

Bayesian Network $\xrightarrow{\text{for}}$ Probabilistic inference (static environment) (calculate probabilities)

Lecture 13

\rightarrow Probabilistic inference with time

Hidden Markov Models

13.1 Markov Chains

Definition 1. A *Markov chain* is a sequence of variables X_1, \dots, X_t where the probability distribution of X_t depends on only a *finite fixed number* of previous variables. (x at time = t)

\rightarrow time = 1
 \rightarrow time = t

Definition 2. A first-order Markov chain is a Markov chain where the probability distribution X_t depends on only the value of X_{t-1} , i.e.

$$\mathbf{P}(X_t | X_1, X_2, \dots, X_{t-1}) = \mathbf{P}(X_t | X_{t-1})$$

$X_i = \text{discrete random variable}$

Definition 3. A *finite first-order Markov chain* is defined as a 3-tuple as:

(1) a finite set of states $Q = \{1, 2, \dots, k\}$ = possible values of X_i

(2) transition probabilities between two states $A = \{a_{ij}\}$ where

$$a_{ij} = p(X_t = j | X_{t-1} = i) \rightarrow \text{transition probability} \rightarrow \text{probability that the state at } t = j \text{ given that the state at } t-1 = i$$

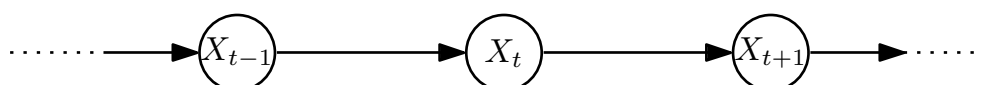
with a constraint that $\sum_{j=1}^k a_{ij} = 1$ for all $i \in Q$,

(3) initial state probabilities $\pi = \{\pi_i\}$ where

$$\pi_i = p(X_0 = i) \rightarrow \text{initial probability} \rightarrow \text{probability that the state at } 0 = i$$

with $\sum_{i=1}^k \pi_i = 1$

The graphical model of a Markov chain is shown below.



Example 1. In one city, a sunny day is 90% likely to be followed by another sunny day. A rainy day is 40% likely to be followed by another rainy day. The weather of day 0 is known to be “sunny”. Write the formal representation of the Markov chain corresponding to this situation. $X_t = \text{weather on day } t$

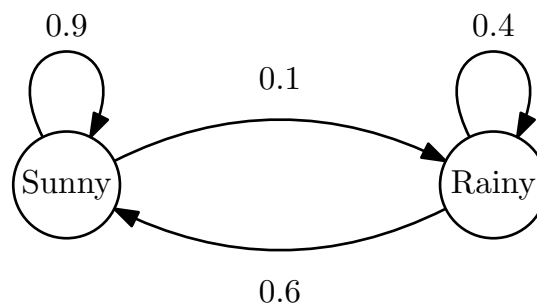
- 1) set of states $Q = \{ 1, 2 \}$
- 2) transition probabilities
- $P(X_t = j | X_{t-1} = i)$
- $P(X_t = 1 | X_{t-1} = 1)$ sunny 0.9
- $P(X_t = 2 | X_{t-1} = 1)$ rainy 0.1
- $P(X_t = 1 | X_{t-1} = 2)$ sunny 0.6
- $P(X_t = 2 | X_{t-1} = 2)$ rainy 0.4
- $A = \begin{bmatrix} 0.9 & 0.1 \\ 0.6 & 0.4 \end{bmatrix}$

- 3) initial probabilities

$\pi = \begin{bmatrix} 1.0 & 0.0 \end{bmatrix}$

sunny rainy

Each Markov chain can be depicted as a *state diagram*. Each state represents a state. Each transition between states is attached with a probability value.



(1=sunny, 2=rainy)

Example 2. Find the probability that the weather of day 2 is “rainy”.

$$\begin{aligned}
 P(X_2=2) &= P(X_0=1, X_1=1, X_2=2) \\
 &\quad + P(X_0=1, X_1=2, X_2=2) \\
 &= P(X_0=1) P(X_1=1 | X_0=1) P(X_2=2 | X_1=1) \\
 &\quad + P(X_0=1) P(X_1=2 | X_0=1) P(X_2=2 | X_1=2) \\
 &= (1.0)(0.9)(0.1) + (1.0)(0.1)(0.4) = 0.13
 \end{aligned}$$

13.2 Hidden Markov Model

Definition 4. A *Hidden Markov Model* (HMM) is a finite Markov chain where its state at time t cannot be directly observed. Each state however emits a symbol with a certain probability. Each HMM is defined as a 5-tuple:

