SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

50.007 Machine Learning, Fall 2020
Homework 3

Due Monday 27 Nov 2020, 11:59pm

Sample Solutions

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:

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

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.

Q1 - 1 pt per error.

Answer The transition probabilities are estimated as:

$$a_{u,v} = \frac{\text{Count}(u; v)}{\text{Count}(u)}$$

	X	Y	Z	STOP
START	3/6	0	3/6	0
X	1/6	0	4/6	1/6
Y	1/4	0	0	3/4
Z	1/8	4/8	1/8	2/8

The emission probabilities are estimated as:

$$b_u(o) = \frac{\text{Count}(u \rightarrow o)}{\text{Count}(u)}$$

	a	b	c	d
X	2/6	2/6	1/6	1/6
Y	2/4	0	1/4	1/4
Z	1/8	5/8	1/8	1/8

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
$(?, ?)$	$(\mathbf{a}, \mathbf{d})$

Q2 - each equation is worth 1 pt.

Answer

- Base case:

$$\pi(0, \texttt{START}) = 1, \quad \text{otherwise} \quad \pi(0, v) = 0 \ \text{ if } \ v \neq \texttt{START} \tag{1}$$

- Moving forward:

$k = 1$

$$\pi(1, X) = a_{\texttt{START},X} \times b_X(b) = 3/6 \times 2/6 = 1/6 \tag{2}$$
$$\pi(1, Y) = a_{\texttt{START},Y} \times b_Y(b) = 0 \tag{3}$$
$$\pi(1, Z) = a_{\texttt{START},Z} \times b_Z(b) = 3/6 \times 1/8 = 1/16 \tag{4}$$

$k = 2$

$$\begin{aligned} \pi(2, X) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,X} \times b_X(b)\} \\ &= \max\{1/6 \times 1/6 \times 1/6, \quad 0, \quad 1/16 \times 1/8 \times 1/6\} \\ &= 1/216 \end{aligned} \tag{5}$$

$$\begin{aligned} \pi(2, Y) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,Y} \times b_Y(b)\} \\ &= \max\{1/6 \times 0 \times 1/4, \quad 0, \quad 1/16 \times 4/8 \times 1/4\} \\ &= 1/128 \end{aligned} \tag{6}$$

$$\begin{aligned} \pi(2, Z) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,Z} \times b_Z(b)\} \\ &= \max\{1/6 \times 4/6 \times 1/8, \quad 0, \quad 1/16 \times 1/8 \times 1/8\} \\ &= 1/72 \end{aligned} \tag{7}$$

$k = 3$

$$\begin{aligned} \pi(3, \texttt{STOP}) &= \max_{u \in \mathcal{T}} \{\pi(2, u) \times a_{u,\texttt{STOP}}\} \\ &= \max\{1/216 \times 1/6, 1/128 \times 3/4, 1/72 \times 2/8\} \\ &= 3/512 \end{aligned} \tag{8}$$

- Backtracking:

$$y_2^* = \arg\max_{v \in \mathcal{T}} \{\pi(2, v) \times a_{v,\texttt{STOP}}\} = Y \tag{9}$$

$$y_1^* = \arg\max_{v \in \mathcal{T}} \{\pi(1, v) \times a_{v,Y}\} = Z \tag{10}$$

Therefore, the optimal sequence is: Z, Y.

3. (20 pts) The HMM discussed in class makes a simple first-order assumption, where the next state only depends on the previous state in the generative process. However, it is possible to extend the model discussed in class to have second-order dependencies. In other words, the HMM can be parameterised in the following way:

$$p(x_1, \ldots, x_n, y_1, y_2, \ldots, y_n) = \prod_{i=1}^{n+1} p(y_i | y_{i-2}, y_{i-1}) \times \prod_{i=1}^{n} p(x_i | y_i)$$

where we define $y_{-1} = y_0 = START$ and $y_{n+1} = STOP$.

In other words, the transition probabilities are charged from $p(y_i|y_{i-1})$ to $p(y_i|y_{i-2}, y_{i-1})$ now. Describe the Viterbi algorithm used for decoding such a second-order HMM model. In other words, describe the dynamic programming algorithm that computes the following efficiently for such an HMM:

$$(y_1^*, y_2^*, \ldots, y_n^*) = \underset{y_1, y_2, \ldots, y_n}{\text{argmax}} \; p(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n)$$

Q3 - each equation in base case is worth 2pts and each equation in Eqs.(13)-(16) is worth 4pts.

