SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

50.007 Machine Learning, Fall 2021
Homework 4

Due Tuesday 30 November 2021, 5pm

This homework will be graded by Zhang Qi

In this homework, we would like to look at the Hidden Markov Model (HMM), one of the most influential models used for structured prediction in machine learning.

1. (10 pts) Assume that we have the following training data available for us to estimate the model parameters:

Count(X) = 5
Count(Y) = 5
Count(Z) = 5

State sequence	Observation sequence
$(\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{X})$	$(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{b})$
$(\mathbf{X}, \mathbf{Z}, \mathbf{Y})$	$(\mathbf{a}, \mathbf{b}, \mathbf{c})$
$(\mathbf{Z}, \mathbf{Y}, \mathbf{X}, \mathbf{Z}, \mathbf{Y})$	$(\mathbf{b}, \mathbf{c}, \mathbf{a}, \mathbf{b}, \mathbf{c})$
$(\mathbf{Z}, \mathbf{X}, \mathbf{Y})$	$(\mathbf{c}, \mathbf{b}, \mathbf{a})$

   Clearly state what are the parameters associated with the HMM. Under the maximum likelihood estimation (MLE), what would be the values for the optimal model parameters? Clearly show how each parameter is estimated exactly.

2. (10 pts) Now, consider during the evaluation phase, you are given the following new observation sequence. Using the parameters you just estimated from the data, find the most probable state sequence using the Viterbi algorithm discussed in class. Clearly present the steps that lead to your final answer.

State sequence	Observation sequence
$(?, ?)$	$(\mathbf{b}, \mathbf{c})$

follow that of the lecture slides

3. (10 pts) The Viterbi algorithm discussed in class can be used for computing the single most probable state sequence for a new observation sequence. Specifically, in the Viterbi algorithm we are interested in finding the optimal sequence using the following formula:

$$(s_1^*, s_2^*) = \arg\max_{s_1, s_2} P(s_1, s_2 | o_1 = \mathbf{b}, o_2 = \mathbf{c})$$

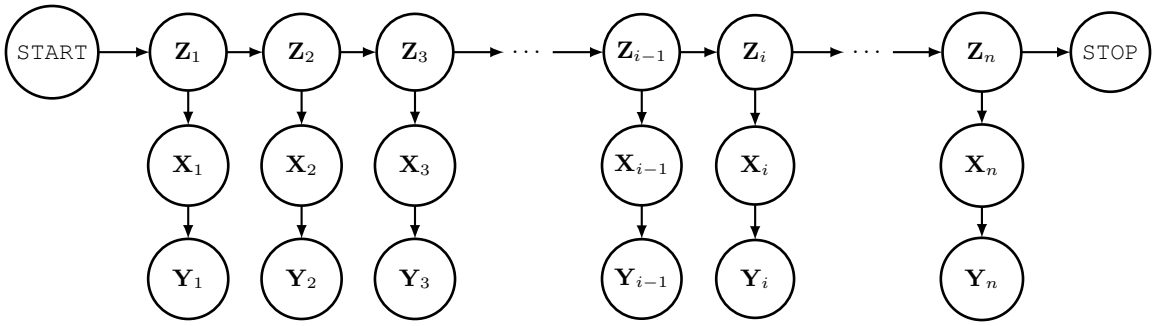
   However, an alternative method to estimate the state sequence is finding the optimal state of index i by solving the optimization problem

joint probability

$$s_i^* = \arg\max_{s_i} P(s_i | o_1 = \mathbf{b}, o_2 = \mathbf{c})$$

   where $P(s_i | o_1 = \mathbf{b}, o_2 = \mathbf{c})$ is the marginal distribution and $i \in \{1, 2\}$. Please use the forward and backward scores for HMM discussed in class to calculate the marginal distribution $P(s_i | o_1 = \mathbf{b}, o_2 = \mathbf{c})$. Clearly present the steps that lead to your final answer.

4. (20 pts) Now consider a slightly different graphical model which extends the HMM (see below). For each state ($\mathbf{Z}$), there is now an observation pair ($\mathbf{X}$, $\mathbf{Y}$), where $\mathbf{Y}$ sequence is generated from the $\mathbf{X}$ sequence.



Assume you are given a large collection of observation pair sequence, and a predefined set of possible states, you would like to estimate the most probable state sequence for each observation pair sequence using an EM algorithm similar to the dynamic programming algorithm discussed in class. Clearly define the forward and backward scores in a way analogous to those defined for HMM, and explain what they mean. Give algorithms for computing the forward and backward scores. Analyze the complexity associated with your algorithms.