- (1) a finite set of states $Q = \{1, 2, \dots, k\}$
- (2) a finite set of symbols $\Sigma = \{b_1, b_2, \dots, b_m\}$
- (3) transition probabilities $A = \{a_{ij}\}$ where

$$a_{ij} = p(X_t = j \mid X_{t-1} = i)$$

with $\sum_{j=1}^k a_{ij} = 1$ for all $i \in Q$,

- (4) initial state probabilities $\pi = \{\pi_i\}$

$$\pi_i = p(X_0 = i)$$

with $\sum_{i=1}^k \pi_i = 1$

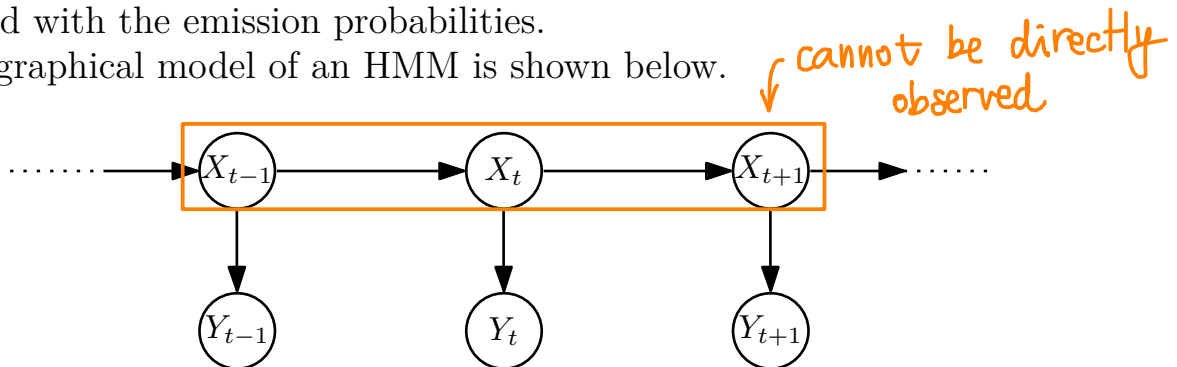
- (5) Emission probabilities $E = \{e_i(b)\}$

$$e_i(b) = p(Y_t = b \mid X_t = i) \leftarrow \text{probability that state } i \text{ emits symbol } b$$

with $\sum_{b \in \Sigma} e_i(b) = 1$ for all $i \in Q$,

Similar to the Markov chain, an HMM can be depicted as a state diagram. Every state transition is labeled with a transition probability. Every state is attached with the emission probabilities.

The graphical model of an HMM is shown below.

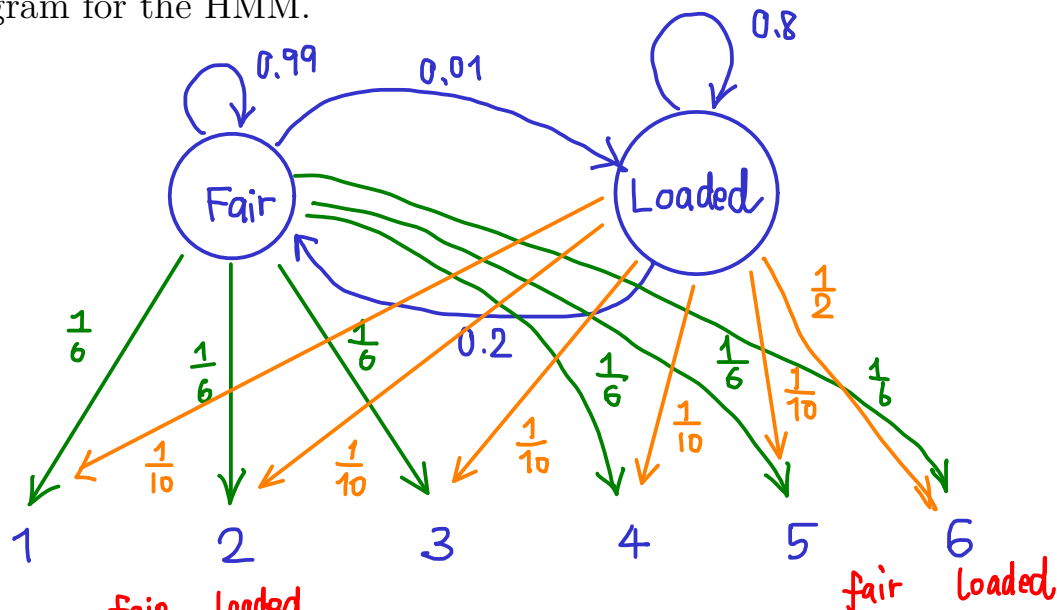


Example 3. A casino has two dice:

(1) Fair die: $p(1) = p(2) = p(3) = p(4) = p(5) = p(6) = \frac{1}{6}$, and

(2) Loaded die: $p(1) = p(2) = p(3) = p(4) = p(5) = \frac{1}{10}$; $p(6) = \frac{1}{2}$

The dealer may switch between the dice with the probability that the dealer switches from the fair die to the loaded die is 0.01, and the probability that the dealer switches from the loaded die to the fair die is 0.2. However, the dealer always starts with the fair die. Write the formal definition and draw the diagram for the HMM.



① $Q = \{ \overset{\text{fair}}{1}, \overset{\text{loaded}}{2} \}$

④ $\pi = [\overset{\text{fair}}{1.0}, \overset{\text{loaded}}{0.0}]$

② $\Sigma = \{ 1, 2, 3, 4, 5, 6 \}$

⑤

③ $A = \begin{matrix} & \overset{\text{fair}}{1} & \overset{\text{loaded}}{2} \\ \overset{\text{fair}}{\text{fair}} & 0.99 & 0.01 \\ \overset{\text{loaded}}{\text{loaded}} & 0.20 & 0.80 \end{matrix}$

$E = \begin{matrix} & \overset{\text{fair}}{1} & \overset{\text{loaded}}{2} & \overset{\text{fair}}{3} & \overset{\text{loaded}}{4} & \overset{\text{fair}}{5} & \overset{\text{loaded}}{6} \\ \overset{\text{fair}}{\text{fair}} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} & \frac{1}{6} \\ \overset{\text{loaded}}{\text{loaded}} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{10} & \frac{1}{2} \end{matrix}$

Example 4. Let λ_2 be an HMM defined as follow:

- (1) States: $Q = \{1, 2, 3\}$
- (2) Observation symbols: $\Sigma = \{a, b, c, d\}$
- (3) Transition probabilities

$$\begin{array}{llll} a_{11} = 0.1 & a_{12} = 0.4 & a_{13} = 0.5 & a_{21} = 0.5 & a_{22} = 0.2 & a_{23} = 0.3 \\ a_{31} = 0.2 & a_{32} = 0.6 & a_{33} = 0.2 & & & \end{array}$$

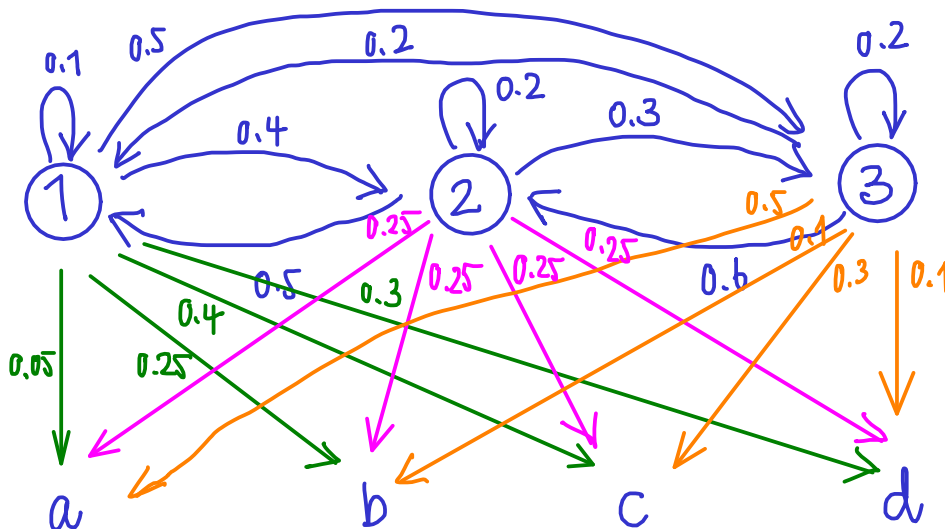
- (4) Emission probabilities

$$\begin{array}{llll} e_1(a) = 0.05 & e_1(b) = 0.25 & e_1(c) = 0.4 & e_1(d) = 0.3 \\ e_2(a) = 0.25 & e_2(b) = 0.25 & e_2(c) = 0.25 & e_2(d) = 0.25 \\ e_3(a) = 0.5 & e_3(b) = 0.1 & e_3(c) = 0.3 & e_3(d) = 0.1 \end{array}$$

- (5) Initial state probabilities

$$a_{01} = 1.0 \qquad a_{02} = 0.0 \qquad a_{03} = 0.0$$

Draw a state diagram showing this HMM λ_2 .

