

6.867: Recitation Handout (Week 11)

November 18, 2016

Contents

1	DNA	2
2	Two-state HMM	4
3	Hidden Where?	4
4	Two by Two	5
5	Missing Observations	6
6	Studiosness	7
7	IO HMM	8
8	Partially Observed Markov Chain	8

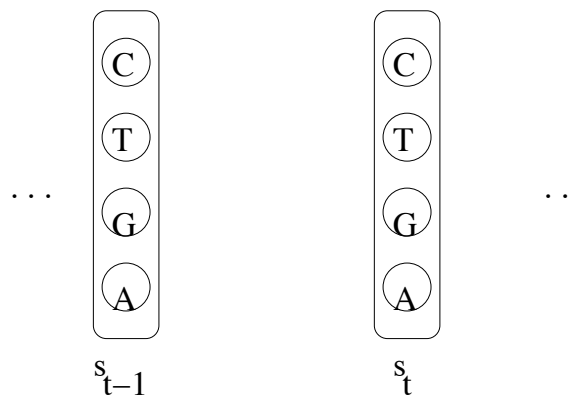
1 DNA

Consider the following pair of observed sequences:

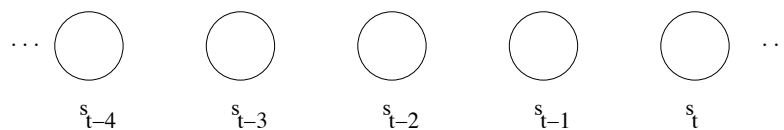
```
Sequence 1 ( $s_t$ ): AATTGGCC AATTGGCC ...
Sequence 2 ( $x_t$ ): 11221122 11221122 ...
Position  $t$ :      01234 ...
```

where we assume that the pattern (indicated with the spaces) will continue forever. Let $s_t \in \{A, G, T, C\}$, $t = 0, 1, 2, \dots$ denote the variables associated with the first sequence, and $x_t \in \{1, 2\}$, $t = 0, 1, 2, \dots$ the variables characterizing the second sequence. So, for example, given the sequences above, the observed values for these variables are $s_0 = A$, $s_1 = A$, $s_2 = T, \dots$ and, similarly, $x_0 = 1$, $x_1 = 1$, $x_2 = 2, \dots$

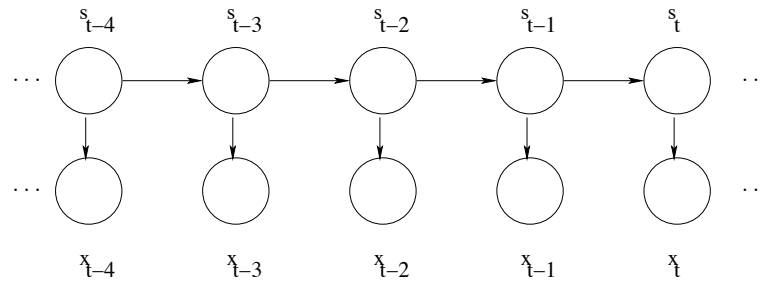
- (a) If we use a simple first-order homogeneous Markov model to predict the first sequence (values for s_t only), what is the maximum likelihood solution that we would find? In the transition diagram below, please draw the relevant transitions and the associated probabilities (this should not require much calculation)



- (b) To improve the Markov model a bit, we would like to define a graphical model that predicts the value of s_t on the basis of the previous observed values s_{t-1}, s_{t-2}, \dots (looking as far back as needed). The model parameters/structure are assumed to remain the same if we shift the model one step. In other words, it is the same graphical model that predicts s_t on the basis of s_{t-1}, s_{t-2}, \dots as the model that predicts s_{t-1} on the basis of s_{t-2}, s_{t-3}, \dots . In the graph below, draw the minimum number of arrows that are needed to predict the first observed sequence (s_t) perfectly (disregarding the first few symbols in the sequence). Since we slide the model along the sequence, you can draw the arrows only for s_t .



- (c) Now, to incorporate the second observation sequence, we will use a standard hidden Markov model:



where again $s_t \in \{A, G, T, C\}$ and $x_t \in \{1, 2\}$. We will estimate the parameters of this HMM in two different ways.

