Pattern Recognition and Machine Learning [2] [1] **Spring 2017**

Lecture 1: Stochastic Processes: Markov Chains

Lecturer: Jisung Lim *Scribes: Jisung Lim*

Scribes: Jisung Lim, *B.S. Candidate of Industrial Engineering in Yonsei University, South Korea.*

Contents

1.1 Probability Space and Random Variable

1.1.1 Probability Space

The underlying frame of reference for random variables or a stochastic process is a probability space. A sample space is given by

$$(\Omega, \mathcal{F}, P) \tag{1.1}$$

where

Ω	sample space (A entire set of possibleoutcomes)
$\mathcal{F}$	σ-algebra (event space)
P	probability measure.

The Kolmogorov's axioms and useful properties are as follows:

1. $A \in \mathcal{F}$, $P(A) \geq 0$, and $P(\Omega \setminus A) = 1 - P(A)$
2. $P(\Omega) = 1$, and $P(\emptyset) = 0$
3. $P(\cup_n A_n) = \sum_n P(A_n)$ for infinite collection of disjoint set A_n.
4. $P(A) \leq P(B)$, $A \subseteq B$
5. $P(A_n) \to P(A)$ if $A_n \uparrow A$ or $A_n \downarrow A$

1.1.2 Measurable space and measurable function

Definition 1.1.1. Suppose $(S, \mathcal{E})$, $(S', \mathcal{E}')$, are measurable spaces with associated σ-algebra and let $f : (S, \mathcal{E}) \to (S', \mathcal{E}')$ be a function on measurable sapce S. The function f is measurable if

$$f^{-1}(A) = \{x \in S : f(x) \in A\} \in \mathcal{E}, \quad \text{for each} \quad A \in \mathcal{E}' \tag{1.2}$$

□

That is, the set of all x's that f maps to A is in $\mathcal{E}$. It is fair to discuss about the operations between functions, especially the composition operations.

Theorem 1.1.1. Let $(S, \mathcal{E})$, $(S', \mathcal{E}')$, $(S'', \mathcal{E}'')$ be measurable spaces with associated σ-algebra. If $f : S \to S'$ and $g : S' \to S''$ are measurable and $f(S) \subseteq S'$ then the composition function $g \circ f : S \to S''$ is also measurable. □

1.1.3 Random variable

At first, in this material, we concerns classical real-valued random variables. A random variable X on a probability space $(\Omega, \mathcal{F}, P)$ is a measurable mapping from Ω to $\mathbb{R}$.

$$X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}) \tag{1.3}$$

The measurability of X ensures that $\mathcal{F}$ contains all set of the form

$$\{X \in B\} = \{\omega \in \Omega : X(\omega) \in B\}, \quad B \in \mathcal{B} \tag{1.4}$$

These are the type of events for which P is defined.

1.1.4 Measurability of random variable and σ-algebra

A measurability of the random variable ensures the probability measure P_X of X to be defined well on the measurable space $(S, \mathcal{E})$. This is because that the probability measure P_X on $(S, \mathcal{E})$ induced by random variable X is defined by

$$P_X(B) := P \circ X^{-1}(B) = P\{\omega \in \Omega : X(\omega) \in B\}, \quad B \in \mathcal{E} \tag{1.5}$$

so as to construct state space $(S, \mathcal{E}, P_X)$ from $(\Omega, \mathcal{F}, P)$.

Here, we may have question about where the σ-algebra comes from. That discussion is beyond the scope of this material, however, we assume that one usually constructs the σ-algebra $\mathcal{F}$ which is large enough so that the random variables of interest are measurable. For instance, if X, Y, and Z are of interest, one can let $\mathcal{F} = \sigma(X, Y, Z)$ be the smallest σ-algebra containing all sets of the form for X, Y, and Z, so that they are measurable.

1.1.5 Distribution function

All of the probability information of X in "isolation" (not associated with other random quantities on the probability space) is contained in its *distribution function*

$$F(x) = P\{X \leq x\}, \quad x \in \mathbb{R} \tag{1.6}$$

1.2 Random Elements and Stochastic Processes

1.2.1 Real to General

A unified way of discussing random vectors, stochastic processes and other random quantities is in terms of random elements. Suppose one is interested in a random element that takes values in a space S with a σ-field $\mathcal{E}$. A *random elements in* $\mathcal{E}$, defined on a probability space $(\Omega, \mathcal{F}, P)$, is a measurable mapping X from Ω to S. Here the X is also called an *S-vauled random variable.*

$$X : (\Omega, \mathcal{F}) \to (S, \mathcal{E}) \quad ; S\text{-valued random variable} \tag{1.7}$$

Here the space S may become a countable set (Markov chain), a Euclidean space (random vector), or a function space with a distance metric (stochastic process).