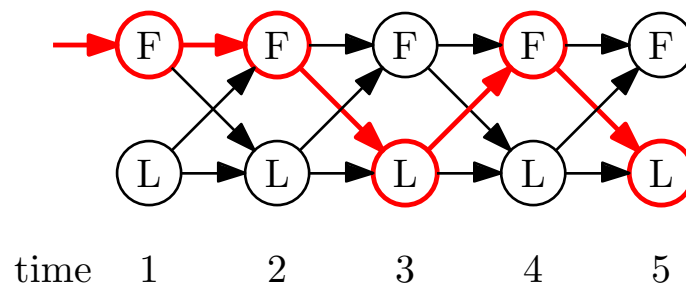


13.2.1 Parse and Trellis

Definition 5. A *parse* is a sequence of visited states at each time.

Each parse is typically depicted by a *trellis* which is a directed graph. Each node in the trellis represents a state at one time. Each edge in the trellis shows a state transition from one time to another.

Example 5. From the HMM for the fair and loaded dice in the previous example, we can draw a trellis for a parse Fair \rightarrow Fair \rightarrow Loaded \rightarrow Fair \rightarrow Loaded as below.



13.2.2 HMM Inference

Evaluation

Evaluation is a task to compute the probability that an observation sequence is generated given an HMM.

Likelihood of a Parse Given an HMM λ , a parse $X = \langle x_1, \dots, x_n \rangle$, and an observation sequence $Y = \langle y_1, \dots, y_n \rangle$, the likelihood of the parse producing the observation sequence (given the HMM λ) can be computed as

$$\begin{aligned}
 & p(\underbrace{y_1, \dots, y_n}_Y, \underbrace{x_1, \dots, x_n}_X \mid \lambda) \\
 &= p(y_1, \dots, y_n, x_1, \dots, x_n) \\
 &= p(x_1) p(y_1 \mid x_1) \prod_{t=2}^n p(x_t \mid x_{t-1}) p(y_t \mid x_t) \tag{13.2.1}
 \end{aligned}$$

$$= p(y_1, \dots, y_{n-1}, x_1, \dots, x_{n-1}) p(y_i \mid x_i) p(x_{i-1} \mid x_i) \tag{13.2.2}$$

initial prob
emission prob
transition prob

Example 6. Let λ_1 be the HMM in the previous example. Given an observation sequence $Y = \langle 1, 1, 6, 6, 6 \rangle$ and a parse $X = \langle F, F, L, F, L \rangle = \langle 1, 1, 2, 1, 2 \rangle$, compute the likelihood of the parse producing the observation sequence $p(Y, X \mid \lambda_1)$.

$$\begin{aligned} p(Y, X \mid \lambda_1) &= p(Y, X) = p(1, 1, 6, 6, 6; F, F, L, F, L) \\ &= p(F) p(1 \mid F) p(F \mid F) p(1 \mid F) p(L \mid F) p(6 \mid L) \\ &\quad p(F \mid L) p(6 \mid F) p(L \mid F) p(6 \mid L) \end{aligned}$$

$$= (1.0) \left(\frac{1}{6}\right) (0.99) \left(\frac{1}{6}\right) (0.01) \left(\frac{1}{2}\right) (0.2) \left(\frac{1}{6}\right) (0.01) \left(\frac{1}{2}\right)$$

$$= 2.29 \times 10^{-8}$$

$$p(6, 6, 6; F, L, L \mid \lambda_1) = ?$$

$$= p(F) p(6 \mid F) p(L \mid F) p(6 \mid L) p(L \mid L) p(6 \mid L)$$

$$= (1.0) \left(\frac{1}{6}\right) (0.01) \left(\frac{1}{2}\right) (0.8) \left(\frac{1}{2}\right)$$

$$= 3.33 \times 10^{-4}$$

Likelihood of a Sequence

Let $Y = \langle y_1, y_2, \dots, y_n \rangle$ be an observation sequence and λ be an HMM,

$$p(Y | \lambda) = p(Y) = \sum_{X \in \mathcal{X}} p(Y, X) \quad (13.2.3)$$

↙ for all possible sequences of states

where \mathcal{X} is a set of all possible parses with length of n .

