e-PGPathshala

Subject : Computer Science

Paper: Machine Learning

Module: Markov and Hidden Markov Models

Module No: CS/ML/34

Quadrant I – e-text

Welcome to the e-PG Pathshala Lecture Series on Machine Learning.In this module we will discuss Markov Model which is the basis of a very important machine learning algorithm called Hidden Markov Model. Both these models are essentially sequential models of learning.

Learning Objectives:

The learning objectives of this module are as follows:

- To understand the concept of Markov Chain
- To explain Markov Chain with an example
- To understand the concept of Hidden Markov Model

34.1 Markov Random Processes

Until now, we have considered data to be i.i.d (independent and Identically distributed). Now we study sequential data such as time-series like stock market, speech, video analysis and ordered like text, genes, etc. If underlying process is unknown we can construct a model to predict the next state in sequence. In general, product rule expresses joint distribution for sequence

$$X_T) = \prod_{t=1}^{T} P(X_t|X_{t-1}, ..., X_1)$$

First-order Markov chain is described as each observation being independent of all previous observations except the most recent.

$$t-1, ..., X_1) = P(X_t|X_{t-1})$$

In this case the Maximum Likelihood parameter estimates are easy.

Now we define the Markovprocessas a simple stochastic process in which the distribution of future states depends only on the present state and not on how it arrived in the present state. A random sequence has the Markov property if its distribution is determined solely by its current state. Any random process having this property is called a Markov random process.For observable state sequences (state is known from data), this leads to a Markov chain model.For non-observable states, this leads to a Hidden Markov Model (HMM).

34.1.1 Markov Model

A discrete (finite) system can be described as consisting of:

- N distinct states.
- Begins (at time t=1) in some initial state(s).
- At each time step (t=1,2,…) the system moves from current to next state (possibly the same as the current state) according to transition probabilities associated with current state.

This kind of system is called a finite, or discrete Markov model. The model has been named after Andrei Andreyevich Markov (1856 -1922).

34.1.2 Markov Property

Now let us discuss a very important property that is known as the Markov property.

Markov Property: The state of the system at time *t*+1 depends only on the state of the system at time *t*. This property allows us to assume that the state at time *t*+1is independent of the states at all times before t (Figure 34.1). This is a simplifying assumption that allows us to compute the states of many sequential problems.

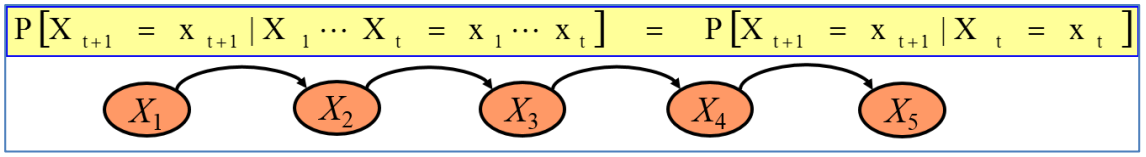


Figure 34.1 Markov Property

Stationary Assumption: In general, a process is called stationary if transition probabilities are independent of t, namely (Figure 34.2)

$$\text{for all t, } P[X_{t+1} = x_j | X_t = x_i] = p_{ij}$$

Figure 34.2 Stationary Property

This means that if system is in state i, the probability thatthe system will next move to state j is p_{ij} , no matter whatthe value of t is.

34.2 Components of the Markov Model

Markov model is a stochastic processes holding the Markov property. The Markov property is also called as memoryless property.This is because the future is independent of past given present. One step transition probability is the basis.We need the following three basic information to define a Markov model which is essentially a state space model:

- Parameter space.
- Statespace – where states represent various conditions of the System
- State transition probability -Transitions betweenstates indicateoccurrences of events

Now let us describe the components of the Markov Model in detail:

First we have a set of states:

$$\{ s_1, s_2, \cdots, s_N \}$$

In general the process moves from one state to another generating a sequence of states given as:

$$s_{i1}, s_{i2}, \cdots, s_{ik}, \cdots$$

Now according to the Markov chain property, the probability of each subsequent state depends only on what was the previous state:

$$P(s_{ik} \mid s_{i1}, s_{i2}, \cdots, s_{ik-1}) = P(s_{ik} \mid s_{ik-1})$$

To define Markov model, the following two probabilities have to be specified:

Transition probabilitywhich is the probability of state $S_i$given $S_j$and defined as:

$$a_{ij} = P(s_i \mid s_j)$$

and**initial probability which the probability of** $S_i$being an Initial state and given as:

$$\pi_i = P(s_i)$$

The output of the process is the set of states at each instant of time.