Answer

- Base case:

$$\pi(-1, \texttt{START}) = 1, \quad \text{otherwise} \quad \pi(-1, v) = 0 \;\text{ if }\; v \neq \texttt{START} \tag{11}$$

$$\pi(0, \texttt{START}) = 1, \quad \text{otherwise} \quad \pi(0, v) = 0 \;\text{ if }\; v \neq \texttt{START} \tag{12}$$

- Moving forward recursively
  For $1 \leq k \leq n$

$$\pi(k, w) = \max_{v \in \mathcal{T}} \{\pi(k-1, v) \times a_{u,v,w} \times b_w(x_k)\} \tag{13}$$

- Final transition
  From y_n to $\texttt{STOP}$, the transition:

$$\pi(n+1, \texttt{STOP}) = \max_{w \in \mathcal{T}} \{\pi(n, w) \times a_{v,w,\texttt{STOP}}\} \tag{14}$$

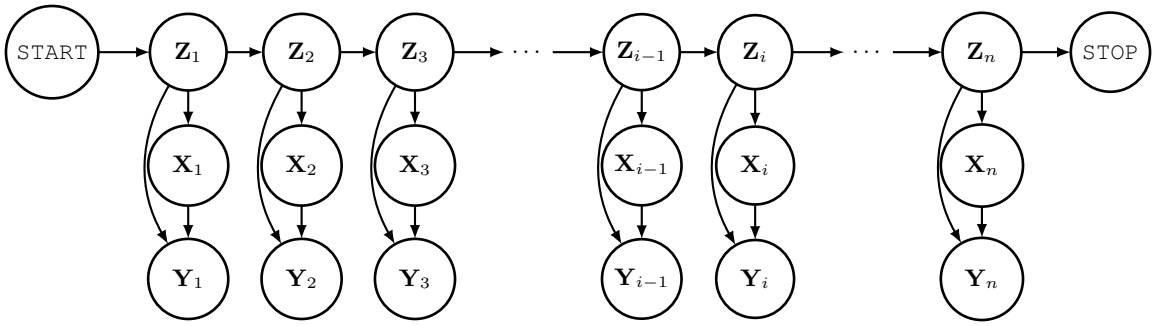
- Backtracking

$$y_n^* = \arg\max_{w \in \mathcal{T}} \{\pi(n, w) \times a_{v,w,\texttt{STOP}}\} \tag{15}$$

For $1 \leq k \leq n-1$,

$$y_k^* = \arg\max_{v \in \mathcal{T}} \{\pi(k, v) \times a_{u,v,y_{k+1}^*}\} \tag{16}$$

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 both the $\mathbf{X}$ sequence and $\mathbf{Z}$ sequence.



Assume you are given a large collection of observation pair sequences, and a predefined set of possible states $\{0, 1, \ldots, N-1, N\}$, where $0 = \texttt{START}$ and $N = \texttt{STOP}$. You would like to estimate the most probable state sequence for each observation pair sequence using an algorithm similar to the dynamic programming algorithm discussed in class. Clearly define the forward and backward scores in a way analogous to HMM. Give algorithms for computing the forward and backward scores. Analyze the time complexity associated with your algorithms (for an observation pair sequence of length n).

Q4 - each equation in Eqs.(17) - (20) and Eqs.(24) and (25) is worth 2 pts. Time complexity is worth 2 pts. In Eqs.(21) - (23), $c_v(y)$ is worth 3 pts, $b_v(x)$ is worth 2 pts, and the whole equation in moving forward part is worth 1 pt.

Answer Assume we have a set of possible states $\{0, 1, \ldots, N-1, N\}$ where $0 = \texttt{START}$ and $N = \texttt{STOP}$.

$$\begin{aligned} & P(x_1, \ldots, x_n, y_1, \ldots, y_n, z_i = u, x_i, \ldots, x_n, y_i, \ldots, y_n; \theta) \\ &= P(x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, z_i = u; \theta) \times P(x_i, \ldots, x_n, y_i, \ldots, y_n | z_i = u; \theta) \\ &= \alpha_u(i)\beta_u(i) \end{aligned} \tag{17}$$

where

$$\alpha_u(i) = P(x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, z_i = u; \theta) \tag{18}$$

$$\beta_u(i) = P(x_i, \ldots, x_n, y_i, \ldots, y_n | z_i = u; \theta) \tag{19}$$

Forward

- Base Case

$$\alpha_u(1) = a_{\texttt{START},u}, \quad \forall \ u \in \{1, \ldots, N-1\} \tag{20}$$

- Moving forward

For $i = 1, \ldots, n-1$:

$$\alpha_u(i+1) = \sum_v \alpha_v(i) \times a_{v,u} \times b_v(x_i) \times c_v(y_i) \tag{21}$$

where

$$b_v(x) = P(x|v) \tag{22}$$

$$c_v(y) = P(y|v, x) \tag{23}$$

Backward

- Base case

$$\beta_u(n) = a_{u,\texttt{STOP}} \times b_u(x_n) \times c_u(y_n) \quad \forall u = 1, \ldots, N-1 \tag{24}$$

- Moving forward
  For $i = n-1, \ldots, 1$:

$$\beta_u(i) = \sum_v a_{u,v} \times b_u(x_i) \times c_u(y_i) \times \beta_v(i+1) \tag{25}$$

At each time step/position, there are N forward (α) and N backward (β) terms to compute. To compute each term, there are $O(N)$ operations. Thus, at each time step/position, there are $O(N^2)$ operations. The length of sentence is n, which is the number of different time steps/positions. Hence, the total complexity is $O(nN^2)$.