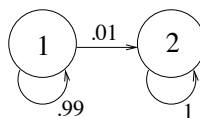
- (A) Treat the pair of observed sequences (s_t, x_t) (given above) as complete observations of the variables in the model and estimate the parameters in the maximum likelihood sense. The initial state distribution $P_0(s_0)$ is set according to the overall frequency of symbols in the first observed sequence (uniform).
- (B) Use only the second observed sequence (x_t) in estimating the parameters, again in the maximum likelihood sense. The initial state distribution is again uniform across the four symbols.

We assume that both estimation processes will be successful relative to their criteria.

- i. What are the observation probabilities $\Pr(x | s)(x \in \{1, 2\}, s \in \{A, G, T, C\})$ resulting from the first estimation approach? (should not require much calculation)
 - ii. Which estimation approach is likely to yield a more accurate model over the second observed sequence (x_t) ? Briefly explain why.
- (d) Consider now the two HMMs resulting from using each of the estimation approaches (approaches A and B above). These HMMs are estimated on the basis of the pair of observed sequences given above. We'd like to evaluate the probability that these two models assign to a new (different) observation sequence 1 2 1 2, i.e., $x_0 = 1, x_1 = 2, x_2 = 1, x_3 = 2$. For the first model, for which we have some idea about what the s_t variables will capture, we also want to know the the associated most likely hidden state sequence. (these should not require much calculation)
- i. What is the probability that the first model (approach A) assigns to this new sequence of observations?
 - ii. What is the probability that the second model (approach B) gives to the new sequence of observations?
 - iii. What is the most likely hidden state sequence in the first model (from approach A) associated with the new observed sequence?

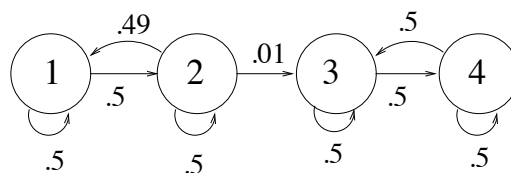
2 Two-state HMM

s	1	2
$P(x=1)$	0	0.1
$P(x=2)$	0.199	0
$P(x=3)$	0.8	0.7
$P(x=4)$	0.001	0.2



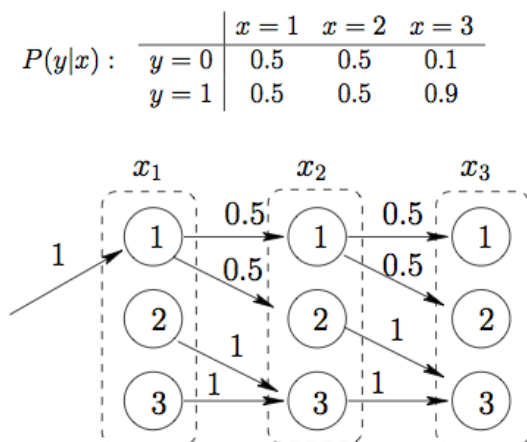
The figure above shows a two-state HMM. The transition probabilities of the Markov chain are given in the transition diagram. The output distribution corresponding to each state is defined over 1, 2, 3, 4 and is given in the table next to the diagram. The HMM is equally likely to start from either of the two states.

- Give an example of an output sequence of length 2 which can not be generated by this HMM.
- We generated a sequence of $6,867^{2016}$ observations from the HMM, and found that the last observation in the sequence was 3. What is the most likely hidden state corresponding to that last observation?
- Consider an output sequence 3 3. What is the most likely sequence of hidden states corresponding to these observations?
- Now, consider an output sequence 3 3 4. What are the *first two states* of the most likely hidden state sequence?
- We can try to increase the modeling capacity of the HMM a bit by breaking each state into two states. Following this idea, we created the diagram below. Can we set the initial state distribution and the output distributions so that this 4-state model, with the transition probabilities indicated in the diagram, would be equivalent to the original 2-state model? If yes, how? If no, why not?