To accomodate these and other spaces as well, we adopt the standard convention that $(S, \mathcal{E})$ is a *Polish space*. That is, S is a metric space that is *complete* and *separable*; and $\mathcal{E}$ is the borel σ-algebra generated by the open sets.

The *probability distribution* of a random element X is S is the probability measure

$$F_X(B) = P\{\omega \in \Omega : X(\omega) \in B\} = P \circ X^{-1}(B), \quad B \in \mathcal{E} \tag{1.8}$$

1.2.2 Stochastic Process

Loosely speaking, a stochastic process is a collection of random variables (or random elements) defined on a single probability space. Hereafter, we will simply use the term "random elements" (which includes random variables), and let $(S, \mathcal{E})$ denote the Polish space where they reside.

A *discrete-time stochastic process* is a collection of random elements given by

$$X = \{X_n \in S : n \in \mathbb{N}\} \tag{1.9}$$

where the random variables X_n are in S and defined on a probability space $(\Omega, \mathcal{F}, P)$.

Here the following terms are as follows

$n \in \mathbb{N}$	time parameter
$(S, \mathcal{E})$	state space of the process
$X_n(\omega) \in S$	state at time n with the outcome ω

Note that, $X = \{X_n\}_{n\in\mathbb{N}}$ is also a random element in the infinite product space S^∞ with the product σ-algebra $\mathcal{E}^\infty$. Its distribution $P(X \in B)$ for $B \in \mathcal{E}^\infty$, is uniquely defined in terms of its *finite-dimensional distributions*, given by

$$P\{X_1 \in B_1, \dots, X_n \in B_n\}, \quad B_j \in \mathcal{E}, n \in \mathbb{N}. \tag{1.10}$$

1.2.3 Summary

- a stochastic process is a family of random variables or random elements defined on a probability space that contains all the probability information about the process.
- We will use the standard convention of suppressing the ω in random elements such as X_n or $X(t)$, and not displaying the underlying probability space $(\Omega, \mathcal{F}, P)$ unless it is essential for the exposition.

1.3 Markov Chains

A sequence of random variables $X_0, X_1, \dots$ with values in a countable set S is a Markov chain if at any time n, the future states (or values) $X_{n+1}, X_{n+2}, \dots$ depend on the history $X_0, X_1, \dots, X_n$ only through the present state X_n.

Markov chains are fundamental stochastic processes that have many diverse application. This is because a Markov chain represents any dynamical system whose states satisfy the recursion

$$X_n = f(X_{n-1}, Y_n), \quad n \in \mathbb{N} \setminus \{0\} \tag{1.11}$$

where $Y_1, Y_2, \ldots$ are iid and f is deterministic function. That is the new state X_n is simply a function of the last state X_{n-1} and an auxiliary random variable Y_n.

1.3.1 Introduction and definition

Let $(\Omega, \mathcal{F}, P)$ be a probability space and X_n's be S-valued random variables given by

$$X_n : (\Omega, \mathcal{F}) \to (S, \mathcal{E}). \tag{1.12}$$

where S be a countable set. We then define the discrete-time stochastic process on S given by

$$X = \{X_n \in S : n \in \mathbb{N}\} \tag{1.13}$$

where S is the state space of the process and n is the time parameter. Here the X is also a random variable which has the form of

$$X : (\Omega, \mathcal{F}) \to (S^\infty, \mathcal{E}^\infty). \tag{1.14}$$

For the measurable space $(S^\infty, \mathcal{E}^\infty)$ induced by X, we can find *finite-dimensional distributions* of the process given by

$$P\{X_0 = i_0, X_1 = i_1, \ldots, X_N = i_N\}, \quad i_0, \ldots, i_N \in S, N \in \mathbb{N} \tag{1.15}$$

These probabilites uniquely determines the probabilities of all events of the process. Consequently, two stochastic processes are equal in distribution if their finite-dimensional distributions are equal.

Various types of stochastic processes are defined by

- specifying the dependency among the variables that determine the finite-dimensional distributions, or
- specifying the manner in which the process evolves over time.

A Markov chain is defined as follows.

