SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

50.007 Machine Learning, Fall 2015
Homework 4

Due Monday 23 Nov 2015, 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.

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

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

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.

Answer The transition probabilities are estimated as:

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

	X	Y	Z	STOP
START	1/2	0	1/2	0
X	2/7	0	3/7	2/7
Y	1/4	0	0	3/4
Z	1/7	4/7	1/7	1/7

The emission probabilities are estimated as:

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

	a	b	c	d
X	3/7	2/7	2/7	0
Y	1/2	0	1/4	1/4
Z	1/7	4/7	1/7	1/7

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

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:

$k = 1$

$$\pi(1, X) = a_{\texttt{START},X} \times b_X(a) = 1 \times 1/2 \times 3/7 = 3/14 \tag{2}$$
$$\pi(1, Y) = a_{\texttt{START},Y} \times b_Y(a) = 0 \times 1/2 = 0 \tag{3}$$
$$\pi(1, Z) = a_{\texttt{START},Z} \times b_Z(a) = 1 \times 1/2 \times 1/7 = 1/14 \tag{4}$$

$k = 2$

$$\begin{aligned}\pi(2, X) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,X} \times b_X(d)\} \\ &= \max\{3/14 \times 2/14 \times 0, \quad 0 \times 1/4 \times 0, \quad 1/4 \times 1/7 \times 0\} \\ &= 0 \end{aligned} \tag{5}$$

$$\begin{aligned}\pi(2, Y) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,Y} \times b_Y(d)\} \\ &= \max\{3/14 \times 0 \times 1/4, \quad 0 \times 0 \times 1/4, \quad 1/14 \times 4/7 \times 1/4\} \\ &= 1/98 \end{aligned} \tag{6}$$

$$\begin{aligned}\pi(2, Z) &= \max_{v \in \mathcal{T}} \{\pi(1, v) \times a_{v,Z} \times b_Z(d)\} \\ &= \max\{3/14 \times 3/7 \times 1/7, \quad 0 \times 0 \times 1/7, \quad 1/14 \times 1/7 \times 1/7\} \\ &= 9/686 \end{aligned} \tag{7}$$

$k = 3$

$$\begin{aligned}\pi(3, \texttt{STOP}) &= \max_{v \in \mathcal{T}} \{\pi(2, v) \times a_{v,\texttt{STOP}}\} \\ &= \max\{0 \times 2/7, 1/98 \times 3/4, 9/686 \times 1/7\} \\ &= 3/392 \end{aligned} \tag{8}$$

- 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 $y_{-1} = y_0 = \texttt{START}$ and $y_{n+1} = \texttt{STOP}$.

In other words, the transition probabilities are changed 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:

$$(y_1^*, y_2^*, \ldots, y_n^*) = \arg\max_{y_1, y_2, \ldots, y_n} p(x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n)$$

Answer The key idea is to consider a pair of states $\langle u, v \rangle$ instead of a single state v at each time step, where v is the current state, and u is the last state (the state at the previous time step).

Define $a_{t,u,v} \equiv p(y_i = v | y_{i-2} = t, y_{t-1} = u)$, and $b_u(o) \equiv p(x_i = o | y_i = u)$.

- Base cases:

$$\pi(0, \langle \texttt{START}, \texttt{START} \rangle) = 1 \tag{11}$$

$$\pi(0, \langle u, v \rangle) = 0, \quad \text{if } \langle u, v \rangle \neq \langle \texttt{START}, \texttt{START} \rangle \tag{12}$$

$$\pi(1, \langle \texttt{START}, v \rangle) = a_{\texttt{START},\texttt{START},v} \times b_v(x_1) \quad \forall v \tag{13}$$

$$\pi(1, \langle u, v \rangle) = 0, \quad \text{if } u \neq \texttt{START} \tag{14}$$

- Moving forward recursively
  For $k \geq 2$

$$\pi(k, \langle u, v \rangle) = \max_{t \in \mathcal{T}} \{\pi(k-1, \langle t, u \rangle) \times a_{t,u,v} \times b_v(x_k)\} \tag{15}$$