For the initial probabilities

$$P(F) = 1.0, \quad \underline{\underline{P(L) = 0.0}}$$

Example 7. Compute $p(1, 4, 6 | \lambda_1)$

$$\begin{aligned} p(1, 4, 6) = & P(1, 4, 6; F, F, F) + \\ & P(1, 4, 6; F, F, L) + \\ & P(1, 4, 6; F, L, F) + \\ & P(1, 4, 6; F, L, L) + \\ & \cancel{P(1, 4, 6; L, F, F)}^0 + \\ & \cancel{P(1, 4, 6; L, F, L)}^0 + \\ & \cancel{P(1, 4, 6; L, L, F)}^0 + \\ & \cancel{P(1, 4, 6; L, L, L)}^0 \end{aligned}$$

$$\begin{aligned} = & P(F) P(1|F) P(F|F) P(4|F) P(F|F) P(6|F) + \\ & P(F) P(1|F) P(F|F) P(4|F) P(L|F) P(6|L) + \\ & P(F) P(1|F) P(L|F) P(4|L) P(F|L) P(6|F) + \\ & P(F) P(1|F) P(L|F) P(4|L) P(L|L) P(6|L) \end{aligned}$$

$$= 0.0047$$

Forward algorithm

Computing $p(Y)$ directly from Equation 13.2.3 is a time-consuming task since many redundant calculations are performed. The dynamic programming approach can be used to speed up the calculation. We define the *forward probability* $f_k(i)$ as the probability of generating the first i symbol in the observation sequence and ending up in state k .

$$\begin{aligned}
 f_k(i) &= p(y_1, \dots, y_i, x_i = k) \\
 &= \sum_{x_1, \dots, x_{i-1}} p(y_1, \dots, y_i, x_1, \dots, x_{i-1}, x_i = k) \\
 &= \sum_l \sum_{x_1, \dots, x_{i-2}} p(y_1, \dots, y_i, x_1, \dots, x_{i-2}, x_{i-1} = l, x_i = k) \\
 &= \sum_l \left[\left[\sum_{x_1, \dots, x_{i-2}} p(y_1, \dots, y_{i-1}, x_1, \dots, x_{i-2}, x_{i-1} = l) \right] a_{lk} e_k(y_i) \right] \\
 &= \sum_l p(y_1, \dots, y_{i-1}, x_{i-1} = l) a_{lk} e_k(y_i) \\
 &= e_k(y_i) \sum_l f_l(i-1) a_{lk}
 \end{aligned} \tag{13.2.4}$$

Handwritten notes: A blue arrow points from the term $a_{lk} e_k(y_i)$ in the third line to the expression $p(x_i=k | x_{i-1}=l)$. Another blue arrow points from the same term to the expression $p(y_i | x_i=k)$.

where $f_k(1) = \pi_k e_k(y_1)$

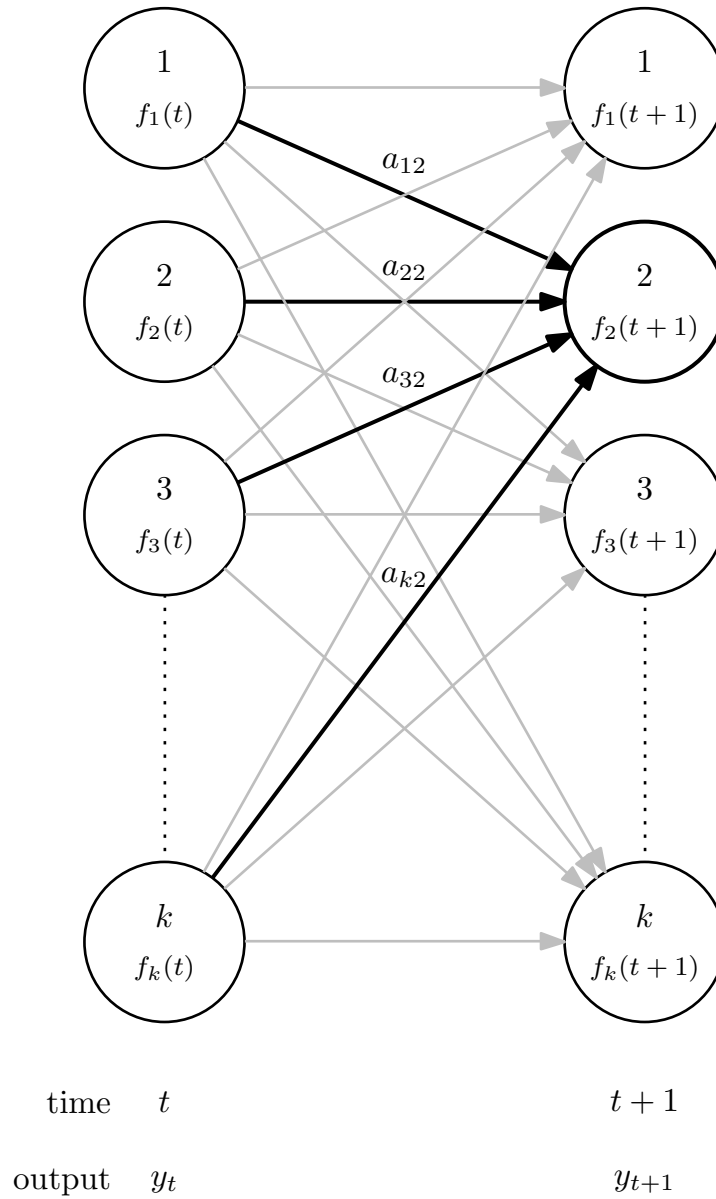
We can then compute $p(Y)$ based on the forward probability as

