SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

01.112 Machine Learning, Spring 2018
Homework 4

Due 4 April 2018, 5pm

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.

Question 1. Assume that we have the following training data available for us to estimate the model parameters:

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

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. *(10 points)*

Answer The transition probabilities are estimated as:

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

	X	Y	Z	STOP
START	0.5	0	0.5	0
X	0	0.4	0.4	0.2
Y	0.2	0	0.2	0.6
Z	0.4	0.6	0	0

The emission probabilities are estimated as:

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

	a	b	c
X	0.2	0.6	0.2
Y	0.4	0	0.6
Z	0.2	0.4	0.4

Question 2. 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. *(10 points)*

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

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:

  $k = 1$

$$\pi(1, X) = a_{\texttt{START},X} \times b_X(a) = 1 \times 0.5 \times 0.2 = 0.1 \tag{2}$$
$$\pi(1, Y) = a_{\texttt{START},Y} \times b_Y(a) = 0 \tag{3}$$
$$\pi(1, Z) = a_{\texttt{START},Z} \times b_Z(a) = 1 \times 0.5 \times 0.2 = 0.1 \tag{4}$$

  $k = 2$

$$\begin{aligned}\pi(2, X) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,X} \times b_X(a)\} \\ &= \max\{0.1 \times 0 \times 0.2, \quad 0 \times 0.2 \times 0.2, \quad 0.1 \times 0.4 \times 0.2\} \\ &= 0.008 \end{aligned} \tag{5}$$
$$\begin{aligned}\pi(2, Y) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,Y} \times b_Y(a)\} \\ &= \max\{0.1 \times 0.4 \times 0.4, \quad 0 \times 0 \times 0.4, \quad 0.1 \times 0.6 \times 0.4\} \\ &= 0.024 \end{aligned} \tag{6}$$
$$\begin{aligned}\pi(2, Z) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,Z} \times b_Z(a)\} \\ &= \max\{0.1 \times 0.4 \times 0.2, \quad 0 \times 0.2 \times 0.2, \quad 0.1 \times 0 \times 0.2\} \\ &= 0.008 \end{aligned} \tag{7}$$

  $k = 3$

$$\begin{aligned}\pi(3, \texttt{STOP}) &= \max_{u \in \mathcal{T}} \{\pi(2, u) \times a_{u,\texttt{STOP}}\} \\ &= \max\{0.008 \times 0.2, 0.024 \times 0.6, 0.008 \times 0\} \\ &= 0.0144 \end{aligned} \tag{8}$$

- 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.

Question 3. The Viterbi algorithm discussed in class is used for finding the optimal $\mathbf{y}$ sequence based on the following:

$$y_1^*, \ldots, y_n^* = \arg\max_{y_1,\ldots,y_n} p(x_1, \ldots, x_n, y_1, \ldots, y_n)$$

Now, consider the problem of part-of-speech tagging. Sometimes we have some prior knowledge about certain tags for certain observations. For example, assume the observation x_i="*the*", we are almost certain that it is not a verb (*i.e.*, we believe $y_i \neq \texttt{V}$). In this case, we would like to do the decoding in the following way, where we would like to incorporate the prior knowledge $y_i \neq \texttt{V}$ (and find optimal values for all other y_k in the sequence, where $k = 1, \ldots, n, k \neq i$):

$$y_1^*, \ldots, y_{i-1}^*, y_{i+1}^*, \ldots, y_n^* = \arg\max_{y_1,\ldots,y_{i-1},y_{i+1},\ldots,y_n} p(x_1, \ldots, x_n, y_1, \ldots, y_{i-1}, y_{i+1} \ldots, y_n | y_i \neq \texttt{V})$$

Clearly describe how to modify the Viterbi algorithm to perform such a new decoding task. *(10 points)*

Answer

- Base case:

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

- Moving forward recursively

  For $2 \leq k \leq i - 1$

$$\pi(k, v) = \max_{u \in \mathcal{T}} \{\pi(k - 1, u) \times a_{u,v} \times b_v(x_k)\} \tag{12}$$