34.3 Markov Model – Examples

Now let us consider an example of the Markov Model scenario where we classify weather into three states (Figure 34.3) as State 1: rain or snow, – State 2: cloudy and State 3: sunny.

Figure 34.3 Three States of the Weather Example

It is given that the weather of some city followed the weather change pattern given in the following table (Figure 34.5):

		Tomorrow		
Today		Rainy	Cloudy	Sunny
	Rainy	0.4	0.3	0.3
	Cloudy	0.2	0.6	0.2
	Sunny	0.1	0.1	0.8

Figure 34.5 Weather Change Pattern

Here we make a major Markov assumption that tomorrow's weather depends only on today's weather. The above information can be represented as a Markov Model which is represented graphically as given in Figure 34.6.

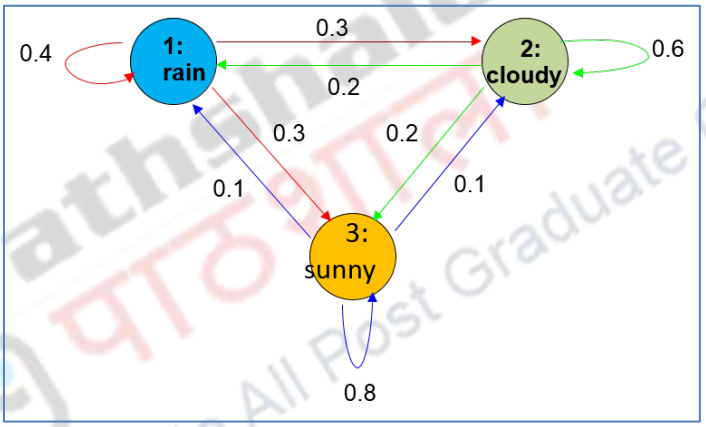


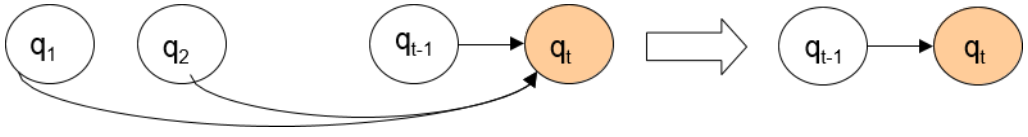
Figure 34.6 Graphical Representation

Here each state corresponds to one observation and the sum of outgoing edge weights is equal to one.

Here we have the observable states as {1,2,….,N}

Observable Sequence is q_1, q_2,……., q_T. In other words these are the stateswe will be observing to make a decision. Because of the 1st order Markov assumptionthe representation is as given in Figure 34.7.

$P(q_t = j \mid q_{t-1} = i, q_{t-2} = k,\ldots)$ reduces to $P(q_t = j \mid q_{t-1} = i)$