$$p(Y) = \sum_k f_k(n) \tag{13.2.5}$$

where n is the length of Y .

The following figure shows how the *forward probability* is computed.

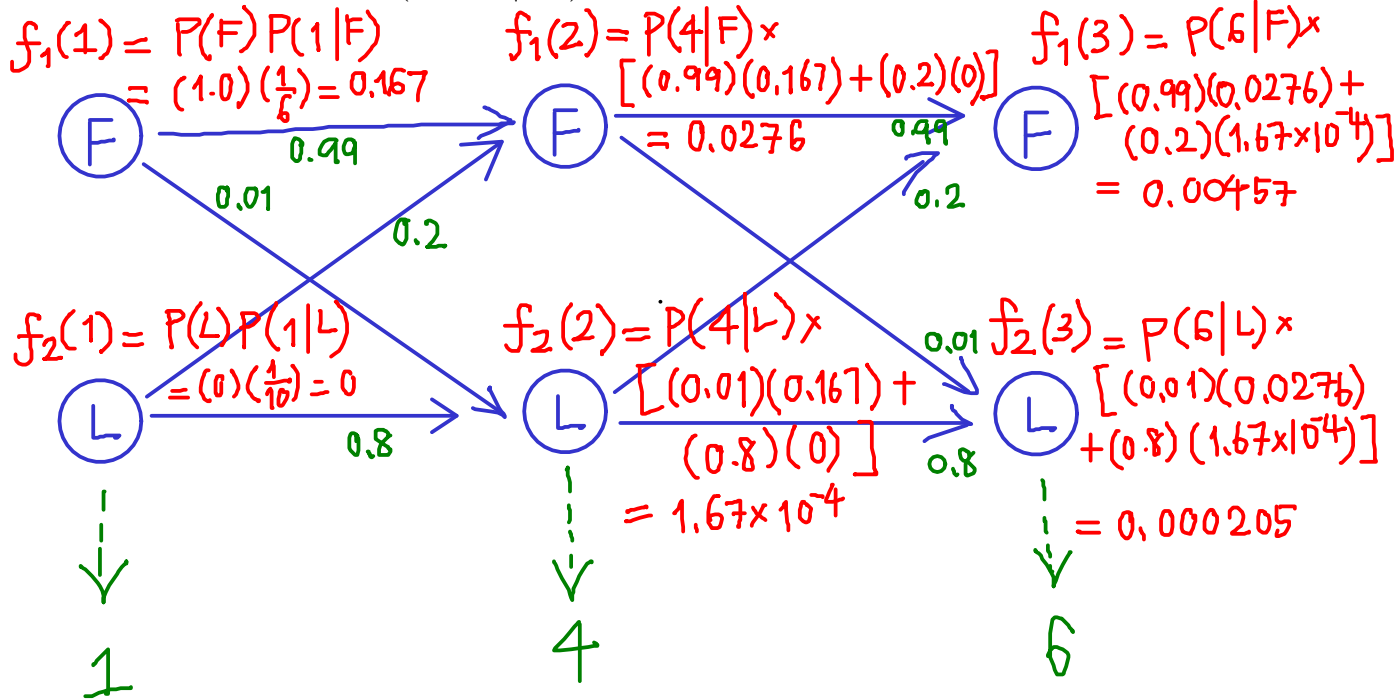
$$f_2(t+1) = e_2(y_{t+1})[f_1(t)a_{12} + \dots + f_k(t)a_{k2}]$$