1. Two parameters:

① Emission probability $[a_{u,v}]$

↳ output obtain from state (probability)

② Transition probability $[b_u(O)]$

↳ The probability of moving from one state to the next state.

$$a_{u,v} = \frac{\text{count}(u,v)}{\text{count}(u)} \qquad b_u(O) = \frac{\text{count}(u \to O)}{\text{count}(u)}$$

- Transition probability

Sanity check (Σ)

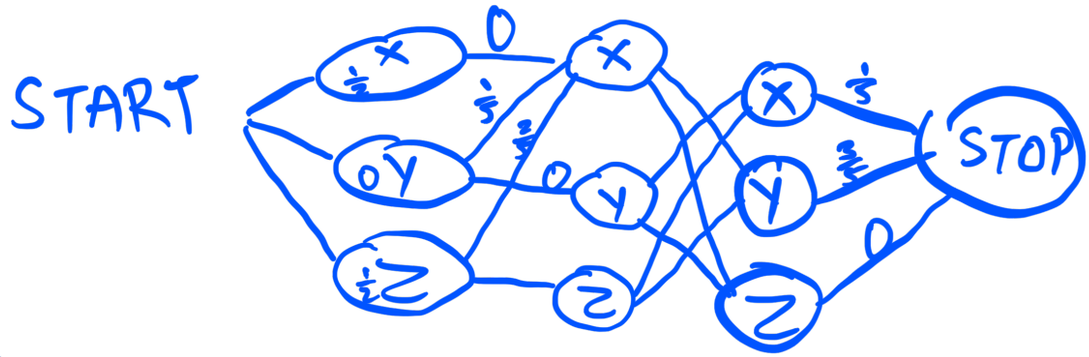
u | v	X	Y	Z	STOP	
START	$\frac{1}{2}$	0	$\frac{1}{2}$	0	1
X	0	$\frac{2}{5}$	$\frac{2}{5}$	$\frac{1}{5}$	1
Y	$\frac{1}{5}$	0	$\frac{1}{5}$	$\frac{3}{5}$	1
Z	$\frac{2}{5}$	$\frac{3}{5}$	0	0	1

The table passes the sanity check.

- Emission probability

u | O	a	b	c	Sanity check (Σ)
X	$\frac{2}{5}$	$\frac{3}{5}$	$\frac{0}{5}$	1
Y	$\frac{1}{5}$	$\frac{0}{5}$	$\frac{4}{5}$	1
Z	$\frac{1}{5}$	$\frac{3}{5}$	$\frac{1}{5}$	1

Q2. Using the back tracking algorithm,



① We know that we cannot move from START to state Y ($a_{START,Y} = 0$)

② Our state before STOP cannot be (Z)

Via Back tracking, we trace a few paths:

ⓐ START → X → Y → STOP

likelihood = $a_{START,X} \times a_{X,Y} \times a_{Y,STOP} \times b_X(b) \times b_Y(c)$

$= \frac{1}{2} \times \frac{2}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{4}{5}$

$= 0.0576$

ⓑ START → Z → Y → STOP

likelihood = $a_{START,Z} \times a_{Z,Y} \times a_{Y,STOP} \times b_Z(b) \times b_Y(c)$

$= \frac{1}{2} \times \frac{3}{5} \times \frac{3}{5} \times \frac{3}{5} \times \frac{4}{5}$

$= 0.0864$

Path ⓑ seems short (But not a very good way to our final)

Initialisation $\pi(0,u) = \begin{cases} 1 & \text{if } u = \text{START} \\ 0 & \text{otherwise} \end{cases}$

① $\pi(1,X) = a_{START,X} \times b_X(b) = \frac{1}{2} \times \frac{3}{5} = \frac{3}{10}$

$\pi(1,Z) = a_{START,Z} \times b_Z(b) = \frac{1}{2} \times \frac{3}{5} = \frac{3}{10}$

$\pi(1,Y) = a_{START,Y} \times b_Y(b) = 0$

② $\pi(2,X) = \max\{\pi(1,u) \times a_{u,X} \times b_X(c)\}$

$= \max\{\pi(1,X) \times a_{X,X} \times b_X(c),$
$\pi(1,Z) \times a_{Z,X} \times b_X(c),$
$\pi(1,Y) \times a_{Y,X} \times b_X(c)\}$