Stationary Bayesian Representation

Figure 34.7 Markov Assumption

We can consider

$$P(q_t = j \mid q_{t-1} = i) = P(q_{t+l} = j \mid q_{t+l-1} = i)$$

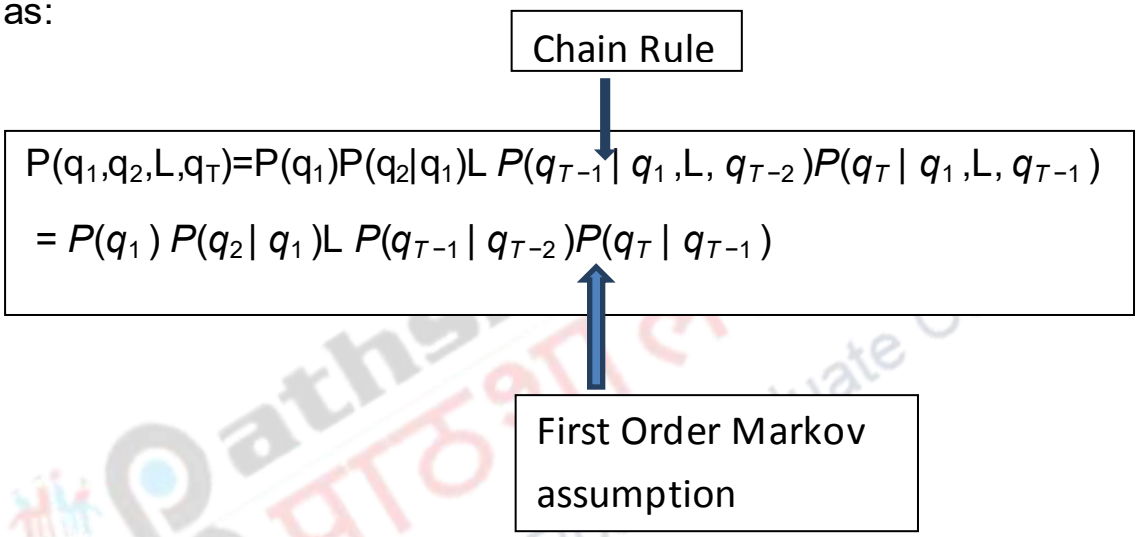
and initial state probability to be

$$\pi_i = P(q_1 = i), \quad 1 \le i \le N$$

Now let us consider Conditional probability

$$P(A, B) = P(A \mid B)P(B)$$

Now the sequence probability of Markov model can be represented as:



Markov Model – Example 1

Let us consider the following question: What is the probability that the weather for the next 7 days will be "sun-sun-rain-rain-sun-cloudy-sun" when today is sunny? (Refer Figure 34.6)

S_1 : *rain*, S_2 : *cloudy*, S_3 : *sunny*

$$P(O \mid \text{model}) = P(S_3, S_3, S_3, S_1, S_1, S_3, S_2, S_3 \mid \text{model})$$

$$= P(S_3) \cdot P(S_3 \mid S_3) \cdot P(S_3 \mid S_3) \cdot P(S_1 \mid S_3)$$

$$\cdot P(S_1 \mid S_1)P(S_3 \mid S_1)P(S_2 \mid S_3)P(S_3 \mid S_2)$$

$$= \pi_3 \cdot a_{33} \cdot a_{33} \cdot a_{31} \cdot a_{11} \cdot a_{13} \cdot a_{32} \cdot a_{23}$$

$$= 1 \cdot (0.8)(0.8)(0.1)(0.4)(0.3)(0.1)(0.2)$$

$$= 1.536 \times 10^{-4}$$

Markov model Matrix - Example 2

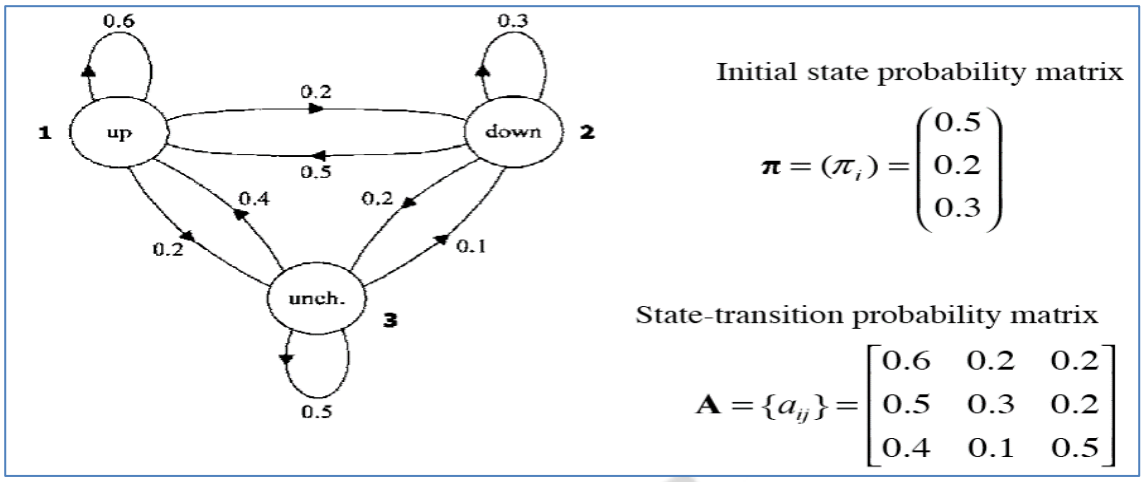


Figure 34.8 Markov Model – Example 2

What is the probability of 5 consecutive up days?