Definition 1.3.1. A stochastic process $X = X_n \in S : n \in \mathbb{N}$ on a countable set S is a *Markov Chain* if, for any $i, j \in S$ and $n \in mathbbN$,

- Markov property; $P\{X_{n+1} = j | X_0, \ldots, X_n\} = P\{X_{n+1} = j | X_n\}$
- Time homogenuity: $P\{X_{n+1} = j | X_n = i\} = p_{ij}$

where p_{ij} is "time-homogeneous" transition probabilities. □

1.3.2 Probabilities of sample path

Here the *transition probability* p_{ij} is a probability that the Markov chain jumps from state i to state j and the matrix $\mathsf{P} = (p_{ij})$ is the *transition matrix* of the chain. This transition probabilities satisfy followings:

1. Let any $i \in S$,

$$\sum_{j \in S} p_{ij} = \sum_{j \in S} P\{X_{n+1} = j | X_n = i\} = 1. \tag{1.16}$$

2. Let $\alpha_i = P\{X_0 = i\}$ be an *initial distribution*. Then, for any $i_0, i_1, \ldots, i_n \in S$ and $n \in \mathbb{N}$,

$$P\{X_0 = i_0, \ldots, X_n = i_n\} = \alpha_{i_0} p_{i_0,i_1} \cdots p_{i_{n-1},i_n}. \tag{1.17}$$

3. The probability of the Markov chain having a sample path in a subset $\mathcal{P} \subset S^{n+1}$ up to time n is given by

$$P\{(X_0, \ldots, X_n) \in \mathcal{P}\} = \sum_{(X_0,\ldots,X_n) \in \mathcal{P}} P\{X_0 = i_0\} p_{i_0,i_1} \cdots p_{i_{n-1},i_n}, \tag{1.18}$$

   or equivalently,

$$P\{(X_0, \ldots, X_n) \in \mathcal{P} | X_0 = i_0\} = \sum_{(i_0,X_1,\ldots,X_n) \in \mathcal{P}} p_{i_0,i_1} \cdots p_{i_{n-1},i_n}. \tag{1.19}$$

4. From the result of third property, we can find the k-step distribution given by

$$P\{X_k = i_k | X_0 = i_0\} = \sum_{i_1 \in S} P_{i_0 i_1} \sum_{i_2 \in S} P_{i_1 i_2} \cdots \sum_{i_{k-1} \in S} P_{i_{k-1} i_k} \tag{1.20}$$

5. From the transition matrix $\mathsf{P} = (p_{ij})$ and its nth product P^n, $n \geq 0$.

   (a) By definition, $\mathsf{P}^0 = 0$ and $\mathsf{P}^n = \mathsf{P}^{n-1}\mathsf{P}$, for $n \geq 1$.

   (b) Let p_{ij}^n denote the (i, j)th entry of P^n.

6. n-step probabilities.

$$P\{X_n = j | X_0 = i\} = p_{ij}^n \tag{1.21}$$

   and letting $\alpha = (\alpha_i)$ be a row vector of initial distributions,

$$P\{X_n = j\} = (\alpha \mathsf{P}^n)_j \tag{1.22}$$

1.3.3 Construction of Markov Chain

This section addresses the following questions. Is there a general framework for constructing or identifying Markov chains? Is there a Markov chain associated with any transition matrix? If so, how is it constructed? Here is a general construction of a Markov chain via a general recursive equation.

Suppose $\{X_n : n \in \mathbb{N}\}$ is a stochastic process on S of the form

$$X_n = f(X_{n-1}, Y_n), \quad n \geq 1, \tag{1.23}$$

where $f : S \times S' \to S$ and $Y_1, Y_2, \dots$ are S'-*valued iid random variables* and S' is independent of X_0. Then X_n is a Markov chain with transition probabilities $p_{ij} = P\{f(i, Y_1) = j\}$, which is given by

$$\begin{aligned} P\{X_{n+1} = j | X_0, \dots X_{n-1}, X_n = i\} &= P\{f(i, Y_{n+1}) = j | X_0, \dots, X_{n-1}, X_n = i\} \\ &= P\{f(i, Y_{n+1}) = j\} = p_{ij} \end{aligned} \tag{1.24}$$