$= \max\{0, 0, 0\}$

$= 0$

$\pi(2,Y) = \max\{\pi(1,u) \times a_{u,Y} \times b_Y(c)\}$

$= \max\{\pi(1,X) \times a_{X,Y} \times b_Y(c),$
$\pi(1,Y) \times a_{Y,Y} \times b_Y(c),$
$\pi(1,Z), a_{Z,Y} \times b_Z(c)\}$

$= \max\{\frac{3}{10} \times \frac{2}{5} \times \frac{4}{5}, 0, \frac{3}{10} \times \frac{3}{5} \times \frac{4}{5}\}$

$= \max\{0.096, 0, 0.144\}$

$= 0.144$ #

$$\pi(2,Z) = \max\{\pi(1,u) \times a_{u,Z} \times b_Z(c)\}$$

$$= \max\{\pi(1,X) \times a_{X,Z} \times b_Z(c)$$
$$\pi(1,Y) \times a_{Y,Z} \times b_Z(c)$$
$$\pi(1,Z) \times a_{Z,Z} \times b_Z(c)\}$$

$$= \max\{\tfrac{3}{10} \times \tfrac{2}{5} \times \tfrac{1}{5}, 0, 0\}$$

$$= 0.024 \#$$

③ $\pi(3, \text{STOP}) = \max\{\pi(2,\pi) \times a_{u,\text{STOP}}\}$

$$= \max\{\pi(2,\pi) \times a_{X,\text{STOP}},$$
$$\pi(2,\pi) \times a_{Y,\text{STOP}},$$
$$\pi(2,\pi) \times a_{Z,\text{STOP}}\}$$

$$= \max\{0, 0.144 \times \tfrac{3}{5}, 0\}$$

$$= 0.0864 \#$$

∴ Final Path

START → Z → Y → STOP

3. From lecture notes 4,

$$s_i^* = \arg\max_{s_i} P(s_i | o_1 = b, o_2 = c)$$

⇕ relate to the problem

To find the optimal tag at the position i given the entire observation sequence of (O_1, O_2)

We can then apply the result,

$$s_i^* = \arg\max_u (\alpha_u(i)\beta_u(i)), \quad i \in \{1, 2\}$$

Using forward probability, we can define $\alpha_u(i)$!

/ Base case:

$$\alpha_u(1) = \alpha_{START,u}$$

Recursive:

$$\alpha_u(j+1) = \sum_v \alpha_v(j) a_{v,u} b_v(x_j)$$

Stop @ $j+1 = 3$

$\alpha_x(1) = \alpha_{START,x} = \frac{1}{2}$
$\alpha_y(1) = \alpha_{START,y} = 0$
$\alpha_z(1) = \alpha_{START,z} = \frac{1}{2}$

} Base case:

$$\begin{aligned}\alpha_x(2) &= \alpha_x(1) \times a_{x,x} \times b_x(b) + \\ &\quad \alpha_y(1) \times a_{y,x} \times b_y(b) + \\ &\quad \alpha_z(1) \times a_{z,x} \times b_z(b) \\ &= 0 + 0 + \left(\frac{1}{2} \times \frac{2}{5} \times \frac{3}{5}\right) \\ &= \frac{3}{25} \\ &= 0.12\end{aligned}$$

$$\begin{aligned}\alpha_y(2) &= \alpha_x(1) \times a_{x,y} \times b_x(b) + \\ &\quad \alpha_y(1) \times a_{y,y} \times b_y(b) + \\ &\quad \alpha_z(1) \times a_{z,y} \times b_z(b) \\ &= \left(\frac{1}{2} \times \frac{2}{5} \times \frac{3}{5}\right) + 0 + \left(\frac{1}{2} \times \frac{3}{5} \times \frac{3}{5}\right) \\ &= 0.3\end{aligned}$$

$$\alpha_2(2) = \alpha_X(1) \times a_{X,Z} \times b_X(b) +$$
$$\alpha_Y(1) \times a_{Y,Z} \times b_Y(b) +$$
$$\alpha_Z(1) \times a_{Z,Z} \times b_Z(b) +$$
$$= \frac{1}{2} \times 0.4 \times 0.6 + 0 + 0$$
$$= 0.12$$