- Sequence is up-up-up-up-up
- I.e., state sequence is 1-1-1-1-1
- P(1,1,1,1,1) =
  - $\pi_1 a_{11} a_{11} a_{11} a_{11} = 0.5 \times (0.6)^4 = 0.0648$

34.4 Hidden Markov Model

A Hidden Markov Model is an extension of a Markov model in which the input symbols are not the same as the states.This means we don't know which state we are in.In HMM POS-tagging for example input symbols are the words, states are the part of speech tags (Figure 34. 9)

Input	Word	The	man	went	to	the	park
Symbol	Tag	Det	Noun	Verb	Prep	Det	Noun

Figure 34.9 POS Example

Hidden Markov Models (HMMs) are also called as probabilistic finite state automata. However to tackle scenarios where states cannot be directly observed, we need an extension called as Hidden Markov Models.

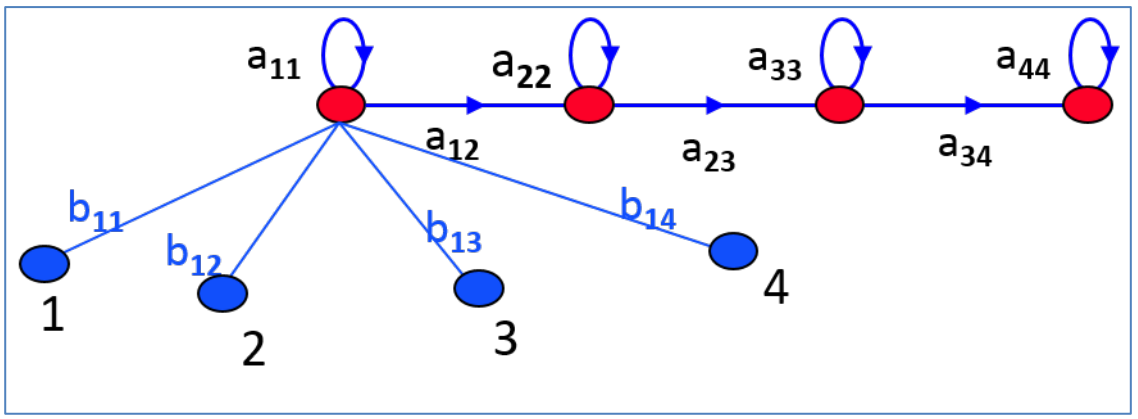


Figure 34.10 Example of Hidden Markov Model

In Figure 34.10, a_{ij} are state transition probabilities and b_{ik} are observation (output) probabilities. Please note that for the observed phenomenon$b_{11} + b_{12} + b_{13} + b_{14} = 1, b_{21} + b_{22} + b_{23} + b_{24} = 1$, etc.

Now considering the example shown in Figure 34.8, we can extend it to become an HMM. Here every state is associated with an additional probability called emission probability to indicate the probability of the output symbols emitted by each state.

