Hidden Markov Models
Lecture 1

J. Tuke and A.B. Rohrlach

School of Mathematical Sciences, University of Adelaide

2015

The Crooked Casino

Consider a crooked casino that has two dice:

- a fair die (F) with PMF:

y	1	2	3	4	5	6
$P(Y = y)$	1/6	1/6	1/6	1/6	1/6	1/6

- a biased die (B) with PMF:

y	1	2	3	4	5	6
$P(Y = y)$	1/10	1/10	1/10	1/10	1/10	1/2

Scenario 1

The casino chooses one die and sticks with it. You observe:

$$k \text{ sixes in } n \text{ rolls.}$$

Denote the number of sixes as Z, what is

$$P(Z = k|F) \text{ and } P(Z = k|B)?$$

Likelihood ratio test

We can also consider the likelihood of each die given the observations:

$$L(F|Z=k) \text{ and } L(B|Z=k).$$

We can then predict the die used using a likelihood ratio test:

$$\frac{L(F|Z=k)}{L(B|Z=k)} > 1 \Rightarrow F, \quad \frac{L(F|Z=k)}{L(B|Z=k)} < 1 \Rightarrow B.$$

[3]

Log likelihood ratio test

Alternatively, we can use the log likelihood ratio test:

$$\begin{aligned} \ell(F|Z=k) - \ell(B|Z=k) > 0 \Rightarrow F, \\ \ell(F|Z=k) - \ell(B|Z=k) < 0 \Rightarrow B, \end{aligned}$$

where $\ell(\cdot) = \log L(\cdot)$.

Tutorial Questions

Calculate the cutoff c such that if $k/n > c$ then you would predict the die is biased.

Scenario 2

What if the casino changes the die used occasionally?

Assume that after each die roll, the casino changes the die 10% of the time.

Naive Approach

We could use a sliding window.

For example let the number of observations be 60 denoted:

$$O_1, O_2, \ldots, O_{60}.$$

We could take a window of the first 10 observations

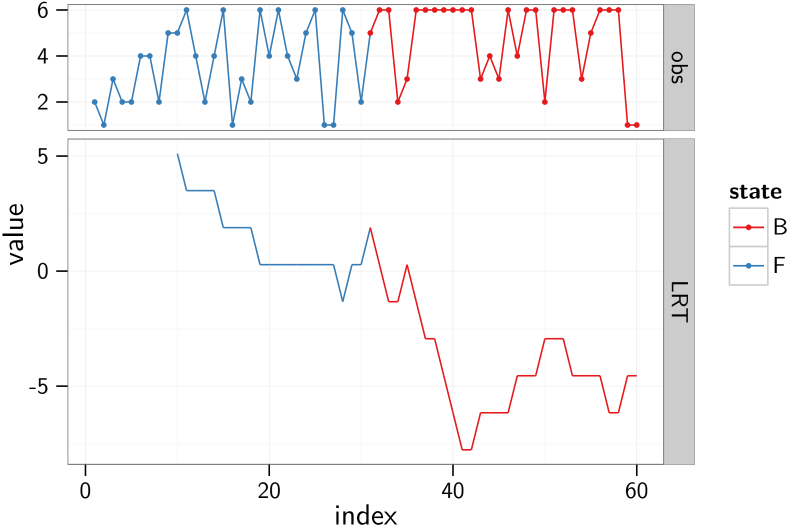
$$O_1, O_2, \ldots, O_{10}$$

and calculate the LRT, LRT_1, then slide the window over by one:

$$O_2, O_3, \ldots, O_{11}$$

and calculate LRT_2, and so on.

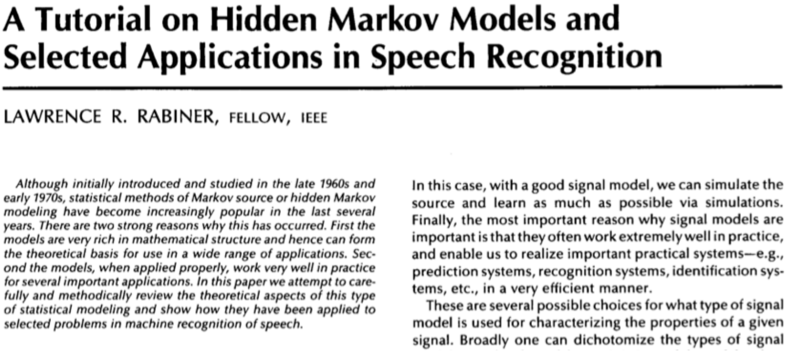
LRT



Hidden Markov Models

Instead we will use Hidden Markov Models.

A good introduction is



A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition

LAWRENCE R. RABINER, FELLOW, IEEE

Although initially introduced and studied in the late 1960s and early 1970s, statistical methods of Markov source or hidden Markov modeling have become increasingly popular in the last several years. There are two strong reasons why this has occurred. First the models are very rich in mathematical structure and hence can form the theoretical basis for use in a wide range of applications. Second the models, when applied properly, work very well in practice for several important applications. In this paper we attempt to carefully and methodically review the theoretical aspects of this type of statistical modeling and show how they have been applied to selected problems in machine recognition of speech.

