Name:	**Student ID:**
Pillar: ISTD/ESD/EPD/ASD	

01.112 Machine Learning
Final Exam
2017 Fall Term

Date: 15 Dec 2017
Start Time: 9:00AM
Duration: 2 hours
(Sample Solutions)

Instructions:

1. Write your name, student ID, and pillar information at the top of this page.
2. This paper consists of 4 main questions and 12 printed pages.
3. The problems are not necessarily in order of difficulty. We recommend that you scan through all the questions first, and then decide on the order to answer them.
4. Write your answers in the space provided.
5. You may refer to your one-sided A4-sized cheat sheet.
6. You are allowed to use non-programmable calculators.
7. You may NOT refer to any other material.
8. You may NOT access the Internet.
9. You may NOT communicate via any means with anyone (aside from the invigilators).

For staff's use:

Qs 1	/8
Qs 2	/40
Qs 3	/20
Qs 4	/12
Total	**/80**

Question 1. (8 points)
Please indicate whether the following statements are true (**T**) or false (**F**).

1. The soft-EM algorithm comes with an guarantee: it will always converge to a global optimum of the objective function, no matter how you do the initialization. *(1 point)*

   Answer : F

2. One task that we can perform with a hidden Markov model (HMM) is to predict the underlying part-of-speech sequence for each input sentence. Under this setting, we can say that the HMM can help us find a mapping between two structured spaces: the input space consists of all possible sentences and the output space consists of all possible part-of-speech tag sequences. *(1 point)*

   Answer : T

3. Hidden Markov models can be regarded as a special class of Bayesian networks. *(1 point)*

   Answer : T

4. The Markov decision process (MDP) discussed in class typically makes some uncertainty assumptions. Specifically, it assumes that the robot does not know exactly its current state at any time. *(1 point)*

   Answer : F

5. The space complexity of the forward-backward algorithm is $O(nT)$ for each instance, where n is the length of the input sequence, and T is the number of possible output states at each position. *(1 point)*

   Answer : T

6. The forward-backward algorithm involves two procedures that perform the calculation of the forward scores and the backward scores respectively. In practice, as these two procedures do not depend on each other, one can choose to calculate the backward scores first and then the forward scores. *(1 point)*

   Answer : T

7. The Viterbi algorithm discussed in class can be extended to support second-order HMMs as we have discussed in homework 4, but the time complexity would become $O(nT^3)$ (n and T are defined above in question 5). *(1 point)*

   Answer : T

8. The Markov blanket of a node A consists of A's parents, children and siblings in the Bayesian network. *(1 point)*

   Answer : F

Question 2. (40 points)
In this problem, we would like to look at the hidden Markov model (HMM).

(a) Assume that we have the following training set available for us to estimate the model parameters:

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

Clearly state what are the parameters associated with the HMM. Under the maximum likelihood estimation (MLE), what would be the optimal model parameters? Fill up the following transition and emission probability tables. Use the space below to clearly show how each **emission parameter** is estimated exactly. *(10 points)*

$a_{u,v}$ $u\backslash v$	**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

$b_u(o)$ $u\backslash o$	**a**	**b**	**c**
X	0.4	0.6	0
Y	0.4	0	0.6
Z	0.2	0.6	0.2

The emission probabilities are estimated as:

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

$$b_X(a) = \frac{\text{Count}(X \to a)}{\text{Count}(X)} = \frac{2}{5} = 0.4$$
$$b_X(b) = \frac{\text{Count}(X \to b)}{\text{Count}(X)} = \frac{3}{5} = 0.6$$
$$b_X(c) = \frac{\text{Count}(X \to c)}{\text{Count}(X)} = \frac{0}{5} = 0$$
$$b_Y(a) = \frac{\text{Count}(Y \to a)}{\text{Count}(Y)} = \frac{2}{5} = 0.4$$
$$b_Y(b) = \frac{\text{Count}(Y \to b)}{\text{Count}(Y)} = \frac{0}{5} = 0$$
$$b_Y(c) = \frac{\text{Count}(Y \to c)}{\text{Count}(Y)} = \frac{3}{5} = 0.6$$
$$b_Z(a) = \frac{\text{Count}(Z \to a)}{\text{Count}(Z)} = \frac{1}{5} = 0.2$$
$$b_Z(b) = \frac{\text{Count}(Z \to b)}{\text{Count}(Z)} = \frac{3}{5} = 0.6$$
$$b_Z(c) = \frac{\text{Count}(Z \to c)}{\text{Count}(Z)} = \frac{1}{5} = 0.2$$

