DATA ANALYTICS

Unit 5: Introduction to Stochastic models and Markov processes (first order)

Bharathi R

Department of Computer Science and Engineering

DATA ANALYTICS

Unit 5:Introduction to Stochastic models and Markov processes (first order)

Bharathi R

Department of Computer Science and Engineering

Introduction Stochastic Process

- Stochastic models are powerful tools which can be used for solving problems which are dynamic in nature, that is, the values of the random variables change with time.

- Stochastic process is defined as a collection of random variables $\{X_n, n \geq 0\}$ indexed by time (however, index can be other than time).

- The value (cash flow) that the random variable X_n can take is called the **state of the stochastic process at time n**.

- The set of all possible values the random variable can take is called the **state space**.

Outline

The following are different stochastic process models classified based on certain properties:

1. Poisson Process
2. Markov Process and Markov Chain (MC)
3. Markov Decision Process
4. Partially Observable Markov Decision Process
5. Semi-Markov Process
6. Random Walk
7. Brownian Motion Process
8. Auto-Regressive and Moving Average Processes

In this chapter, we will be discussing **Poisson process, compound Poisson process, Markov chains, and Markov decision process (MDP) with applications.**

1.Poisson Process : Examples

Generally we would like to count the number of events that occur over a period of time. Following are few examples of counting process:

1. Retail stores would like to predict footfall (number of customers visiting the store) over a period of time.
2. Call centres would like to predict the number of calls they receive over a period of time.
3. Number of customer arrivals at banks, airports, restaurants, and any service centres.
4. Demand for spare parts of capital equipment caused due to failure of parts over a period of time.
5. Number of insurance claims received at an insurance company.

1.Poisson Process

Homogeneous Poisson Process (HPP) is a stochastic counting process $N(t)$ with the following properties:

$N(0) = 0$, that is the number of events by time t = 0 is zero.

$N(t)$ has independent increments. That is if $t_0 < t_1 < t_2 < \ldots < < t_n$, then $N(t_1) – N(t_0)$, $N(t_2) – N(t_1)$, …, $N(t_n) – N(t_{n–1})$ are independent.

The number of events by time t, N(t), follows a Poisson distribution, that is

$$P[N(t) = n] = \frac{e^{-\lambda t} \times (\lambda t)^n}{n!}$$

1. Poisson Process

Cumulative distribution of number of events by time t in a Poisson process is given by

$$P[N(t) \le n] = \sum_{i=0}^{n} P[N(t) = i] = \sum_{i=0}^{n} \frac{e^{-\lambda t} \times (\lambda t)^i}{i!}$$

The mean, $E[N(t)]$, and variance, $\text{Var}[N(t)]$, of a Poisson process $N(t)$ are given by

