COMP 5212 Machine Learning (2024 Spring)
Homework 3 Solution:
Hand out: April 18, 2024
Due: April 30, 2024, 11:59 PM
Total Points: 77

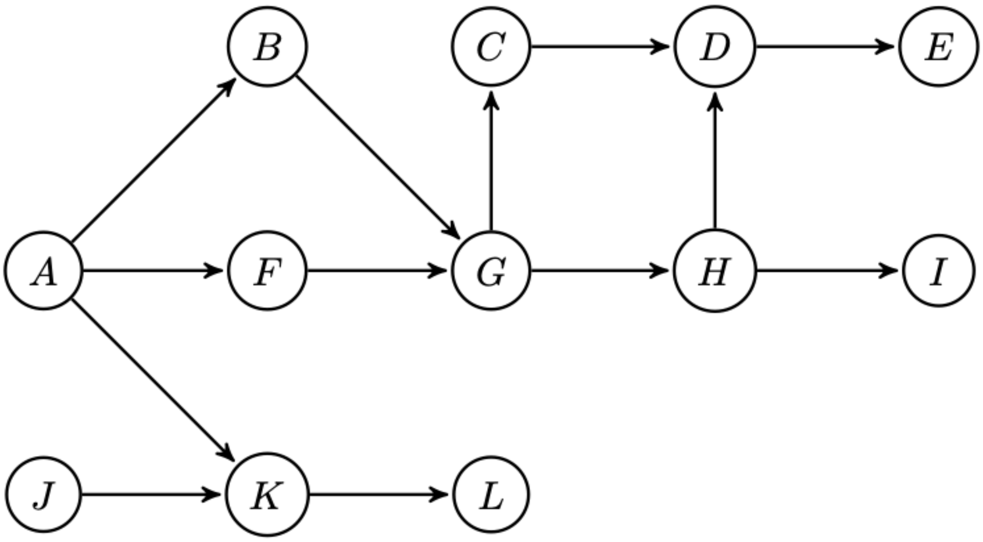
Your solution should contain below information at the top of its first page.

1. Your name
2. Your student id number

Some Notes:

- Homeworks will not be easy, please start early.
- For this homework, please submit a single PDF file to Canvas. Make sure to prepare your solution to each problem on a separate page.
- The total points of this homework is 100, and a score of 100 already gives you full grades on this homework. There are possibly bonus questions, you can optionally work on them if you are interested and have time.
- You can choose either using LaTeX by inserting your solutions to the problem pdf, or manually write your solutions on clean white papers and scan it as a pdf file – in the case of handwriting, please write clearly and briefly, we are not responsible to extract information from unclear handwriting. **We highly recommend you use LaTeX for the sake of any misunderstandings about the handwriting.** If your submission is a scan of a handwritten solution, make sure that it is of high enough resolution to be easily read. At least 300dpi and possibly denser.
- We encourage students to work in groups for homeworks, but the students need to write down the homework solutions or the code independently. In case that you work with others on the homework, please write down the names of people with whom you've discussed the homework. You are not allowed to copy, refer to, or look at the exact solutions from previous years, online, or other resources.
- **Late Policy:** 3 free late days in total across the semester, for additional late days, 20% penalization applied for each day late. **No assignment will be accepted more than 3 days late**.
- Please refer to the Course Logistics page for more details on the honor code and logisitcs. We have zero tolerance — in the case of honor code violation for a single time, you will fail this course directly.

1 Probabilistic Graphical Models (24 pts)



Which of the following statements are true given the network above? For true statements, brief explain why. For false statements, show one active trail using the Bayes Ball Algorithm.

1. $P(H, J) = P(H)P(J)$

True. The ball is blocked at K.

2. $P(H, J|L) = P(H|L)P(J|L)$

False. JKLKAFGH.

3. $P(C, I|F) = P(C|F)P(I|F)$

False. CGHI.

4. $P(C, I|G, E) = P(C|G, E)P(I|G, E)$

False. CDEHI.

5. $P(A, D|B) = P(A|B)P(D|B)$

False. AFGHD.