Forward Algorithm

```
1 k = 2
2
3 a = [[0.99, 0.01],
4       [0.20, 0.80]]
5
6 e = [{'1':0.166, '2':0.166, '3':0.166,
7       '4':0.166, '5':0.166, '6':0.166},
8       {'1':0.100, '2':0.100, '3':0.100,
9       '4':0.100, '5':0.100, '6':0.500}]
10
11 pi = [1.0, 0.0]
12
13 Y = '146'
14
15 f = [[0.0 for i in enumerate(Y)] for j in range(k)]
16
17 # Initialization
18 for i in range(k):
19     f[i][0] = pi[i]*e[i][Y[0]]
20
21 # Induction
22 for t in range(1, len(Y)):
23     for i in range(k):
24         s = 0.0
25         for l in range(k):
26             s += f[l][t-1]*a[l][i]
27         f[i][t] = e[i][Y[t]]*s
28
29 # Termination
30 t = len(Y)-1
31 p = 0.0
32 for i in range(k):
33     p += f[i][t]
34
35 print(p)
```

Example 8. Compute $p(1, 4, 6 \mid \lambda_1)$ using the forward algorithm.

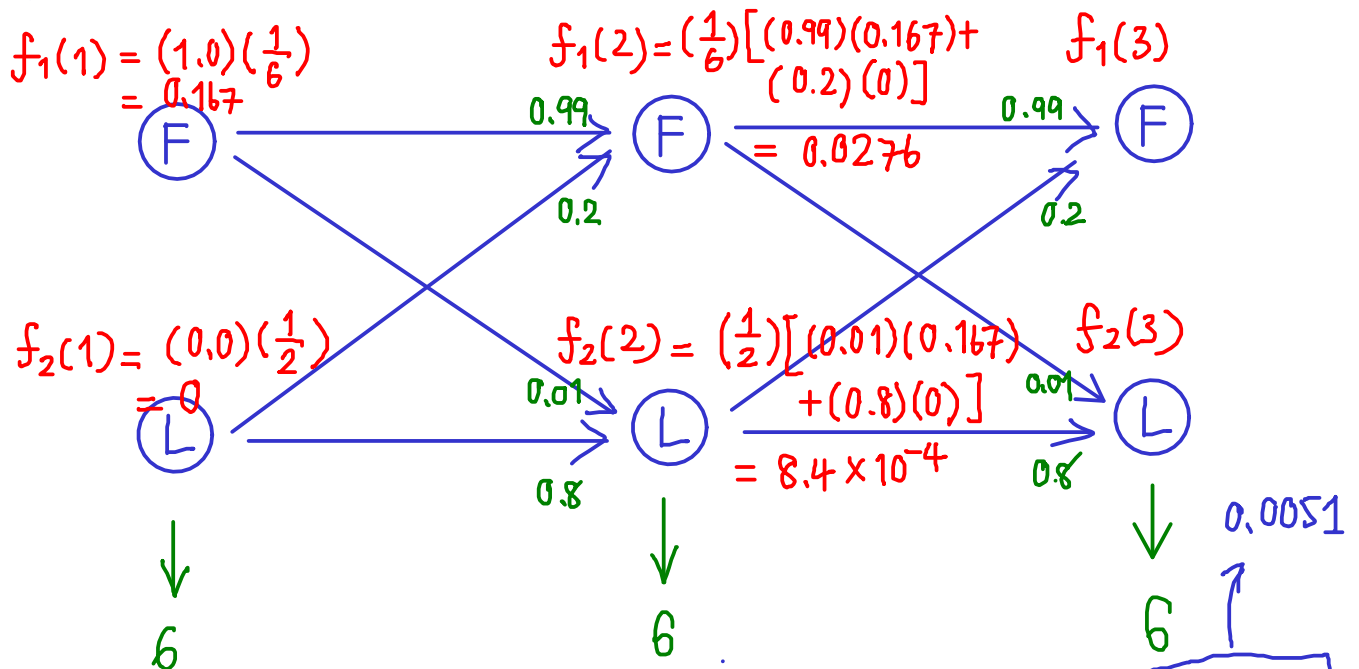


$$P(1, 4, 6 \mid \lambda_1) = f_1(3) + f_2(3)$$

$$= 0.00457 + 0.000205$$

$$= 0.00478$$

$$P(6, 6, 6 \mid \lambda_1) = ?$$



$$f_1(3) = (\frac{1}{6}) [(0.99)(0.0276) + (0.2)(8.4 \times 10^{-4})] = 0.0046$$

$$f_2(3) = (\frac{1}{2}) [(0.01)(0.0276) + (0.8)(8.4 \times 10^{-4})] = 0.00047$$

13.2.3 Decoding

Decoding is a task to find a parse for a given observation sequence that maximizes the likelihood of the parse.