(b) Now, consider in the test phase, you are given the following new observation sequence and the following parameters, 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)*

$a_{u,v}$ $u\backslash v$	**X**	**Y**	**Z**	STOP
START	0.5	0.4	0.1	0.0
X	0.1	0.3	0.3	0.3
Y	0.2	0.2	0.2	0.4
Z	0.1	0.6	0.1	0.2

$b_u(o)$ $u\backslash o$	**a**	**b**	**c**
X	0.3	0.3	0.4
Y	0.2	0.3	0.5
Z	0.2	0.4	0.4

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

Answer

- Base case:

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

- Moving forward:

  $k = 1$

$$\pi(1, X) = a_{\texttt{START},X} \times b_X(a) = 0.5 \times 0.3 = 0.15 \tag{2}$$
$$\pi(1, Y) = a_{\texttt{START},Y} \times b_Y(a) = 0.4 \times 0.2 = 0.08 \tag{3}$$
$$\pi(1, Z) = a_{\texttt{START},Z} \times b_Z(a) = 0.1 \times 0.2 = 0.02 \tag{4}$$

  $k = 2$

$$\begin{aligned}\pi(2, X) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,X} \times b_X(c)\} \\ &= \max\{0.15 \times 0.1 \times 0.4, \quad 0.08 \times 0.2 \times 0.4, \quad 0.02 \times 0.1 \times 0.4\} \\ &= 0.0064 \end{aligned} \tag{5}$$

$$\begin{aligned}\pi(2, Y) &= \max_{u \in \mathcal{T}} \{\pi(1, u) \times a_{u,Y} \times b_Y(c)\} \\ &= \max\{0.15 \times 0.3 \times 0.5, \quad 0.08 \times 0.2 \times 0.5, \quad 0.02 \times 0.6 \times 0.5\} \\ &= 0.0225 \end{aligned} \tag{6}$$

$$\begin{aligned}\pi(2, Z) &= \max_{v \in \mathcal{T}} \{\pi(1, v) \times a_{v,Z} \times b_Z(c)\} \\ &= \max\{0.15 \times 0.3 \times 0.4, \quad 0.08 \times 0.2 \times 0.4, \quad 0.02 \times 0.1 \times 0.4\} \\ &= 0.018 \end{aligned} \tag{7}$$

  $k = 3$

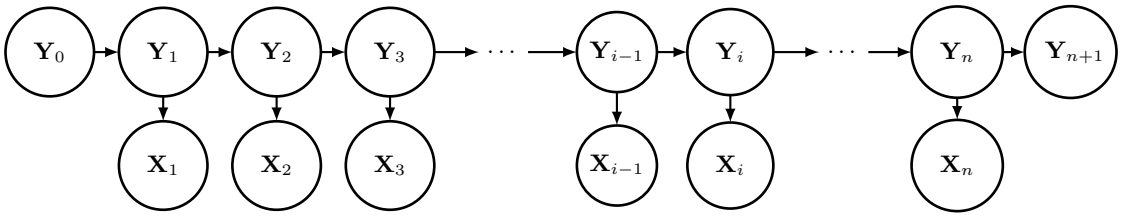
$$\begin{aligned}\pi(3, \texttt{STOP}) &= \max_{v \in \mathcal{T}} \{\pi(2, v) \times a_{v,\texttt{STOP}}\} \\ &= \max\{0.0064 \times 0.3, 0.0225 \times 0.4, 0.018 \times 0.2\} \\ &= 0.009 \end{aligned} \tag{8}$$

- Backtracking:

$$y_2^* = \arg\max_{v \in \mathcal{T}} \{\pi(2, v) \times a_{v,\texttt{STOP}}\} = Y \tag{9}$$
$$y_1^* = \arg\max_{v \in \mathcal{T}} \{\pi(1, v) \times a_{v,Y}\} = X \tag{10}$$

  Therefore, the optimal sequence is: X, Y.