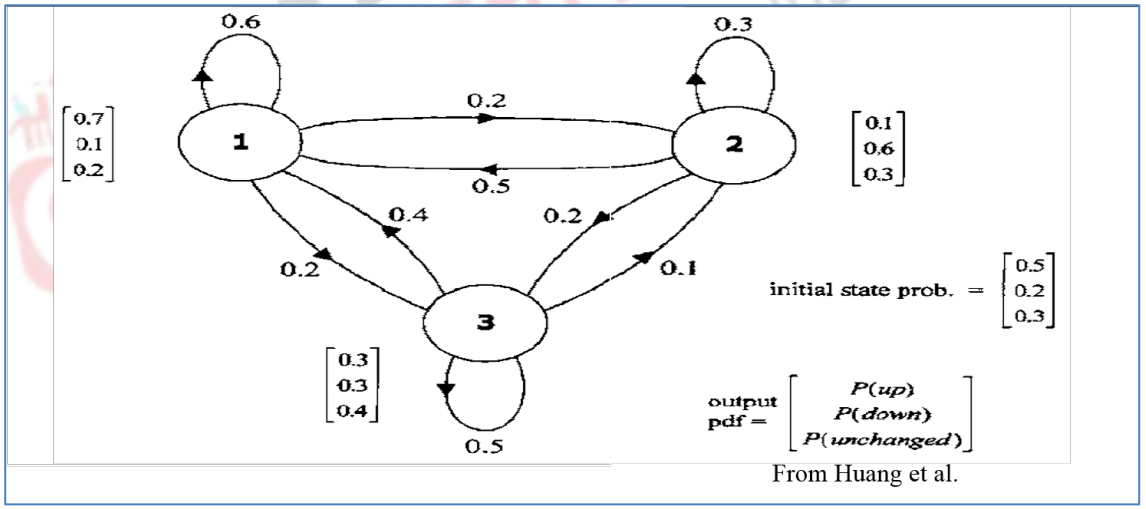


Figure 34.11 HMM Model

34.5 Hidden Markov Model – The Urn Example

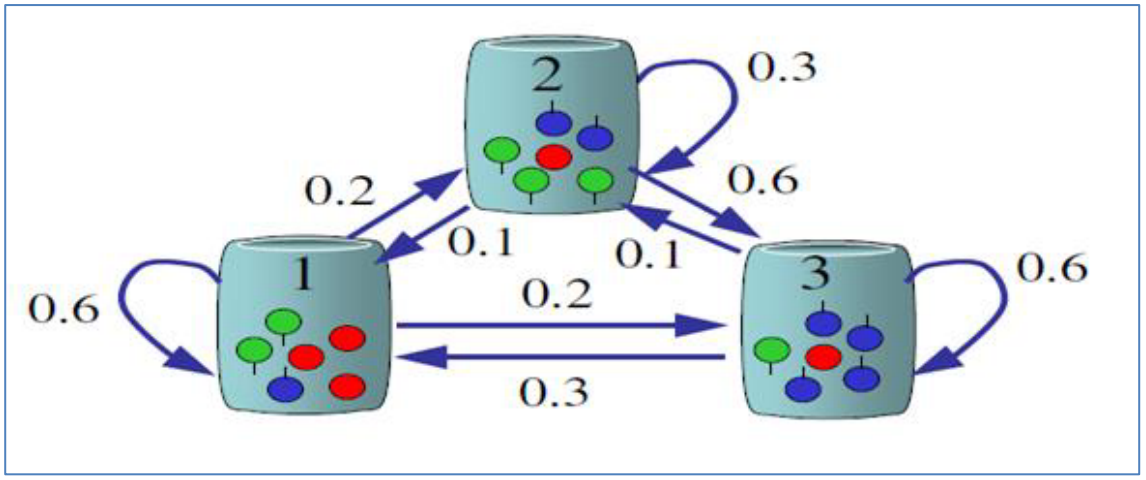


Figure 34.12 Urn Example

The example assumes we have N urns containing color balls of M distinct colors. Each urn can contain different number of balls of each color. Here we assume that we have 3 urns each containing 6 balls, which can be of 3 distinct colors. The HMM generation process is a sequential process (Figure 34.13) where

- Step 1: Pick initial urn according to some random process
- Step 2: Randomly pick a ball from the urn and then replace it
- Step 3: Select another urn according to a random selection process
- Step 4: Repeat Steps 2 & 3

Here we have the Markov process as $\{q(t)\}$ and the output process as $\{f(x|q)\}$.

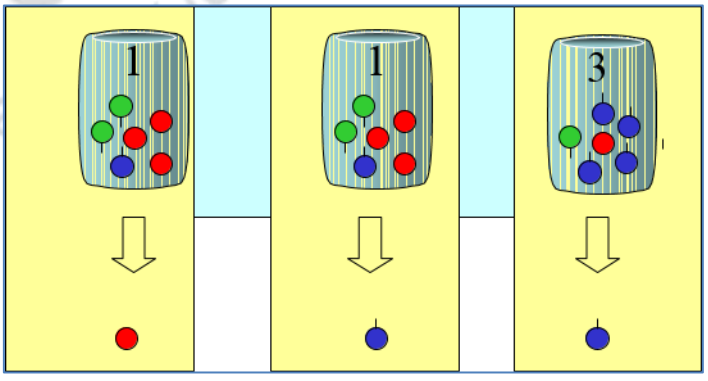


Figure 34.13 HMM Generation Process

Now the next question is what is hidden. Let us assume that we can see the color of the ball selected each time but however we do not know which urn is selected or the sequence in which urn selection takes place (state transition) is hidden (Figure 34.14).