6. $P(B, F) = P(B)P(F)$

False. BAF.

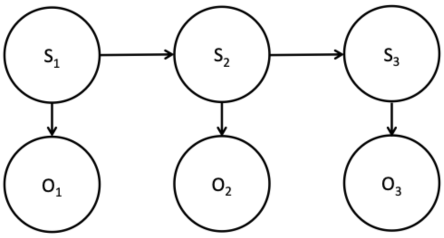
7. $P(C, K|B, F) = P(C|B, F)P(K|B, F)$

True. The ball is blocked at F or B.

8. $P(E, K|L) = P(E|L)P(K|L)$

False. KAFGHDE.

2 Hidden Markov Models (53 pts)



1. [**5 points**] Assume that we have the Hidden Markov Model (HMM) depicted in the figure above. If each of the states can take on k different values and a total of m different observations are possible for each state, how many parameters are required to fully define this HMM? Justify your answer.

HMM parameters have three components: initial probability distribution, transition probability distribution and the emission probability distribution. There are a total of k states, so k parameters are required to define the initial p probability distribution (we will ignore all of the -1s for this problem to make things cleaner).For the transition distribution, we can transition from any one of k states to any of the k states (including staying in the same state) , so k^2 parameters are required. Then, we need a total of km parameters for the emission probability distribution, since each of the k states can emit each of the m observations.
Thus, the total number of parameters required are k + k^2 + km. Note that the number of parameters does not depend on the length of the HMMs.

2. [**10 points**] In class, we have learned the forward algorithm to compute the probability of observations, $P(\{O_t\}_{t=1}^T)$. Please try to derive a different algorithm to compute $P(\{O_t\}_{t=1}^T)$ in the backward direction. Similar to the forward algorithm in the slides, you should indicate the initialization, the computation equations in the backward direction at each step, and the termination. (While the figure above only shows a sequence length of 3, please derive a general form here assuming the sequence length is T)

Algorithm 1 Backward Algorithm For Evaluation Problem

```
Initialise $\beta_T^k = 1$ for all $k$
for $t = T - 1$ to 1 do
    $\beta_t^k = \sum_i \beta_{t+1}^i p(S_{t+1} = i|S_t = k)p(O_{t+1}|S_{t+1} = i)$ for all k
end for
$P(\{O_t\})_{t=1}^T) = \sum_k \beta_1^k p(O_1|S_1 = k)p(S_1 = k)$
```

3. [**10 points**] Similar to the last question, in class, we have learned the forward-backward algorithm to compute the posterior distribution $P(S_t = k|\{O_t\}_{t=1}^T)$. Please try to derive a backward-forward version of it – where you eliminate the variables on the left side of S_t in the

backward direction, and the right side of S_t in the forward direction. Similar to that in the slides, you should indicate the initialization, the computation equations, and the termination. (While the figure above only shows a sequence length of 3, please derive a general form here assuming the sequence length is T)

Algorithm 2 Backward-Forward Algorithm For Decoding Problem

```
Initialise $\alpha_{t-1}^i = P(S_t = k|S_{t-1} = i)P(O_{t-1}|S_{t-1} = i)$ for all $i$
Initialise $\beta_{t+1}^i = P(S_{t+1} = i|S_t = k)P(O_{t+1}|S_{t+1} = i)$ for all $i$
for $x = t + 2$ to $T$ do
    $\beta_x^i = \sum_j \beta_{x-1}^j P(S_x = i|S_{x-1} = j)p(O_x|S_x = i)$
end for
for $x = t - 2$ to $1$ do
    $\alpha_x^i = \sum_j \alpha_{x+1}^j P(S_{x+1} = j|S_x = i)P(O_x|S_x = i)$
end for
Further Define $\alpha_0 = \sum_i \alpha_1^i P(S_1 = i)$
Further Define $\beta_{T+1} = \sum_i \beta_T^i$
$P(S_t = k, \{O_t\}_{t=1}^T) = \alpha_0 \beta_{T+1} \cdot P(O_t|S_t = k)$
$P(S_t = k|\{O_t\}_{t=1}^T) = \frac{P(S_t=k,\{O_t\}_{t=1}^T)}{P(\{O_t\}_{t=1}^T)} = \frac{\alpha_0 \beta_{T+1} \cdot P(O_t|S_t=k)}{\sum_k \alpha_0 \beta_{T+1} \cdot P(O_t|S_t=k)}$
For special case: if t = 1: We set $\alpha_0 = P(S_1 = k)$.
If t = T: we set $\beta_{T+1} = 1$.
```