For Backward score calculation,

Base case:

$$\beta_u(n) = a_{u,STOP}\, b_u(x_n)$$

Recursive case:

$$\beta_u(j) = \sum_v a_{u,v}\, b_u(x_j)\, \beta_v(j+1)$$

$$\beta_X(2) = a_{X,STOP} \times b_X(c) = 0$$

$$\beta_Y(2) = a_{Y,STOP} \times b_Y(c) = 0.6 \times \frac{4}{5} = 0.48$$

$$\beta_Z(2) = a_{Z,STOP} \times b_Z(c) = 0$$

$$\beta_X(1) = a_{X,X} \times b_X(b) \times \beta_X(2) +$$
$$a_{X,Y} \times b_X(b) \times \beta_Y(2) +$$
$$a_{X,Z} \times b_X(b) \times \beta_Z(2)$$
$$= 0 + \frac{2}{5} \cdot \frac{3}{5} \cdot 0.48 + 0$$
$$= 0.1152$$

$$\beta_Y(2) = a_{Y,X} \times b_Y(b) \times \beta_X(2) +$$
$$a_{Y,Y} \times b_Y(b) \times \beta_Y(2) +$$
$$a_{Y,Z} \times b_Y(b) \times \beta_Z(2)$$
$$= 0 + 0 + 0$$
$$= 0$$

$$\beta_Z(2) = a_{Z,X} \times b_Z(b) \times \beta_X(2)$$
$$a_{Z,Y} \times b_Z(b) \times \beta_Y(2)$$
$$a_{Z,Z} \times b_Z(b) \times \beta_Z(2)$$
$$= 0 + \frac{3}{5} \times \frac{3}{5} \times 0.48 + 0$$
$$= 0.1728$$

s_1^* = most possible states!
$= \arg\max_u \{\alpha_u(1) \cdot \beta_u(1)\}$

$= \arg\max_u \{\alpha_x(1) \cdot \beta_x(1), \alpha_y(1) \cdot \beta_y(1), \alpha_z(1) \cdot \beta_z(1)\}$

$= \arg\max_u \{0.0576, 0, 0.0864\}$

$= z$

$s_2^* = \arg\max_u (\alpha_u(2) \cdot \beta_u(2)$

$= \arg\max_u \{\alpha_x(2) \times \beta_x(2), \alpha_y(2) \times \beta_y(2), \alpha_z(2) \times \beta_z(2)\}$

$= \arg\max_u \{0, 0.144, 0\}$

$= y$

Q4.

Using soft EM, we can use the results to derive the following for our use.

$$a_{v,u} = \frac{\text{count}(v,u)}{\text{count}(v)} \qquad b_v(x_i) = \frac{\text{count}(v \to x_i)}{\text{count}(v)}$$

$$\underbrace{c_v(y_i)}_{\text{low emission for each layer.}} = \frac{\text{count}(v \to y_i)}{\text{count}(v)}$$

∴ Using the above, we can compute the forward and backward scores respectively.

- Forward score,

$\alpha_u(i)$ is the sum of scores of all paths from start to state (i,u)

$$\alpha_u(1) = a_{\text{START},u}$$

$$\alpha_u(i+1) = \sum_v \alpha_v(i)\, a_{v,u}\, b_v(x_i)\, c_v(x_i) \quad \forall u \in 1, \ldots, N-1, \; i = 1, \ldots, n-1$$

- Backward score,

$\beta_u(i)$ is the sum of scores of all paths from state (i, u) to STOP

Base case:

$\beta_u(n) = a_{u,\text{STOP}}\, b_u(x_n) \cdot c_u(y_n) \quad \forall u \in 1, \ldots, N-1$

Recursive case:

$$\beta_u(i) = \sum_v a_{u,v}\, b_u(x_i)\, \beta_v(i+1) \cdot c_u(y_i) \quad \forall u \in 1, \ldots, N-1,\ i = n-1, \ldots, 1$$

For each layer, $O(T)$ time complexity!

$\beta_u(i)$ is $O(T)$ time complexity.

$\beta_u(i) = \sum_v \left(a_{u,v} \cdot b_u(x_i) \cdot c_u(y_i) \cdot \beta_v(i+1) \right]$

T nodes 1 1 1 1

$= O(T)$

4 steps for each node to the other node.

$$O(T^2 n) = O(T) \cdot O(T) \cdot O(n)$$

Explanation: ⓐ for each node to connect to T other nodes in computation ($O(T)$)

ⓑ For all T nodes in each layer to connect to T other nodes, in computation ($O(T) \times O(T)$)

For n layers, to repeat ⓐ and ⓑ in each layer,

Total time complexity $= O(T) \times O(T) \times O(n)$

$$= O(T^2 n)$$

*We discount 4 as constants do not matter too much in time complexity (Refer to introduction to Algorithms (50.004))

For space complexity, we store T values in each layer.

∴ For n layers,

Space complexity $= O(T \times n) = O(Tn)$