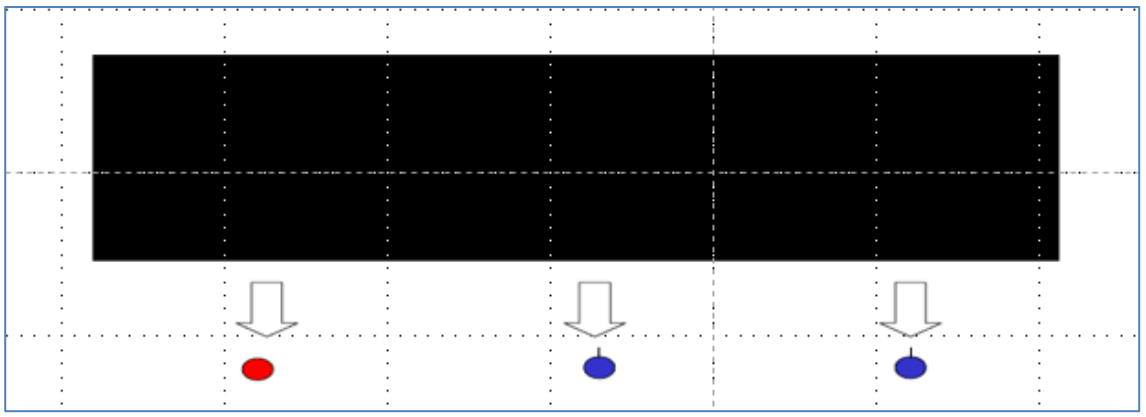


Figure 34.14 Hidden Information

HMM Assumptions:

There are some assumptions we make regarding the HMM model. The two assumptions are:

- **Markov assumption:** the state transition depends only on the origin and destination states
- **Output-independent assumption:** all observation frames are dependent only on the state that generated them, not on neighbouring observation frames

These assumptions allow us to now define the dependency structure of the HMM model. 1-st order Markov assumption of transition gives the following

$$P(q_t \mid q_1, q_2, \mathrm{L}, q_{t-1}) = P(q_t \mid q_{t-1})$$

Conditional independency of observation parameters

$$P(X_t \mid q_t, X_1, \mathrm{L}, X_{t-1}, q_1, \mathrm{L}, q_{t-1}) = P(X_t \mid q_t)$$

The output-independent assumption gives rise to the following structure (Figure 34.15).

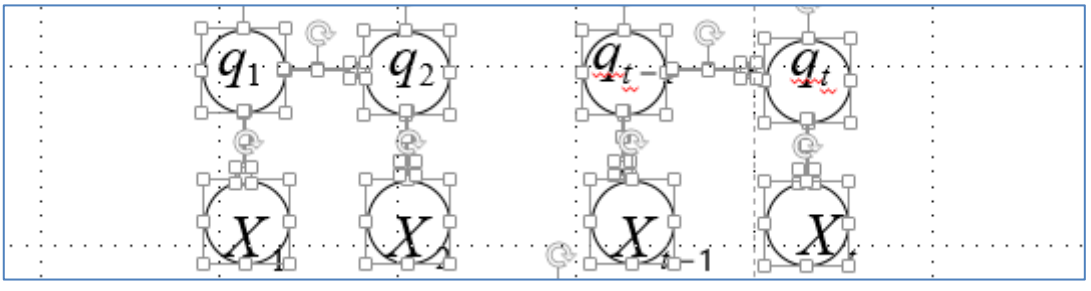


Figure 34.15 Bayesian Network

Figure 34.16 outlines the definition of the HMM model. Here N indicates the total number of states in the model while M indicates the number of symbols observable in the states. Now as we can see the HMM model is defined using

three probabilities A, which is the state transition probability, B, which is the observation probability and Π the initial state distribution.

HMM: Definition

- Notation: $\lambda = (A, B, \Pi)$

(1) N: Number of states

(2) M: Number of symbols observable in states

$$V = \{v_1, \cdots, v_M\}$$

(3) A: State transition probability distribution

$$A = \{a_{ij}\}, \quad 1 \le i, j \le N$$

(4) B: Observation symbol probability distribution

$$B = \{b_i(v_k)\}, \quad 1 \le i \le N, 1 \le j \le M$$

(5) Π: Initial state distribution

$$\pi_i = P(q_1 = i), \quad 1 \le i \le N$$

Figure 34.16 The HMM Definition

Figure 34.17 shows the HMM formalism. Here {*S, K*, P, *A, B*} is the model and *S* : {$s_1 \ldots s_N$} are the values for the hidden states and *K* : {$k_1 \ldots k_M$} are the values for the observations