Suppose that we have binary states (labeled A and B) and binary observations (labeled 0 and 1) and the initial, transition, and emission probabilities are as given in the table in Figure 1.

State	$P(S_1)$
A	0.99
B	0.01

(a) Initial probs.

S_1	S_2	$P(S_2\|S_1)$
A	A	0.99
A	B	0.01
B	A	0.01
B	B	0.99

(b) Transition probs.

S	O	$P(O\|S)$
A	0	0.8
A	1	0.2
B	0	0.1
B	1	0.9

(c) Emission probs.

Figure 1: Probabilities Table

4. [**7 points**] Using the forward algorithm, compute the probability that we observe the sequence $O_1 = 0, O_2 = 1,$ and $O_3 = 0$. Show your work (i.e., show each of your alphas, you can directly use the equations from the lecture).

$$\begin{aligned}
\alpha_1^A &= 0.8 \cdot 0.99 = 0.792 \\
\alpha_1^B &= 0.1 \cdot 0.001 = 0.001 \\
\alpha_2^A &= 0.2(0.792 \cdot 0.99 + 0.001 \cdot 0.01) = 0.156818 \\
\alpha_2^B &= 0.9(0.792 \cdot 0.01 + 0.001 \cdot 0.99) = 0.008019 \\
\alpha_3^A &= 0.8(0.156818 \cdot 0.99 + 0.008019 \cdot 0.01)\ 0.124264 \\
\alpha_3^B &= 0.1(0.156818 \cdot 0.01 + 0.008019 \cdot 0.99) = 0.000950699 \\
P(\{O_T\}_{t=1}^T) &= 0.1252147
\end{aligned}$$

5. [**7 points**] Using the backward algorithm you derived in Question 2 to compute the probability that we observe the aforementioned sequence ($O_1 = 0, O_2 = 1,$ and $O_3 = 0$). You can directly use the equation you just derived.

$$\begin{aligned}
\beta_3^A &= 1 \\
\beta_3^B &= 1 \\
\beta_2^A &= 0.99 \cdot 0.8 \cdot 1 + 0.01 \cdot 0.1 \cdot 1 = 0.793 \\
\beta_2^B &= 0.01 \cdot 0.08 \cdot 1 + 0.99 \cdot 0.01 \cdot 1 = 0.107 \\
\beta_1^A &= 0.8(0.156818 \cdot 0.99 + 0.008019 \cdot 0.01)\ 0.124264 \\
\beta_1^B &= 0.1(0.156818 \cdot 0.01 + 0.008019 \cdot 0.99) = 0.000950699 \\
P(\{O_T\}_{t=1}^T) &= 0.1252147
\end{aligned}$$

6. [**3 points**] Using the backward-forward algorithm you derived in Question 3, compute (and report) the most likely setting for each state.

For $P(S_1 = A|\{O_t\}_{t=1}^T)$:

$$\begin{aligned}
\beta_2^A &= P(S_2 = A|S_1 = A)P(O_2|S_2 = A) = 0.99 \cdot 0.2 = 0.198 \\
\beta_2^B &= P(S_2 = B|S_1 = A)P(O_2|S_2 = B) = 0.01 \cdot 0.9 = 0.009 \\
\beta_3^A &= 0.198 \cdot 0.99 \cdot 0.8 + 0.009 \cdot 0.01 \cdot 0.8 = 0.156888 \\
\beta_3^B &= 0.198 \cdot 0.01 \cdot 0.1 + 0.009 \cdot 0.99 \cdot 0.1 = 0.001089 \\
\beta_4 &= 0.156888 + 0.001089 = 0.157977 \\
\alpha_0 &= P(S_1 = A) = 0.99\text{special case for t=1} \\
P(S_1 = A, \{O_t\}_{t=1}^T) &= 0.157977 \cdot 0.99 \cdot P(O_1|S_1 = A) = 0.1251178
\end{aligned}$$

