Hidden Markov Model

Gaurav Kar

18111025

1 Introduction

Before delving into what the Hidden Markov Model is, let's understand the Markov Chain. A Markov Chain is a model or a type of random process that explains the probabilities of sequences of random variables, commonly known as states. Each of the states can take values from some set. In other words, we can explain it as the probability of being in a state, which depends on the previous state. We use the Markov Chain when we need to calculate the probability for a sequence of observable events. However, in most cases, the chain is hidden or invisible, and each state randomly generates 1 out of every k observations visible to us. Now, we will define the Hidden Markov Model. he Hidden Markov Model (HMM) is an analytical Model where the system being modeled is considered a Markov process with hidden or unobserved states. Machine learning and pattern recognition applications, like gesture recognition speech handwriting, are applications of the Hidden Markov Model.

HMM, Hidden Markov Model enables us to speak about observed or visible events and hidden events in our probabilistic model. The HMM is based on augmenting the Markov chain. A Markov chain is a model that tells us something about the probabilities of sequences of random variables, states, each of which can take on values from some set. These sets can be words, or tags, or symbols representing anything, like the weather. A Markov chain makes a very strong assumption that if we want to predict the future in the sequence, all that matters is the current state. The states before the current state have no impact on the future except via the current state. It's as if to predict tomorrow's weather you could examine today's weather but you weren't allowed to look at yesterday's weather. Hidden Markov models (HMMs), named after the Russian mathematician Andrey Andreyevich Markov, who developed much of relevant statistical theory, are introduced and studied in the early 1970s. They were first used in speech recognition and have been successfully applied to the analysis of biological sequences since late 1980s. Nowadays, they are considered as a specific form of dynamic Bayesian networks, which are based on the theory of Bayes. HMMs are statistical models to capture hidden information from observable sequential symbols (e.g., a nucleotidic sequence). They have many applications in sequence analysis, in particular to predict exons and introns in genomic DNA, identify functional motifs (domains) in proteins (profile

HMM), align two sequences (pair HMM). In a HMM, the system being modelled is assumed to be a Markov process with unknown parameters, and the challenge is to determine the hidden parameters from the observable parameters. A good HMM accurately models the real world source of the observed real data and has the ability to simulate the source. A lot of Machine Learning techniques are based on HMMs have been successfully applied to problems including speech recognition, optical character recognition, computational biology and they have become a fundamental tool in bioinformatics: for their robust statistical foundation, conceptual simplicity and malleability, they are adapted fit diverse classification problems. In Computational Biology, a hidden Markov model (HMM) is a statistical approach that is frequently used for modelling biological sequences. In applying it, a sequence is modelled as an output of a discrete stochastic process, which progresses through a series of states that are 'hidden' from the observer. Each such hidden state emits a symbol representing an elementary unit of the modelled data, for example, in case of a protein sequence, an amino acid. In the following sections, we first introduce the concepts of Hidden Markov Model as a particular type of probabilistic model in a Bayesian framework; then, we describe some important aspects of modelling Hidden Markov Models in order to solve real problems, giving particular emphasis in its use in biological context. To show the potentiality of these statistical approaches, we present the stochastic modelling of an HMM, defining first the model architecture and then the learning and operating algorithms. In this work we illustrate, as example, applications in computational biology and bioinformatics and, in particular, the attention is on the problem to find regions of DNA that are methylated or un-methylated (CpG-islands finding). The parameter learning task in HMMs is to find, given an output sequence or a set of such sequences, the best set of state transition and emission probabilities. The task is usually to derive the maximum likelihood estimate of the parameters of the HMM given the set of output sequences. No tractable algorithm is known for solving this problem exactly, but a local maximum likelihood can be derived efficiently using the Baum–Welch algorithm or the Baldi–Chauvin algorithm. The Baum–Welch algorithm is a special case of the expectation-maximization algorithm. If the HMMs are used for time series prediction, more sophisticated Bayesian inference methods, like Markov chain Monte Carlo (MCMC) sampling are proven to be favorable over finding a single maximum likelihood model both in terms of accuracy and stability. Since MCMC imposes significant computational burden, in cases where computational scalability is also of interest, one may alternatively resort to variational approximations to Bayesian inference, e.g Indeed, approximate variational inference offers computational efficiency comparable to expectation-maximization, while yielding an accuracy profile only slightly inferior to exact MCMC-type Bayesian inference.