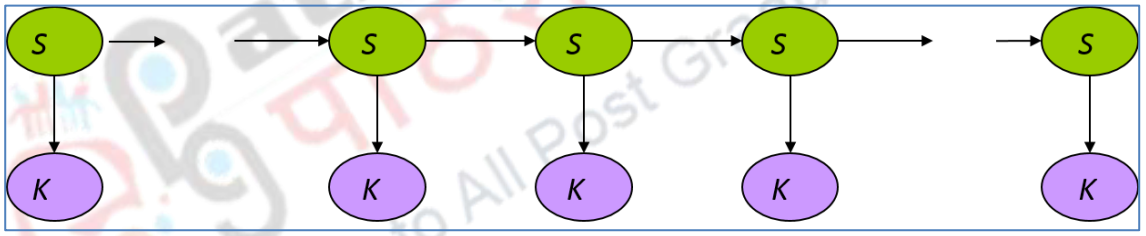


Figure 34.17 The Basic Formalism

The complete formalism is given in Figure 34.18. Here the model is defined by 5 parameters {*S, K*, P, *A, B*} where P = {p_i} are the initial state probabilities, *A* = {a_{ij}} are the state transition probabilities and *B* = {b_{ik}} are the observation state probabilities

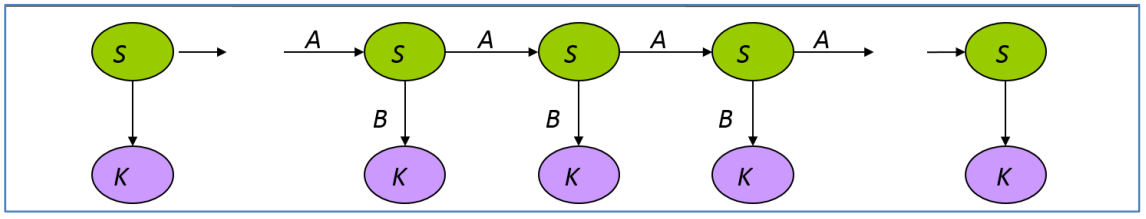


Figure 34.19 The Complete Formalism

Now let us revisit the Urn example and explain it using the formalism as shown in Figure 34.20. Here number of states N=3, the number of observations possible M=3 that is V={R,G,B}.

The initial state distribution is given as πFig 34.20 (a) and the state transition probability A = { aij } – is the transition probability of going from state I to state j and is given by the state transition matrix (Figure 34.20 (b)). Similarly the observation symbol probability distribution B { b_i (v_k) } is the probability of seeing symbol v_k in state b_i (Figure 34.20 (c)).