1.3.4 Stationary distribution

Definition 1.3.2. A probability measure π on S is a *stationary distribution* for the Markov chain X_n (or for P) if

$$\pi_i = \sum_j \pi_j p_{ji}, \quad i \in S, \tag{1.25}$$

or equivalently in matrix notation,

$$\pi = \pi \mathsf{P} \tag{1.26}$$

where $\pi = (\pi_i : i \in S)$ is a low vector. In general, η with $\sum_i \eta_i \leq \infty$ that satisfies $\eta = \eta \mathsf{P}$ is an *invariant measure* for P. □

Definition 1.3.3. A stochastic process $\{X_n : n \geq 0\}$ on a general state space is *stationary* if, for any $n \geq 0$,

$$(X_n, \dots, X_{n+k}) \overset{\mathrm{d}}{=} (X_0, \dots, X_k), \quad k \geq 1 \tag{1.27}$$

That is, the finite-dimensional distributions of X_n remain the same if the time is shifted by any amount n. A stationary process is sometimes said to be a process that is in *equilibrium* or in *steady state.* □

Theorem 1.3.1. The following statements are equivalent for the Markov chain X_n.

1. X_n is stationary.
2. $X_n \overset{\mathrm{d}}{=} X_0, \quad n \geq 1$.
3. The distribution of X_0 is a stationary distribution.

□

1.3.5 Ergodicity

Theorem 1.3.2. The Markov chain is called ergodic if it is irreducible, and its states are positive recurrent and aperiodic. □

Theorem 1.3.3. (Stationary Distribution) An irreducible Markov chain X_n has a positive stationary distribution if and only if all of its states are positive recurrent. In that case, the stationary distribution is unique. In that case, the stationary distribution is unique and has the following form: For any fixed $i \in S$,

$$\pi_j = \frac{E_i\left[\sum_{n=0}^{\tau_i - 1} \mathbf{1}(X_n = j)\right]}{\mu_i} \quad j \in S \tag{1.28}$$

where $\tau_i = \min\{n \in \mathbb{N} \setminus \{0\} : X_n = i\}$ and $\mu_i = E_i[\tau_i]$. Or equivalently, $\pi_j = 1/\mu_j$, $j \in S$. □

Theorem 1.3.4. An irreducible aperiodic Markov chain is ergodic if and only if it has a stationary distribution. In this case, the stationary distribution is positive and has the form shown in Theorem 1.3.3. □

Theorem 1.3.5. If a Markov chain is ergodic, then its stationary distribution is its limiting distribution, which is positive. □

Theorem 1.3.6. For an irreducible, aperiodic Markov chain, the following statements are equivalent.

1. The chain is ergodic.
2. The chain has a stationary distribution.
3. The chain has a limiting distribution.

□

1.3.6 Strong Laws of Large Number

Theorem 1.3.7. (Classical SLLN) If $Y_1, Y_2, \ldots$ are iid random variables with a mean that may be infinite, then

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} Y_i = E[Y_1] \quad a.s. \tag{1.29}$$

□

Theorem 1.3.8. The average time between visits to state i is

$$\lim_{n\to\infty} \frac{1}{\tau_n} = \frac{1}{\pi_i} \tag{1.30}$$

The average number of visits to state i is

$$\lim_{n\to\infty} \frac{1}{N_i(n)} = \pi_i \tag{1.31}$$

□

Theorem 1.3.9. For the ergodic Markov chain X_n with stationary distribution π and any $f : S \to \mathbb{R}$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \sum_i f(i)\pi_i \tag{1.32}$$

□

1.3.7 Markov Chain Monte Carlo

There are a variety of statistical problems in which one uses Monte Carlo simulations for estimating expectations of the form

$$\mu = \sum_{i \in S} g(i)\pi_i \tag{1.33}$$

where π is a specified probability measure and $g : S \to \mathbb{R}$. A standard Markov chain Monte Carlo approach is to construct an ergodic Markov transition matrix whose stationary distribution is π. Then for a sample path $X_1, \ldots, X_n$ of the chain, an estimator for μ is

$$\hat{\mu}_n = \frac{1}{n} \sum_{i=1}^{n} g(X_i) \tag{1.34}$$

By theorem 1.3.9 for Markov chains, $\hat{\mu} \to \mu$, and so $\hat{\mu}$ is a consistent estimator of μ.

References

[1] Babatunde A. Ogunnaike. *Random Phenomena: Fundamentals of Probability and Statistics for Engineers*. CRC Press, 2009.

[2] Georgia Institute of Technology Richard Serfozo. *Basics of Applied Stochastic Processes*. Springer-Verlag Berlin Heidelberg, 2009.