$$E[N(t)] = \lambda t$$
$$\text{Var}[N(t)] = \lambda t$$

1.Poisson Process

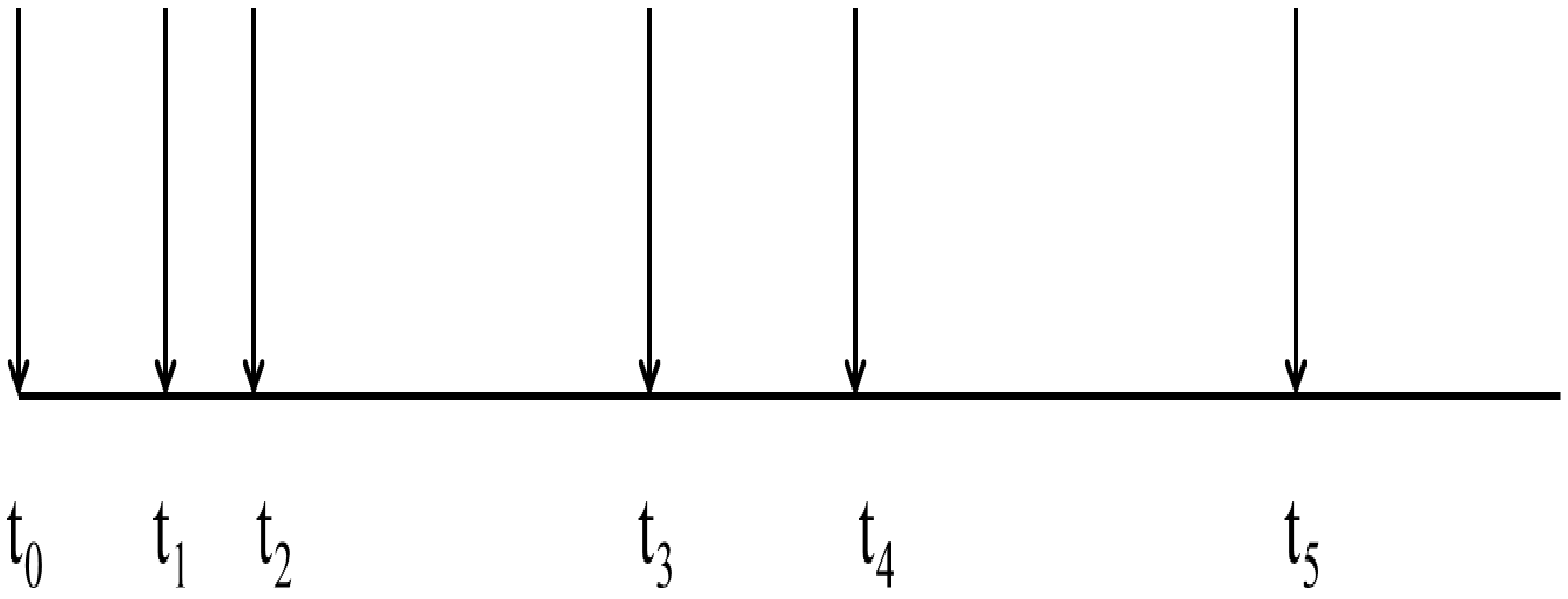

Figure above shows a Poisson process of events in which $(t_1 – t_0)$, $(t_2 – t_1)$, $(t_3 – t_2)$ are time between events.

In the case of Poisson process, the time between events follows an exponential distribution with parameter l, that is the time between events have a density function $f(t) = \lambda e^{-\lambda t}$ and cumulative distribution function $F(t) = 1 - e^{-\lambda t}$.

Example

Johny Sparewala (JS) is a supplier of aircraft flight control system spares based out of Mumbai, India. The demand for hydraulic pumps used in the flight control system follows a Poisson process. Sample data (50 cases) on time between demands (measured in number of days) for hydraulic pumps are shown in Table 16.1

TABLE 16.1 Time between demands (in days) for hydraulic pumps

104	90	45	32	12	6	30	23	58	118
80	12	216	71	29	188	15	88	88	94
63	125	108	42	77	65	18	25	30	16
92	114	151	10	26	182	175	189	14	11
83	418	21	19	73	31	175	14	226	8

Example Continued

(a) Calculate the expected number of demand for hydraulic pump spares for next two years.

(b) Johny Sparewala would like to ensure that the demand for spares over next two years is met in at least 90% of the cases from the spares stocked (called fill rate) since lead time to manufacture a part is more than 2 years. Calculate the inventory of spares that would give at least 90% fill rate.

Solution

(a) To calculate the expected number of demand for spares for two years, we have to estimate the parameter λ of the Poisson distribution. The maximum likelihood estimate of λ is given by

$$\hat{\lambda} = \frac{1}{\frac{1}{n}\sum_{i=1}^{n} X_i} = 0.0125$$

where X_i is the time between failure of i^{th} case and $\frac{1}{n}\sum_{i=1}^{n} X_i$ is the mean time between failure.