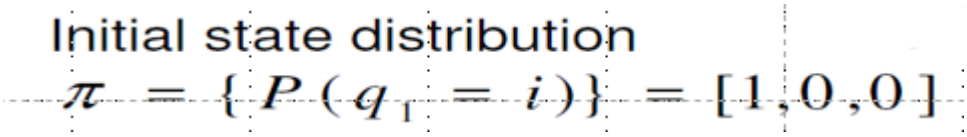


Figure 34.20 (a) Urn Example – Initial State Distribution

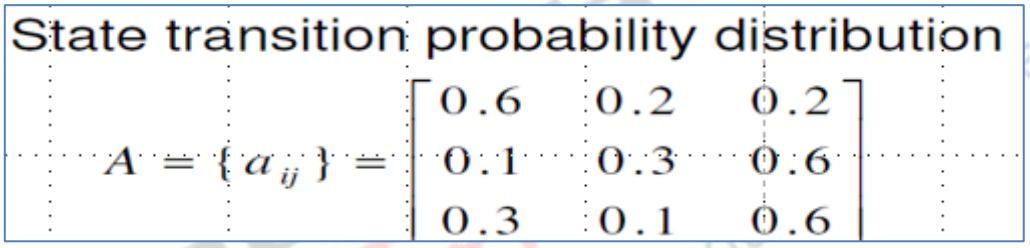


Figure 34.20 (b) Urn Example – State Transition Probability Distribution

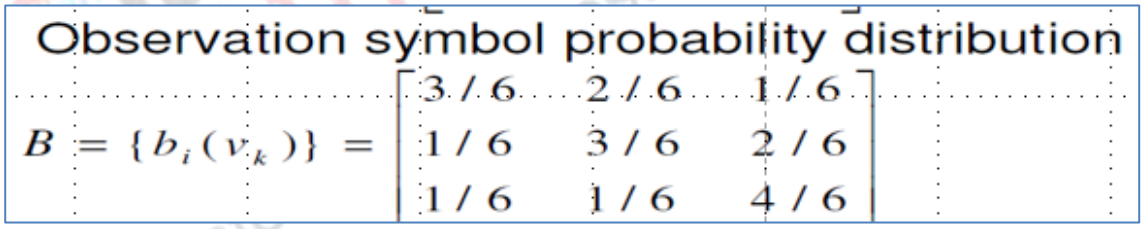


Figure 34.20 (c) Urn Example – Observation Symbol Probability Distribution

Let us discuss another example (sentence example) of HMM given in Figure 34.21. Here we have four states, the probability of emitting each symbol (here dog and eats) and the transition probability between states. The initial probability is 1 at S and 0 for all other states. Therefore we start at S (probability 1) , we can go to either state N or state V with probability 0.5. Let us assume we go to state N. Now the probability of emitting dog at state N is 0.9. Next let us assume we go to state V with probability 0.8. At V we generate eats with probability 0.9. Now we go from V back to N (probability 0.7) where we emit dog with probability 0.9. Now from N let us assume we go to state /S (probability 0.1). Thus the probability of following this sequence is 0.02 (Figure 34.21).

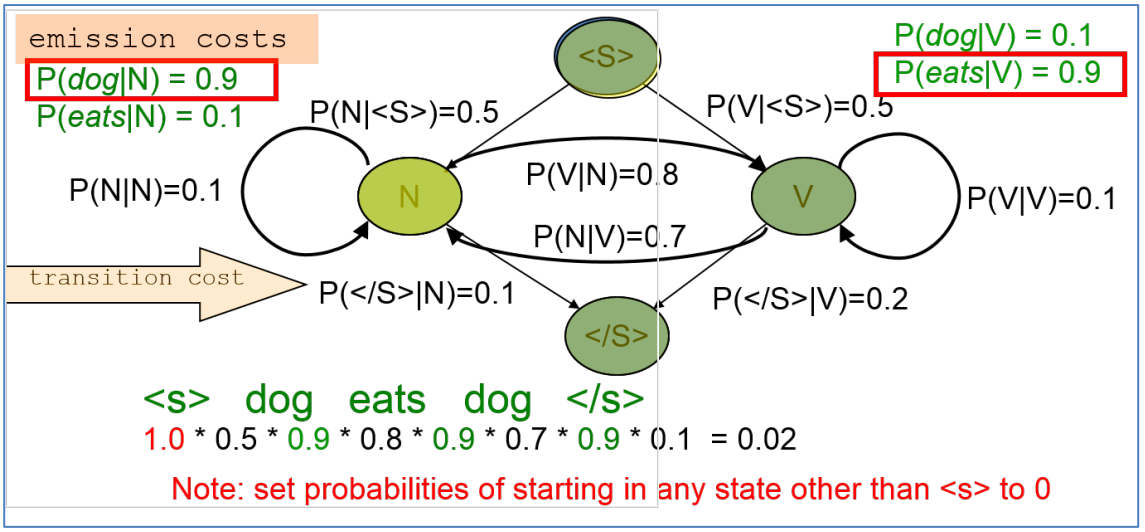


Figure 34.21 Sentence Example

34.5 Advantage of HMM on Sequential Data

HMM is a natural model structure which represents a doubly stochastic process where transition parameters model *temporal* variability and output distribution model *spatial* variability. It is efficient and good modelling tool forsequences with temporal constraints and spatial variability along the sequence, and this is able to model many real world complex processes. Efficient evaluation, decoding and training algorithms are available which are mathematically strong and computationally efficient. Moreover it is a proven technology with successful stories in many applications

34.6 Successful Application Areas of HMM

Some of the successful areas where HMM has been used is listed below:

- On-line handwriting recognition
- Speech recognition
- Gesture recognition
- Language modeling
- Motion video analysis and tracking
- Optical character recognition.
- Stock price prediction
- Flood predictions
- Modeling of coding/noncoding regions in DNA,
- Multiple sequence alignment,

Outlined here are details of some of the applications such as speech recognition where we need to recognize spoken words and phrases, text processing where we need to parse raw records into structured records, bioinformatics where we need to predict protein sequences and in finance domain where we make stock market forecasts by carrying out price pattern prediction or we can compare shopping services.

Summary

- Explained the concept of Markov Chain
- Outlined Markov Chain with an example
- Discussed the concept of Hidden Markov Model with examples