Similarly, we get

$$P(S_1 = B, \{O_t\}_{t=1}^T) = 9.6923 \cdot 10^{-5}$$

For $P(S_2 = A|\{O_t\}_{t=1}^T)$:

$$\begin{aligned}
\alpha_1^A &= P(S_2 = A|S_1 = A)P(O_1|S_1 = A) = 0.99 \cdot 0.8 = 0.792 \\
\alpha_1^B &= P(S_2 = A|S_1 = B)P(O_1|S_1 = B) = 0.01 \cdot 0.1 = 0.001 \\
\alpha_0 &= 0.792 \cdot 0.99 + 0.001 \cdot 0.01 = 0.78409 \\
\beta_3^A &= P(S_3 = A|S_2 = A)P(O_3|S_3 = A) = 0.99 \cdot 0.8 = 0.792 \\
\beta_3^B &= P(S_3 = B|S_2 = A)P(O_3|S_3 = B) = 0.01 \cdot 0.1 = 0.001 \\
\beta_4 &= 0.792 + 0.001 = 0.793 \\
P(S_2 = A, \{O_t\}_{t=1}^T) &= \alpha_0 \cdot \beta_4 \cdot P(O_2|S_2 = A) = 0.1243567
\end{aligned}$$

Similarly, we get

$$P(S_2 = B, \{O_t\}_{t=1}^T) = 0.000858033$$

For $P(S_3 = A|\{O_t\}_{t=1}^T)$:

$$\begin{aligned}\alpha_2^A &= P(S_3 = A|S_2 = A)P(O_2|S_2 = A) = 0.99 \cdot 0.2 = 0.198\\ \alpha_2^B &= P(S_3 = A|S_2 = B)P(O_2|S_2 = B) = 0.01 \cdot 0.9 = 0.009\\ \alpha_1^A &= 0.198 \cdot 0.99 \cdot 0.8 + 0.009 \cdot 0.01 \cdot 0.8 = 0.156888\\ \alpha_1^B &= 0.198 \cdot 0.01 \cdot 0.1 + 0.009 \cdot 0.99 \cdot 0.1 = 0.001089\\ \alpha_0 &= 0.156888 \cdot 0.99 + 0.001089 \cdot 0.01 = 0.15533001\\ \beta_4 = 1P(S_3 = A, \{O_t\}_{t=1}^T) &= 0.15533001 \cdot 1 \cdot P(O_3|S_3 = A) = 0.124264\end{aligned}$$

Similarly, we get

$$P(S_3 = B, \{O_t\}_{t=1}^T) = 0.000950699$$

Then by comparing each two probability, the most likely setting for each state is A,A,A.

7. [**9 points**] Use the Viterbi algorithm to compute (and report) the most likely sequence of states. Show your work (i.e., show each of your Vs).

The Viterbi algorithm predicts that the most likely sequence of states is A, A, A. The relevant com putations are:

$$\begin{aligned}V_1^A &= 0.99 \cdot 0.8 = 0.792\\ V_1^B &= 0.01 \cdot 0.1 = 0.001\\ V_2^A &= 0.2 \cdot 0.792 \cdot 0.99 = 0.156816\\ V_2^B &= 0.9 \cdot 0.792 \cdot 0.01 = 0.007128\\ V_3^A &= 0.8 \cdot 0.156816 \cdot 0.99 = 0.1241983\\ V_3^B &= 0.1 \cdot 0.007128 \cdot 0.99 = 0.000705672\end{aligned}$$

8. [**2 points**] Is the most likely sequence of states the same as the sequence comprised of the most likely setting for each individual state? Provide a 1-2 sentence justification for your answer. In this case, the two are the same. This is because the probability of changing states is so low. However, in general, the two need not be the same.