Given an HMM λ and an observation sequence $Y = \langle y_1, \dots, y_n \rangle$, find

$$X^* = \operatorname{argmax}_X p(Y, X | \lambda) \quad (13.2.6)$$

Similar to the forward algorithm, we can use the dynamic programming approach to perform the task. The algorithm is called “*Veterbi algorithm*”. We define

$$V_k(i) = e_k(y_i) \max_{l=1, \dots, k} [V_l(i-1) a_{lk}] \quad (13.2.7)$$

where $V_k(1) = \pi_k e_k(y_1)$

Example 9. Find the parse X that maximizes $p(1, 4, 6 \mid \lambda_1)$

Veterbi Algorithm

```

1 k = 2
2
3 states = ['Fair', 'Loaded']
4
5 a = [[0.99, 0.01],
6       [0.20, 0.80]]
7
8 e = [{'1':0.166, '2':0.166, '3':0.166,
9       '4':0.166, '5':0.166, '6':0.166},
10      {'1':0.100, '2':0.100, '3':0.100,
11       '4':0.100, '5':0.100, '6':0.500}]
12
13 pi = [1.0, 0.0]
14
15 Y = '146'
16
17 V = [[0.0 for i in enumerate(Y)] for j in range(k)]
18
19 W = [[0 for i in enumerate(Y)] for j in range(k)]
20
21 # Initialization
22 for i in range(k):
23     V[i][0] = pi[i]*e[i][Y[0]]
24
25 # Induction
26 for t in range(1, len(Y)):
27     for i in range(k):
28         v = []
29         for l in range(k):
30             v.append(V[l][t-1]*a[l][i])
31
32         V[i][t] = e[i][Y[t]]*max(v)
33         W[i][t] = v.index(max(v))
34
35 # Termination
36 v = []
37 t = len(Y)-1
38 for i in range(k):
39     v.append(V[i][t])
40 vi = v.index(max(v))
41 X = states[vi]
42
43 for i in range(t-1, -1, -1):
44     vi = W[vi][i]
45     X = states[vi] + ' ' + X
46
47 print(X)

```


13.2.4 Learning

Given a training set

$$\mathcal{D} = \{(X_i, Y_i)\}_{i=1}^N$$

where each X_i is a parse, and Y_i is an observation sequence generated from X_i .

This *learning* task is to estimate the parameters (A, B, π) that best explains the training set. We can use a technique called “*maximum likelihood estimation (MLE)*” to find the parameters.

$$a_{ij} = p(X_t = j \mid X_{t-1} = i) = \frac{\text{Count}(i \rightarrow j)}{\text{Count}(i)} \quad (13.2.8)$$

where $\text{Count}(i \rightarrow j)$ returns the number of $i \rightarrow j$ transitions found in the parses in \mathcal{D} . Similarly, $\text{Count}(i)$ is the number of state i in \mathcal{D} .

$$e_k(b) = p(Y_t = b \mid X_t = k) = \frac{\text{Count}(b, k)}{\text{Count}(k)} \quad (13.2.9)$$

where $\text{Count}(b, k)$ is the number of times that a symbol b is emitted from state k , and $\text{Count}(k)$ is the number of state k .

$$\pi_i = p(X_1 = i) = \frac{\text{Count}(X_1 = i)}{N} \quad (13.2.10)$$

where $\text{Count}(X_1 = i)$ is the number of parses having i as the first state.

Example 10. From the following examples, estimate the parameters of an HMM λ_3 with $Q = \{1, 2, 3, 4\}$ and $\Sigma = \{a, b, c, d\}$:

Parse	Observation sequence
1, 2, 1, 3, 4, 4	a, b, b, a, c, d
2, 1, 3, 4, 4	b, b, a, c, d
1, 1, 1, 2, 2, 4	c, d, d, c, a, a
2, 3, 3, 3, 4, 4, 1	c, a, b, c, d, a, b

References

- Lawrence R. Rabiner, “A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition”, Proceedings of the IEEE, Vol. 77, No. 2, February 1989.