The expected number of demand for spares, $E[N(t)]$, for 2 years (2 × 365 days) is given by

$$E[N(t)] = E[N(2 \times 365)] = \hat{\lambda} \times t = 0.0125 \times 2 \times 365 = 9.125$$

Solution

(b) To ensure that the demand for spares is met 90% of the time, we have to calculate smallest k such that

$$\sum_{i=0}^{k} \frac{e^{-\hat{\lambda}t} \times (\hat{\lambda}t)^i}{i!} \geq 0.90$$

Table 16.2 shows density and cumulative distribution function values of Poisson process for different values of k.

TABLE 16.2 Poisson density and distribution function for different values of k

k	Poisson Density	Cumulative	k	Poisson Density	Cumulative
0	0.0001	0.0001	11	0.0996	0.7907
1	0.0010	0.0011	12	0.0758	0.8665
2	0.0045	0.0056	13	0.0532	0.9197
3	0.0138	0.0194	14	0.0347	0.9543
4	0.0315	0.0509	15	0.0211	0.9754
5	0.0574	0.1083	16	0.0120	0.9875
6	0.0873	0.1956	17	0.0065	0.9939
7	0.1138	0.3095	18	0.0033	0.9972
8	0.1298	0.4393	19	0.0016	0.9988
9	0.1316	0.5709	20	0.0007	0.9995
10	0.1201	0.6911	21	0.0003	0.9998

Solution

Smallest value of k for which the cumulative probability is greater than 0.90 is 13. That is, JS should stock 13 spares to ensure that they meet demand for spares in 90% of the cases over a two-year period. The probability density function of Poisson distribution with mean 9.125 is shown in Figure 16.2.