In this case, with a good signal model, we can simulate the source and learn as much as possible via simulations. Finally, the most important reason why signal models are important is that they often work extremely well in practice, and enable us to realize important practical systems—e.g., prediction systems, recognition systems, identification systems, etc., in a very efficient manner.

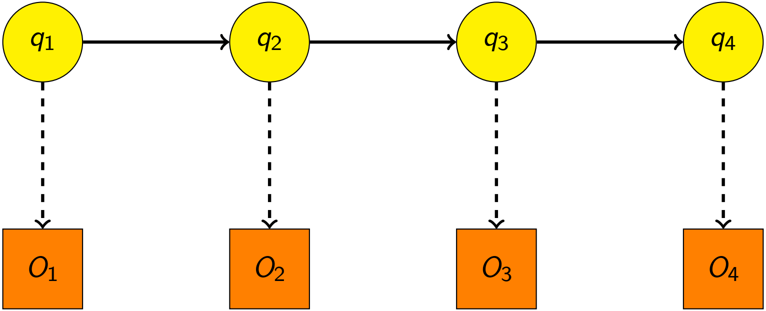
These are several possible choices for what type of signal model is used for characterizing the properties of a given signal. Broadly one can dichotomize the types of signal

[9]

Hidden Markov Models

A Hidden Markov Model consists of a Markov Model, for which the states are not observed, but instead the observations are a probabilistic function of the state.

Hidden Markov Model



[11]

Elements of Hidden Markov Model (HMM)

States

We denote the possible states of the Markov Chain by

$$S = \{S_1, S_2, \ldots, S_N\}.$$

We also denote the state at time t by q_t.

Elements of Hidden Markov Model (HMM)

Observations

We have M possible observations for each state which we denote as:

$$V = \{v_1, v_2, \ldots, v_M\}.$$

Elements of Hidden Markov Models (HMM)

State transistion probability distribution

We denote this as $A = \{a_{ij}\}$, where

$$a_{ij} = P(q_{t+1} = S_j | q_t = S_i), \quad 1 \leq i, j \leq N.$$

Elements of Hidden Markov Models (HMM)

Observation symbol probability distribution

We denote the observation symbol probability distribution in state j by $B = \{b_j(k)\}$, where

$$b_j(k) = P(v_k \text{ at } t | q_t = S_j), \quad 1 \leq j \leq N, 1 \leq k \leq M.$$

[15]

Elements of Hidden Markov Models (HMM)

Initial state distribution

we denote this as $\boldsymbol{\pi} = \{\pi_i\}$, where

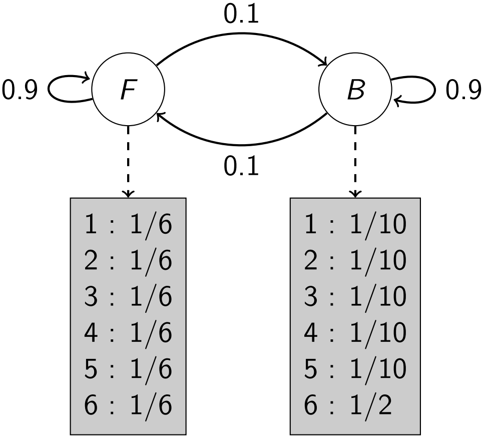
$$\pi_i = P(q_1 = S_i), \quad 1 \leq i \leq N.$$

[16]

Elements of Hidden Markov Models (HMM)

What are the elements for the casino example Scenario 2.

Casino Example



1 : 1/6
2 : 1/6
3 : 1/6
4 : 1/6
5 : 1/6
6 : 1/6

1 : 1/10
2 : 1/10
3 : 1/10
4 : 1/10
5 : 1/10
6 : 1/2

[18]

Simulating an HMM

```
procedure SIM_HMM($N, M, A, B, \boldsymbol{\pi}$)
    $q_1 \leftarrow S_i$ according to $\boldsymbol{\pi}$
    for $t \leftarrow 1 : T$ do
        $O_t \leftarrow v_k$ according to $b_i(k)$
        $q_{t+1} \leftarrow S_j$ according to $a_{ij}$
    end for
```

Coding challenge 1

Simulate 100 observations for the casino example.

Useful R command to look up: `sample()`.

Coding challenge 2

Is your code working? Questions to consider:

- What is the stationary distribution of the Markov chain describing the states (F, B)? Are you getting the correct proportions in your simulation?
- What frequency of observations should you get for each state? Are you obtaining these?
- Use a goodness of fit test to test that your observed frequencies are not statistically different from expected.

  Hint: `chisq.test()`

Coding challenge 3

Coding regions in DNA

The nucleotide sequences in DNA can be classified as coding (transcribed and translated as protein) and non-coding (does not result in direct protein production).

It is known that if a nucleotide is non-coding there is a 50% chance that the next nucleotide is coding, while if a nucleotide is coding, there is a 20% chance that the next nucleotide is non-coding.

For the non-coding nucleotides there is an equal chance that they will be A, C, G or T. For the coding nucleotides, the frequency of A and C are the same, the frequency of G and T are the same, and the frequency of A is twice the frequency of G.

Write the HMM for this problem, simulate it, test it.