Markov Chains: Fundamentals and Applications in Machine Learning

Burak Yuva
EE410 Sabanci University - Term Project
burakyuva@sabanciuniv.edu

Abstract

The present report explores Markov Chains (MC) within the context of information theory and machine learning. The report aims to: (i) motivate fundamentals of MC with clear examples and formal definitions, (ii) provide key mathematical and practical properties of MC, and (iii) illustrate applicability of MC within the machine learning literature through the Markov Decision Processes (MDP) framework for reinforcement learning, and the Hidden Markov Models (HMM) framework for speech and language processing.

1 Introduction

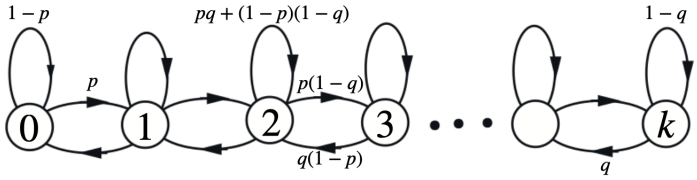
1.1 Markov Processes vs. Markov Chains

Markov Processes (MP) are a fundamental class of stochastic processes wherein the future is independent of the past given the present. Let's consider the following abstract definition for a function that describes a state change: $X^{'} = f(X)$ where X is the old state and X' is the new state. For example, Newton's laws of motion from classical mechanics describe such state changes driven by force, where each state can be position, velocity, or acceleration. MP, on the other hand, considers the evolution of a system in the presence of some random noise ϵ: $X^{'} = f(X, \epsilon)$.

Markov Chains (MC), deriving directly from MP, are discrete-time stochastic models with applications in Bayesian statistics, finance, information theory, and machine learning.

1.2 A First Look at Markov Chains

Compared to other simpler stochastic processes which have memoryless states (e.g. Bernouilli Processes, Poisson Processes, etc.), Markov Chains capture dependencies between different states. A simple and intuitive example to describe MC is the *checkout counter model*. In this model, customers arrive and get in queue, get served one at a time, and depart the queue once they are served. The key properties of this model are: (i) functions in discrete time steps: $t = 0, 1, \ldots$, (ii) customer arrivals are modeled with $B(p)$, (iii) customer service times are modeled with $G(q)$, (iv) X_t denotes the state at step t, and is the number of customers in queue at time t, and let's assume that the maximum capacity of the counter is k. A graphical representation with attached transition probabilities is as follows:



A typical question we would like to ask in such a scenario is: What is the probability that a customer is departing at time t? In order to answer this, we have to write out steps and transitions. There exists 3 transitions for all states except beginning and ending states: (i) a customer arrives and no customer departs which increases the state by one with probability $p(1-q)$, (ii) a customer departs and no customer arrives which decreases the state by one with probability $q(1-p)$, and (iii) either a customer arrives and a customer departs simultaneously or no arrival or departure happens with probability $pq + (1-p)(1-q)$.

1.3 Markov Chains: Formalism

Now that we have covered an explicit discrete time finite state MC, we can build up towards a more generalized definition. We define X_n as the state after n transitions with the following assumptions: (i) it belongs to a finite set, and (ii) X_0 is either given or random.

The most fundamental property of a MC is the **Markov Property** (also known as **Markov Assumption**), and it declares that the past does not matter for future predictions given the current state. It can be formalized as follows:

$$p_{i,j} = P(X_{n+1} = j | X_n = i, x_{n-1}, ..., x_0) = P(X_{n+1} = j | X_n = i)$$

where the current state is i, and $p_{i,j}$ denote the transition probability from state i to j.

For this proposal to work reasonably well, the state has to capture a sufficient degree of task-dependent information. Therefore, a typical model specification for a MC consist of (i) identification of possible states, (ii) identification of possible transitions, and (iii) estimation of transition probabilities.

1.3.1 n-step Transition Probabilities

If we have a MC system which has an initial state of i, the state occupancy probabilities n steps into the future is given by:

$$r_{ij}(n) = P(x_n = j | X_0 = i)$$

We can convert this to *recursive representation*:

$$r_{ij}(n) = \sum_{k=1} mr_{ik}(n-1)p_{kj} State occupancy probabilities, given initial state$$

i:$r_{ij}(n)$ = P($X_n = j | X_0 = i$)

Key recursion $r_{ij}(n) =$

A Appendix

A.1 Notation

$B(p)$: Bernouilli distribution with success probability of p, i.e. the probability mass function (PMF) is $f(k;p) = p^k(1-p)^{1-k}$.

$G(p)$: Geometric distribution with success probability of p, i.e. the probability that the kth trial is the first success is $P(X = k) = (1-p)^{k-1}p$.