^{(1)} = P \cdot P = P^2$
3. By induction, $P^{(n)} = P^{(n-1+1)} = P^{n-1} \cdot P = P^n$

Transition Graph

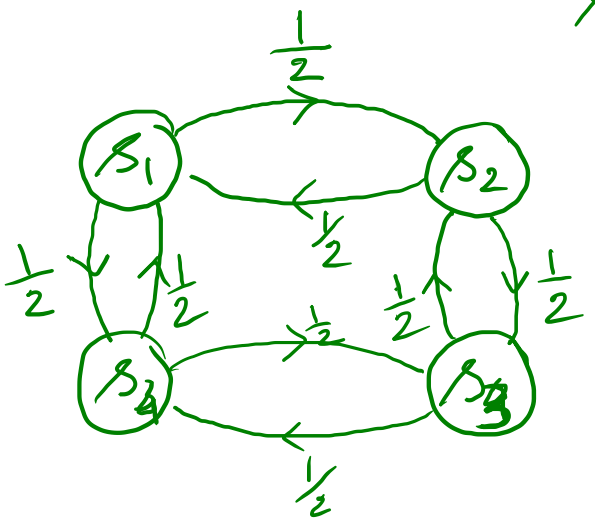
A useful way to picture a MC

Nodes: representing the states of the MC

Edges/Arrows between the nodes: representing transition probabilities

Ex 1: (Random Walker)

$$P = \begin{array}{c} \\ s_1 \\ s_2 \\ s_3 \\ s_4 \end{array} \begin{array}{c} \begin{array}{cccc} s_1 & s_2 & s_3 & s_4 \end{array} \\ \begin{bmatrix} 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \\ 0 & \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & 0 \end{bmatrix} \end{array}$$

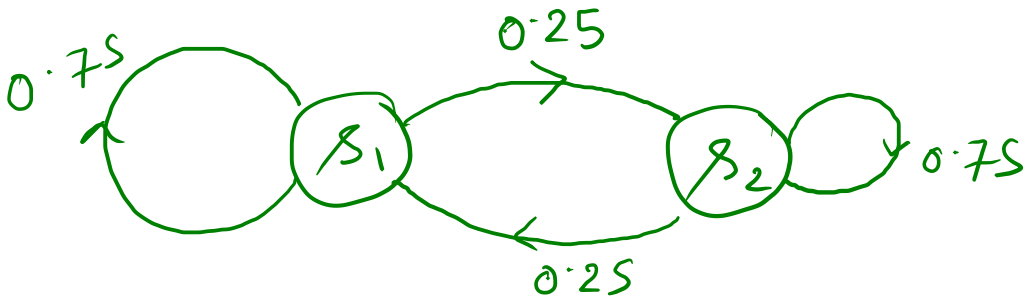


MC: Examples

Ex 2: (Gothenburg weather)

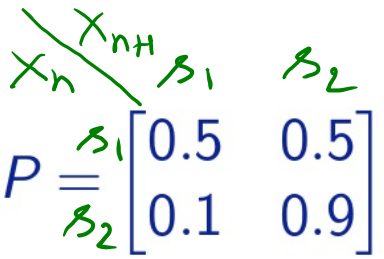
X_n \ X_{n+1}

$$P = \begin{array}{c} \\ S_1 \\ S_2 \end{array} \begin{array}{c} \begin{array}{cc} S_1 & S_2 \end{array} \\ \begin{bmatrix} 0.75 & 0.25 \\ 0.25 & 0.75 \end{bmatrix} \end{array}$$



Examples

Ex 3: (Los Angeles weather)



$$P = \begin{matrix} S_1 \\ S_2 \end{matrix} \begin{bmatrix} 0.5 & 0.5 \\ 0.1 & 0.9 \end{bmatrix}$$