(c) Recall that in the HMM discussed in class, the graphical model can be illustrated as follows:

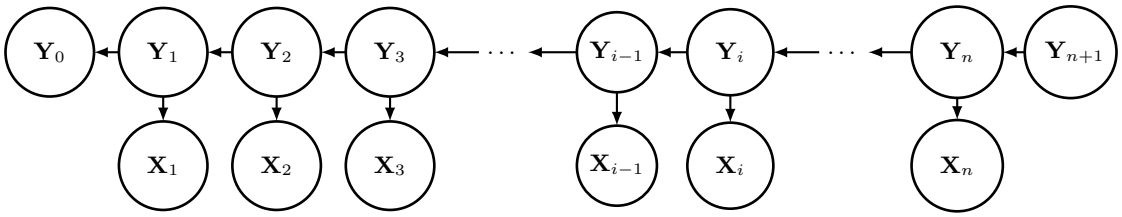


The joint probability for each input-output sequence pair is defined as:

$$p(X_1, \ldots, X_n, Y_0, Y_1, \ldots, Y_n, Y_{n+1}) = \prod_{j=0,1,\ldots,n} p(Y_{j+1}|Y_j) \prod_{j=1,2,\ldots,n} p(X_j|Y_j)$$

where $Y_0 = \texttt{START}$ and $Y_{n+1} = \texttt{STOP}$. Here, the first group of terms are the transition probabilities ($a_{u,v}$) and the second group of terms are emission probabilities ($b_u(o)$).

Now, imagine we would like to take an alternative approach to model how the input and output sequences are jointly generated. Specifically, we assume the input and output sequences are generated from the right to the left (*i.e.*, the last state is generated first, followed by the last observation, ...). The graphical illustration of the generative procedure is now as follows:



This means the joint probability is now defined as follows:

$$p(X_1, \ldots, X_n, Y_0, Y_1, \ldots, Y_n, Y_{n+1}) = \prod_{j=n,n-1,\ldots,0} p(Y_j|Y_{j+1}) \prod_{j=n,n-1,\ldots,1} p(X_j|Y_j)$$

where $Y_0 = \texttt{START}$ and $Y_{n+1} = \texttt{STOP}$. Here, the first group of terms are the transition probabilities ($a_{u,v}$) and the second group of terms are emission probabilities ($b_u(o)$).

Now, based on the above new assumption, estimate the model parameters **based on the same training set given in (a)**. *(5 points)*

$a_{u,v}$ $u\backslash v$	**X**	**Y**	**Z**	STOP
START	0.25	0.75	0	0
X	0	0.2	0.4	0.4
Y	0.4	0	0.6	0
Z	0.4	0.2	0	0.4

$b_u(o)$ $u\backslash o$	**a**	**b**	**c**
X	0.4	0.6	0
Y	0.4	0	0.6
Z	0.2	0.6	0.2

(d) Under this reversed HMM described in (c), find the most probable state sequence for the observation sequence $(\mathbf{a}, \mathbf{c})$ based on the following probability tables.

$a_{u,v}$ $u\backslash v$	**X**	**Y**	**Z**	STOP
START	0.2	0.3	0.5	0.0
X	0.2	0.2	0.4	0.2
Y	0.2	0.1	0.2	0.5
Z	0.4	0.3	0.1	0.2

$b_u(o)$ $u\backslash o$	**a**	**b**	**c**
X	0.4	0.5	0.1
Y	0.4	0.1	0.5
Z	0.2	0.6	0.2

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

(Hint: here, you may want to run an alternative version of the Viterbi algorithm, which runs from the right to the left, followed by a back-tracking from the left to the right.) *(10 points)*

Answer

- Base case:

$$\pi(3, \texttt{START}) = 1, \quad \text{otherwise} \quad \pi(3, v) = 0 \ \text{ if } \ v \neq \texttt{START} \tag{11}$$

- Moving backward:

$k = 2$

$$\pi(2, X) = a_{\texttt{START},X} \times b_X(\mathbf{c}) = 0.2 \times 0.1 = 0.02 \tag{12}$$
$$\pi(2, Y) = a_{\texttt{START},Y} \times b_Y(\mathbf{c}) = 0.3 \times 0.5 = 0.15 \tag{13}$$
$$\pi(2, Z) = a_{\texttt{START},Z} \times b_Z(\mathbf{c}) = 0.5 \times 0.2 = 0.1 \tag{14}$$