- Final transition
  From $\langle y_{n-1}, y_n \rangle$ to $\texttt{STOP}$, the transition:

$$\pi(n+1, \langle u, \texttt{STOP} \rangle) = \max_{t \in \mathcal{T}} \{\pi(n, \langle t, u \rangle) \times a_{t,u,\texttt{STOP}}\} \tag{16}$$

- Backward tracking

$$y_n^* = \arg\max_u \{\pi(n+1, \langle u, \texttt{STOP}\rangle)\} \tag{17}$$

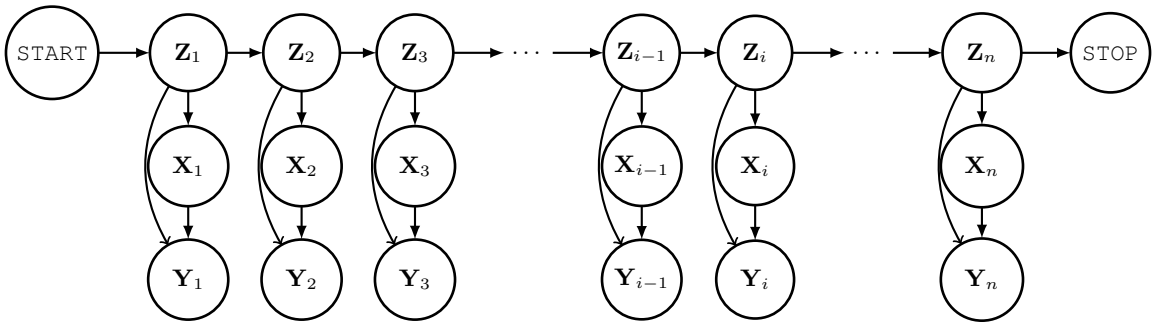
$$\begin{aligned} y_k^* &= \arg\max_t \{\pi(k+1, \langle t, y_{k+1}^*\rangle) \times a_{t,y_{k+1}^*,y_{k+2}^*} \times b_{y_{k+2}^*}(x_{k+2})\} \\ &= \arg\max_t \{\pi(k+1, \langle t, y_{k+1}^*\rangle) \times a_{t,y_{k+1}^*,y_{k+2}^*}\} \end{aligned} \tag{18}$$

Note: alternatively, if you do not compute $\pi(n+1, \langle u, \texttt{STOP}\rangle)$, that is also fine. You can obtain the last two optimal states from the following:

$$\langle y_{n-1}^*, y_n^*\rangle = \arg\max_{u,v} \{\pi(n, \langle u, v\rangle) \times a_{u,v,\texttt{STOP}}\} \tag{19}$$

Why is it so? Think about it! Also think about whether this alternative approach would be faster!

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).

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

$$\begin{aligned} &P(x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, z_i = u, x_i, \ldots, x_n, y_i, \ldots, y_n; \theta) \\ &= P(x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, z_i = u; \theta) \times P(x_i, \ldots, x_n, y_i, \ldots, y_n | z_i = u; \theta) \\ &= \alpha_u(i)\beta_u(i) \end{aligned} \tag{20}$$

where

$$\alpha_u(i) = P(x_1, \ldots, x_{i-1}, y_1, \ldots, y_{i-1}, z_i = u; \theta) \tag{21}$$

$$\beta_u(i) = P(x_i, \ldots, x_n, y_i, \ldots, y_n | z_i = u; \theta) \tag{22}$$

Forward

- Base Case

$$\alpha_u(1) = a_{\texttt{START},u}, \quad \forall \ u \in \{1, \ldots, N-1\} \tag{23}$$

- Moving forward
  For $i = 2, \ldots, n$:

$$\alpha_u(i+1) = \sum_v \alpha_v(i) \times a_{v,u} \times b_v(x_i) \times c_{v,x_i}(y_i) \tag{24}$$

where

$$c_{v,x}(y) = P(y|v,x) \tag{25}$$

Backward

- Base case

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

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

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

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 $O(nN^2)$.