Probability for Deep Learning
Part 4

Johar M. Ashfaque

I. INTRODUCTION

We will consider sequences of dependent random variables. For example, daily temperatures will form a sequence of time-ordered random variables and clearly the temperature on day one is not independent of the temperature on day two.

A stochastic process $\{X_t : t \in T\}$ is a collection of random variables.The variables X_t take values in some set $\mathcal{X}$ called the state space. The set T is called the index set and for our purposes can be thought of as time. The index set can either be discrete or continuous.
Example 1. Let $\mathcal{X} = \{\text{sunny}, \text{cloudy}\}$. A typical sequence might be

$$\text{sunny}, \text{sunny}, \text{cloudy}, \text{sunny}, \text{cloudy}, \cdots.$$

This process has a discrete state space and a discrete index set.
Example 2. A sequence of IID random variables can be written as $\{X_t : t \in T\}$ where $T = \{1, 2, 3, ...\}$. Hence a sequence of IID random variables is an example of a stochastic process.

If $X_1, ..., X_n$ are random variables then we can write the joint density as

$$f(x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i|\text{past}_i)$$

where past_i refers to all the variables before X_i.

II. MARKOV CHAINS

The simplest stochastic process is a Markov chain in which the distribution of X_t depends only on X_{t-1}. We will assume that the state space is discrete $\mathcal{X} = \{1, ..., N\}$ and that the index set is $T = \{0, 1, 2, ...\}$.

Definition II.1 *The process $\{X_n : n \in T\}$ is a Markov chain if*

$$\mathbb{P}(X_n = x|X_0, ..., X_{n-1}) = \mathbb{P}(X_n = x|X_{n-1})$$

for all n and for all $x \in \mathcal{X}$.

For the Markov chain, the joint probability is

$$f(x_1, ..., x_n) = f(x_1)f(x_2|x_1)...f(x_n|x_{n-1}).$$

A. Transition Probabilities

The key quantities of a Markov chain are the probabilities of jumping from one state into another.

Definition II.2 *We call*

$$\mathbb{P}(X_{n+1} = j|X_n = i)$$

the transition probabilities. If the transition probabilities do not change with time, we call the chain homogeneous. In this case, we define

$$p_{ij} = \mathbb{P}(X_{n+1} = j | X_n = i).$$

The matrix $\mathbf{P}$ *whose* (i,j) *element is* p_{ij} *called the transition matrix.*

We will only be interested in homogeneous Markov chains. Notice that $\mathbf{P}$ has two properties

- $p_{ij} \geq 0$
- $\sum_i p_{ij} = 1$

Each row is a probability mass function. A matrix with these properties is called a stochastic matrix. Let

$$p_{ij}(n) = \mathbb{P}(X_{m+n} = j | X_m = i)$$

be the probability of going from state i to state j in n steps. Let $\mathbf{P}_n$ be the matrix whose (i,j) element is $p_{ij}(n)$. These are called the n-step transition probabilities.

Theorem II.3 *(The Chapman-Kolmogorov Equations) The* n*-step transition probabilities satisfy*

$$p_{ij}(m+n) = \sum_k p_{ik}(m) p_{kj}(n).$$

This statement of the theorem is nothing more than the equation for matrix multiplication. Hence

$$\mathbb{P}_{m+n} = \mathbb{P}_m \mathbb{P}_n.$$

1) Transition Matrix: $\mathbb{P}(i,j) = \mathbb{P}(X_{n+1} = j \| X_n = i)$.
2) n-step Matrix, $\mathbb{P}_n(i,j) = \mathbb{P}(X_{m+n} = j \| X_m = i)$.

TABLE I. Summary

B. States

The states of a Markov chain can be classified according to various properties.

Definition II.4 *We say that* i *reaches* j *or* j *is accessible from* I *if* $p_{ij}(n) > 0$ *for some* n *and we write* $i \to j$*. If* $i \to j$ *and* $j \to i$ *then we write* $i \leftrightarrow j$ *and we say that* i *and* j *communicate.*

Theorem II.5 *The communication relation satisfies the following properties:*

1. $i \leftrightarrow i$.
2. *If* $i \leftrightarrow j$ *then* $j \leftrightarrow i$.
3. *If* $i \leftrightarrow j$ *and* $j \leftrightarrow k$ *then* $i \leftrightarrow k$.
4. *The set of states* $\mathcal{X}$ *can be written as a disjoint union of classes* $\mathcal{X} = \mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots$ *where two states* i *and* j *communicate with each other if and only if they are in the same class.*

If all states communicate with each other then the chain is called irreducible. A set of states is closed if once you enter that set of states you never leave. A closed set consisting of a single state is called an absorbing state.

Example. Let $\mathcal{X} = \{1, 2, 3, 4\}$ and

$$\mathbf{P} = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} & 0 & 0 \\ \frac{2}{3} & \frac{1}{3} & 0 & 0 \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

The classes are $\{1, 2\}$, $\{3\}$ and $\{4\}$. State 4 is an absorbing state.

Theorem II.6 *A state i is recurrent if and only if*

$$\sum_n p_{ii}(n) = \infty.$$

A state i is transient if and only if

$$\sum_n p_{ii}(n) < \infty.$$

Theorem II.7 *Some facts.*
- *If state i is recurrent and $i \leftrightarrow j$ then j is recurrent.*
- *If state i is transient and $i \leftrightarrow j$ then j is transient.*
- *A finite Markov chain must have at least one recurrent state.*
- *The states of a finite, irreducible Markov chain are all recurrent.*

III. POISSON PROCESSES

One of the most studied and useful stochastic processes is the Poisson process. It arises when we count occurrences of events over time. For example, traffic accidents, radioactive decay etc. As the name suggests the Poisson process is related to the Poisson distribution.

Definition III.1 *A Poisson process is a stochastic process*

$$\{X_t : t \in [0, \infty)\}$$

with state space

$$\mathcal{X} = \{0, 1, 2, ...\}$$

such that
- $X(0) = 0$.
- *For any $0 = t_0 < t_1 < t_2 < \cdots < t_n$, the increments*

  $$X(t_1) - X(t_0), X(t_2) - X(t_1), \cdots, X(t_n) - X(t_{n-1})$$

  are independent.
- *There is a function $\lambda(t)$ called an intensity function. This is to say the number of events in any interval of length t is a Poisson random variable with parameter (or mean) λt.*

Definition III.2 *A Poisson process with intensity function $\lambda(t) \equiv \lambda$ for some $\lambda > 0$ is called a homogeneous Poisson process with rate λ. In this case*

$$X(t) \sim Poisson(\lambda t).$$