  If $k = i$

$$\pi(i, v) = \max_{u \in \mathcal{T}} \{\pi(i - 1, u) \times a_{u,v} \times b_v(x_i)\}, \forall v \neq \texttt{V} \quad \text{otherwise} \quad \pi(i, \texttt{V}) = 0 \tag{13}$$

  For $k > i$

$$\pi(k, v) = \max_{u \in \mathcal{T}} \{\pi(k - 1, u) \times a_{u,v} \times b_v(x_k)\} \tag{14}$$

- Final transition

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

$$\pi(n + 1, \texttt{STOP}) = \max_{v \in \mathcal{T}} \{\pi(n, v) \times a_{v,\texttt{STOP}}\} \tag{15}$$

- Backtracking

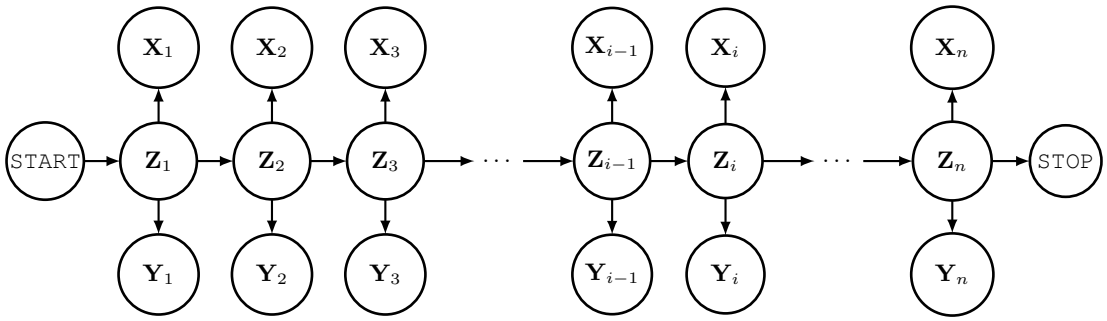
$$y_n^* = \arg\max_{v \in \mathcal{T}} \{\pi(n, v) \times a_{v,\texttt{STOP}}\} \tag{16}$$

$$\tag{17}$$

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

$$y_k^* = \arg\max_{u \in \mathcal{T}} \{\pi(k, u) \times a_{u,y_{k+1}^*}\} \tag{18}$$

Question 4. 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 $\mathbf{X}$ sequence is generated from the $\mathbf{Z}$ sequence and $\mathbf{Y}$ sequence is also generated from the $\mathbf{Z}$ 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 $0 = \textbf{START}$ and $N = \textbf{STOP}$. 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). *(10 points)*

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

$$\begin{aligned} &P(x_1, \ldots, x_n, y_1, \ldots, y_n, z_i = u, x_i, \ldots, x_n, y_i, \ldots, y_n; \theta) \\ &= P(x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, z_i = u; \theta) \times P(x_i, \ldots, x_n, y_i, \ldots, y_n | z_i = u; \theta) \\ &= \alpha_u(i)\beta_u(i) \end{aligned} \tag{19}$$

where

$$\alpha_u(i) = P(x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, z_i = u; \theta) \tag{20}$$

$$\beta_u(i) = P(x_i, \ldots, x_n, y_i, \ldots, y_n | z_i = u; \theta) \tag{21}$$

Forward

- Base Case

$$\alpha_u(1) = a_{\texttt{START},u}, \quad \forall \ u \in \{1, \ldots, N-1\} \tag{22}$$

- Moving forward

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

$$\alpha_u(i+1) = \sum_v \alpha_v(i) \times a_{v,u} \times b_v(x_i) \times c_v(y_i) \tag{23}$$

where

$$b_v(x) = P(x|v) \tag{24}$$

$$c_v(y) = P(y|v) \tag{25}$$

Backward

- Base case

$$\beta_u(n) = a_{u,\texttt{STOP}} \times b_u(x_n) \times c_u(y_n) \quad \forall u = 1, \ldots, N-1 \tag{26}$$

- Moving forward

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

$$\beta_u(i) = \sum_v a_{u,v} \times b_u(x_i) \times c_u(y_i) \times \beta_v(i+1) \tag{27}$$

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 $O(nN^2)$.