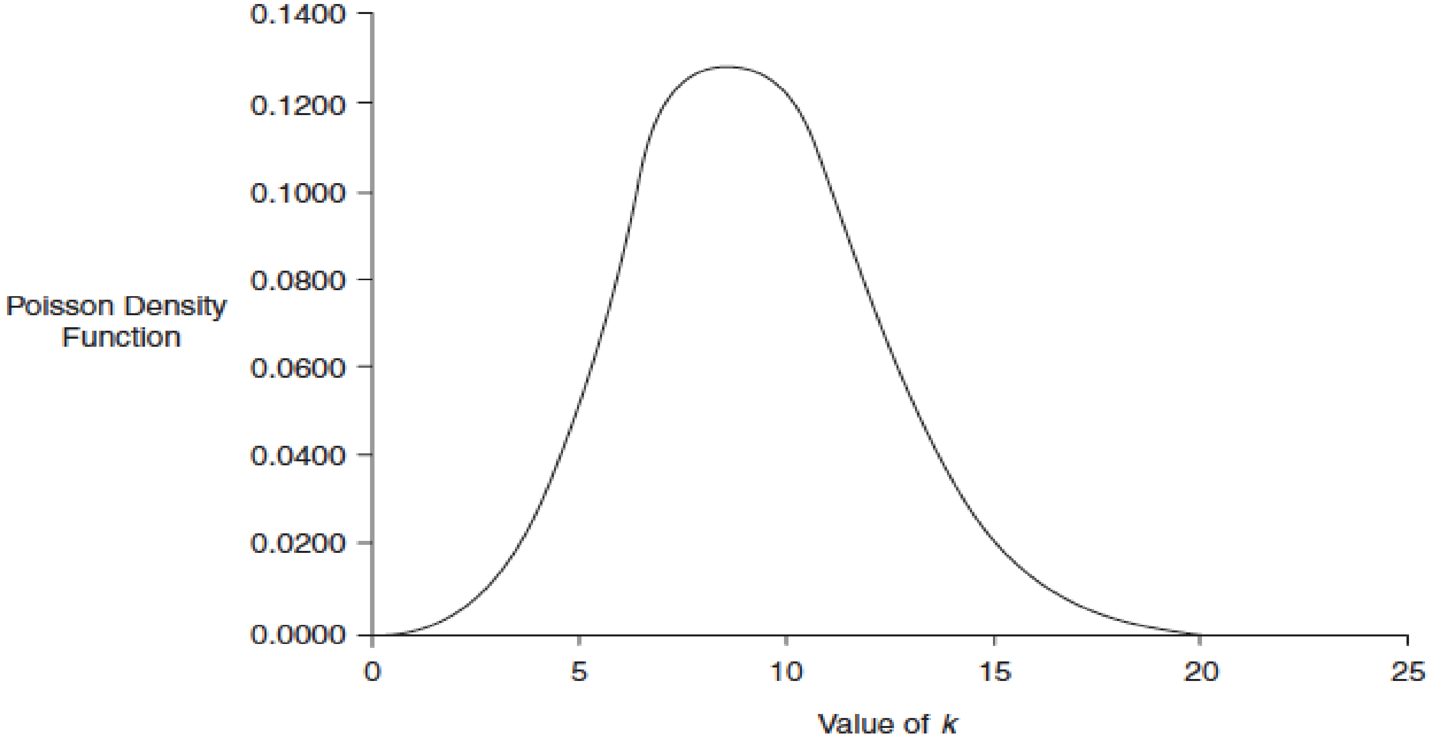

FIGURE 16.2 Poisson process density function.

2. Compound Poisson Process

Compound Poisson process is a stochastic process $X(t)$ where the arrival of events follows a Poisson process and each arrival is associated with another independent and identically distributed random variable Y_i.

Compound Poisson process $X(t)$ is a continuous-time stochastic process defined as

$$X(t) = \sum_{k=1}^{N(t)} Y_k$$

where $N(t)$ is a Poisson process with mean λt and Y_i are independent and identically distributed random variables with mean $E(Y_i)$ and variance $\text{Var}(Y_i)$.

Example

Customers arrive at an average rate of 12 per hour to withdraw money from an ATM and the arrivals follow a Poisson process. The money withdrawn are independent and identically distributed with mean and variance INR 4200 and 2,50,000, respectively. If the ATM has INR 6,00,000 cash, what is the probability that it will run out of cash in 10 hours?

Solution

Solution:

The mean and standard deviation of the compound Poisson process $X(t)$ can be calculated as described below:

Mean of compound Poisson process is

$$\mu_{X(t)} = \lambda t \times E(Y_i) = 12 \times 10 \times 4200 = 5{,}04{,}000$$

Variance of compound Poisson process is

$$\sigma^2_{X(t)} = \lambda t \times (\mathrm{Var}(Y_i) + [E(Y_i)]^2) = 12 \times 10 \times (250000 + 4200^2) = 21468 \times 10^5$$

Standard deviation of compound Poisson process is

$$\sigma_{X(t)} = \sqrt{\sigma^2_{X(t)}} = \sqrt{21468 \times 10^5} = 46333.57$$

Probability that the cash withdrawal will exceed INR 6,00,000 is given by

$$P(X(t) \geq 6{,}00{,}000) = P\left(Z \geq \frac{6{,}00{,}000 - 504000}{46333.57}\right) = P(Z \geq 2.0719) = 0.0191$$

That is, there is approximately 2% chance that the ATM will run out of cash in 10 hours.

3. Markov Chains

The condition $P\left[X_{n+1}=j \mid X_0=i_0, X_1=i_1, \ldots, X_n=i\right]=P\left[X_{n+1}=j \mid X_n=i\right]$
is called Markov property named after the Russian mathematician *A A Markov*.

If the state space *S* is discrete then the stochastic process

$$\{X_n, n = 0, 1, 2, \ldots\}$$

that satisfies the condition is called a Markov chain

One-Step Transition Probabilities of Markov Chain

Let $\{X_n, n = 0, 1, 2, \ldots\}$ be a Markov chain with state space S. Then the conditional probability

$$P\left[X_{n+1} = j \mid X_0 = i_0, X_1 = i_1, \ldots, X_n = i\right] = P\left[X_{n+1} = j \mid X_n = i\right] = P_{ij}$$

is called the one-step transition probability. P_{ij} gives conditional probability of moving from state i to stage j in one period.