$k = 1$

$$\begin{aligned}\pi(1, X) &= \max_{u \in \mathcal{T}} \{\pi(2, u) \times a_{u,X} \times b_X(\mathbf{a})\} \\ &= \max\{0.02 \times 0.2 \times 0.4, \quad 0.15 \times 0.2 \times 0.4, \quad 0.1 \times 0.4 \times 0.4\} \\ &= 0.016\end{aligned} \tag{15}$$

$$\begin{aligned}\pi(1, Y) &= \max_{u \in \mathcal{T}} \{\pi(2, u) \times a_{u,Y} \times b_Y(\mathbf{a})\} \\ &= \max\{0.02 \times 0.2 \times 0.4, \quad 0.15 \times 0.1 \times 0.4, \quad 0.1 \times 0.3 \times 0.4\} \\ &= 0.012\end{aligned} \tag{16}$$

$$\begin{aligned}\pi(1, Z) &= \max_{v \in \mathcal{T}} \{\pi(2, v) \times a_{v,Z} \times b_Z(\mathbf{a})\} \\ &= \max\{0.02 \times 0.4 \times 0.2, \quad 0.15 \times 0.2 \times 0.2, \quad 0.1 \times 0.1 \times 0.2\} \\ &= 0.006\end{aligned} \tag{17}$$

$k = 0$

$$\begin{aligned}\pi(0, \texttt{STOP}) &= \max_{v \in \mathcal{T}} \{\pi(1, v) \times a_{v,\texttt{STOP}}\} \\ &= \max\{0.016 \times 0.2, 0.012 \times 0.5, 0.006 \times 0.2\} \\ &= 0.006\end{aligned} \tag{18}$$

- Backtracking:

$$y_1^* = \arg\max_{v \in \mathcal{T}} \{\pi(1, v) \times a_{v,\texttt{STOP}}\} = Y \tag{19}$$

$$y_2^* = \arg\max_{v \in \mathcal{T}} \{\pi(2, v) \times a_{v,Y}\} = Z \tag{20}$$

  Therefore, the optimal sequence is: Y, Z.

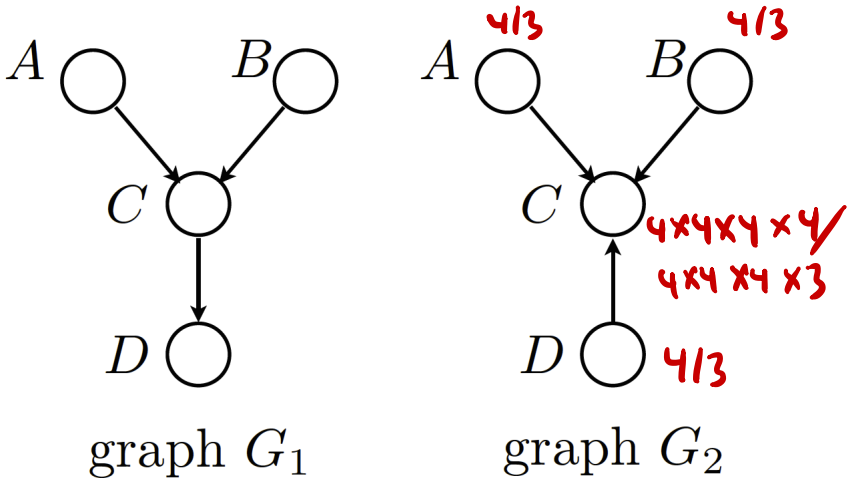
(e) Now, imagine that you have already implemented the training and decoding procedures for the original standard HMM. You would like to implement such procedures for the reversed HMM. Before you started the implementation of the new model's training and decoding procedures from scratch, someone told you that there were tricks that allow you to build such a reversed HMM very quickly without involving a significant amount of coding efforts. Try to come up with a good method that allows you to build such a reversed HMM very quickly. Describe the idea clearly and concisely. *(5 points)*.

Answer

(a) During training, simply pass the reversed state sequences and observation sequences to the implemented training procedure.

(b) During decoding, first reverse the given input sentence and pass it to implemented decoder for decoding (using the learned model parameters above), and then output the reversed output state sequence.

Question 3. (20 points)
This question deals with traffic congestion, an important problem in Singapore. Suppose we have four discrete random variables, **A**, **B**, **C**, and **D** corresponding to levels of traffic congestion in different locations in Singapore. Each of the random variables takes values in $\{0, 1, 2, 3\}$ denoting different levels of congestion respectively. We entertain here two alternative Bayesian network models over these variables. These are given as graphs G_1 and G_2 below.



(a) List one independence property that holds for graph G_1 but NOT for graph G_2. *(4 points)*

(Express your answers in the following form: "X is independent of Y given Z" or "X is independent of Y" where X, Y, Z are random variables.)

Possible answers are:
A independent of D given C
B independent of D given C
A, B independent of D given C
A independent of D given C, B
B independent of D given A, C

(b) List one independence property that holds for graph G_2 but NOT for graph G_1. *(4 points)*

(Express your answers in the following form: "X is independent of Y given Z" or "X is independent of Y" where X, Y, Z are random variables.)

A independent of D
B independent of D

(c) How many independent parameters do we need in order to specify the joint distribution associated with graph G_1? *(5 points)*

$P(A), P(B)$: 3 parameters each, 6 in total
$P(C|A,B)$: $3 \times 4^4 = 48$ parameters
$P(D|C)$: $3 \times 4 = 12$ parameters
total = 66

(d) Suppose we are given a dataset consisting of n complete observations of the four variables (n days worth of traffic assessments). We find maximum-likelihood parameter estimates for each of our alternative models, G_1 and G_2. The resulting maximum log-likelihood values turned out to be practically the same. Which model should we choose? Please choose (by ticking) one of the following answers. *(3 points)*

( ) G_1

( ) G_2

( ) either/both

Ans: G_1

(e) Which of the following rationales is correct for answering part (d)? *(4 points)*

(Note that the BIC is defined as a penalized likelihood.)

( ) Since the two models attain the same log-likelihood values, they should be considered equally good.

( ) The BIC score for G_1 would be larger than for G_2.

( ) The BIC score for G_2 would be larger than for G_1.

( ) Because the two models make different independence assumptions about the variables, yet attain the same log-likelihood of the data, we cannot statistically decide between them.

Ans: 2nd one, i.e., the BIC score for G_1 would be larger than for G_2.

Question 4. (12 points)
Consider the following Markov decision process (MDP). It has states $\{0, 1, 2, 3, 4, 5\}$ with 5 as the starting state. In every state, you can take one of two possible actions: walk (W) or jump (J). The Walk action decreases the state by one. The Jump action has probability 0.5 of decreasing the state by two, and probability 0.5 of leaving the state unchanged. Actions will not decrease the state below zero: you will remain in state 0 no matter which action you take (i.e., state 0 is a terminal state). Jumping in state 1 leads to state 0 with probability 0.5 and state 1 with probability 0.5. This definition leads to the following transition functions:

- For states $k \geq 1$, $T(k, W, k-1) = 1$
- For states $k \geq 2$, $T(k, J, k-2) = T(k, J, k) = 0.5$
- For state $k = 1$, $T(k, J, k-1) = T(k, J, k) = 0.5$
- For state $k = 0$, $T(k, J, k) = T(k, W, k) = 1$

The reward gained when taking an action is the distance travelled squared, i.e., $R(s, a, s') = (s - s')^2$. The discount factor is $\gamma = 0.6$.

(a) Suppose we initialize $Q_0^*(s, a) = 0$ for all $s \in \{0, 1, 2, 3, 4, 5\}$ and $a \in \{J, W\}$. Evaluate the Q-values $Q_1^*(s, a)$ after exactly one iteration of the Q-Value Iteration Algorithm. Write your answers in the table below. *(3 points)*

	$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
J	0	0.5	2	2	2	2
W	0	1	1	1	1	1

(b) What is the policy that we would derive from $Q_1^*(s, a)$? Answer by filling in the action that should be taken at each state in the table below. *(3 points)*

$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
W	J	J	J	J

(c) What are the values $V_1^*(s)$ corresponding to $Q_1^*(s, a)$? *(3 points)*

$s = 0$	$s = 1$	$s = 2$	$s = 3$	$s = 4$	$s = 5$
0	1	2	2	2	2

(d) Will the policy change after the second iteration? If your answer is "yes", briefly describe how. *(3 points)*

Ans: No.