One-Step Transition Probabilities of Markov Chain

P_{ij} gives conditional probability of moving from state i to stage j in one period.
One-step transition probabilities between all states in the state space are expressed in the form of one-step transition probability matrix as shown below

$\mathbf{P} = \mathbf{P_{ij}} =$

	1	2	…	n
1	P_{11}	P_{12}	…	P_{1n}
2	P_{21}	P_{22}	…	P_{2n}
…	…	…	…	…
n	P_{n1}	P_{n2}	…	P_{nn}

m-step transition probability

An *m*-step transition probability in a Markov chain is given by

$$P_{ij}^{(m)} = P(X_{n+m} = j \mid X_n = i)$$

The m-step transition probability can be written as

$$P_{ij}^{(m)} = \sum_{r=1}^{n} P_{ir}^{k} \times P_{rj}^{(m-k)}, \quad 0 < k < m$$

Estimation of One-Step Transition Probabilities of Markov Chain

Transition probabilities of a Markov chain are estimated using maximum likelihood estimate (MLE) from the transition data (Anderson and Goodman, 1957).

The MLE estimate of the transition probability P_{ij} (probability of moving from state i to state j in one step) is given by

$$\hat{P}_{ij} = \frac{N_{ij}}{\sum_{k=1}^{m} N_{ik}}$$

where N_{ij} is number of cases in which $X_n = i$ (state at time n is i) and $X_{n+1} = j$ (state at time $n + 1$ is j).

Hypothesis Tests for Markov Chain: Anderson Goodman Test

The null and alternative hypotheses to check whether the sequence of random variables follows a Markov chain is stated below

H_0: The sequences of transitions (X_1, X_2, …, X_n) are independent (zero-order Markov chain)

H_A: The sequences of transitions (X_1, X_2, …, X_n) are dependent (first-order Markov chain)

The corresponding test statistic is

$$\chi^2 = \sum_{i=1}^{m} \sum_{j=1}^{n} \left(\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \right)$$

where

O_{ij} = Observed number of transitions from state i to state j in one period.

E_{ij} = Expected number of transitions from state i to state j assuming independence.

Testing Time Homogeneity of Transition Matrices: Likelihood Ratio Test

Anderson and Goodman (1957) suggested a likelihood ratio test for checking whether the transition probability matrices are time homogeneous. The null and alternative hypotheses of the likelihood ratio tests are

$$H_0: P_{ij}(t) = P_{ij}, \; t = 1, 2, 3, 4, \text{ and } 5$$
$$H_A: P_{ij}(t) \neq P_{ij}, \; t = 1, 2, 3, 4, \text{ and } 5$$

The test statistic is a likelihood test ratio statistic and is given by (Anderson and Goodman, 1957):

$$\lambda = \prod_{t} \prod_{i,j} \left[\frac{\hat{P}_{ij}}{\hat{P}_{ij}(t)} \right]^{n_{ij}(t)}$$

Summary

1.Most problems in analytics are dynamic in nature and thus require collection of random variables to model the problem.

2. Stochastic process is a collection of random variables usually indexed by time t and used while modelling problems that are not independent and identically distributed.

3. Poisson process is a counting process that is used in decision-making scenarios such as capacity planning and spare parts demand forecasting. Compound Poisson process can be used to study problems such as cash replenishments at ATMs, total insurance claims, etc.

4. Markov chain is one of the most powerful models in analytics with applications across industry sectors. Google's PageRank algorithm is based on Markov chain.

5. Asset availability, market share, customer retention probability, and customer lifetime value are few applications of Markov chain in analytics.

References

Text Book:

Chapter 16.1 -16.4.4 “Business Analytics,

The Science of Data-Driven Decision Making”, by U. Dinesh Kumar, Wiley 2017

THANK YOU

Prof Bharathi R
Department of Computer Science
rbharathi@pes.edu