3 Hidden Where?

Suppose we have a hidden Markov model with three possible states and transitions described in the figure below. The states at times $t = 1, 2, 3$ are represented by variables x_1 , x_2 , and x_3 , respectively. The distributions governing binary (0/1) outputs y_1 , y_2 , and y_3 are state dependent but do not depend on time. In other words, we only need to define $P(y|x)$ common to all time points. Specifically,



- (a) There are only three possible hidden state sequences over the three time points $t = 1, 2, 3$. What are they?
- (b) What is the most likely hidden state sequence given the observed sequence $(y_1 = 0, y_2 = 0, y_3 = 0)$? Is the answer unique?
- (c) Suppose we only receive observations for the first two time points, i.e., (y_1, y_2) , and train the HMM with the EM algorithm. The HMM is initialized with the parameters illustrated above. Select which of the following parameters could change as a result of running the EM algorithm:
- () transitions from state $x = 1$, i.e., $P_{1j} = P(x_{t+1} = j | x_t = 1), j = 1, 2, 3$
 - () transitions from state $x = 3$, i.e., $P_{3j} = P(x_{t+1} = j | x_t = 3), j = 1, 2, 3$
 - () output distribution from state $x = 1$, i.e., $P(y|x = 1), y = 0, 1$
 - () output distribution from state $x = 3$, i.e., $P(y|x = 3), y = 0, 1$

4 Two by Two

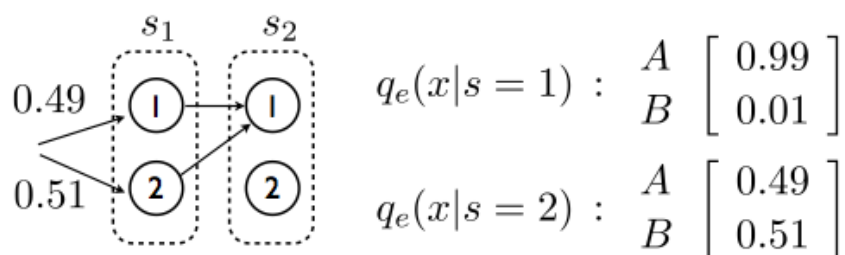


Figure 1: hidden Markov model

Consider a homogeneous hidden Markov model illustrated in Figure 2. There are two values for the hidden state, $s = 1$ and $s = 2$, and two possible output symbols, $x = A$ and $x = B$.

- What is the most frequent output symbol (A or B) to appear in the first position of sequences generated from this HMM?
- What is the sequence of three output symbols that has the highest probability of being generated from this HMM model?
- Suppose we generate output sequences of length three from this HMM and also reveal the first state s_1 . So, each example that we sample is a value assignment to variables s_1, x_1, x_2 , and x_3 . The special structure in the above HMM permits us to write the following two conditional probabilities as products of $q_e(x|s)$. Provide the expressions.

$$P(x_1, x_2, x_3 | s_1 = 1) =$$

$$P(x_1, x_2, x_3 | s_1 = 2) =$$

5 Missing Observations

Say we have a standard HMM, which generates sequences of length $m = 4$, that is the model takes the form

$$p(x_1, x_2, x_3, x_4, s_1, s_2, s_3, s_4) = t(s_1) \prod_{j=2}^4 t(s_j | s_{j-1}) \prod_{j=1}^4 e(x_j | s_j)$$

where $t(s_1)$ is the initial probability distribution over the states, $t(s_j | s_{j-1})$ are the state transition probabilities and $e(x_j | s_j)$ are the observation probabilities. Assume now that we condition on $X_2 = 1$ and $X_4 = 1$. We can then consider the conditional distribution

$$p(x_1, x_2, x_3, x_4, s_1, s_2, s_3, s_4 | X_2 = 1, X_4 = 1)$$

- Define a factor graph over the random variables $X_1, X_3, S_1, S_2, S_3, S_4$ that correctly models the conditional distribution

$$p(x_1, x_2, x_3, x_4, s_1, s_2, s_3, s_4 | X_2 = 1, X_4 = 1)$$

You should: (1) show the graph underlying the Markov random field; (2) specify any potential functions ψ on edges in the graph, as functions of the parameters $t(s_1)$, $t(s_i | s_{i-1})$ and $e(x_i | s_i)$.

- Assume that we have an HMM of length m :

$$p(x_1, x_2, \dots, x_m, s_1, s_2, \dots, s_m) = t(s_1) \prod_{j=2}^m t(s_j | s_{j-1}) \prod_{j=1}^m e(x_j | s_j)$$

Define a dynamic programming algorithm, that given an input sequence $x_1 \dots x_m$, allows us to efficiently calculate

$$\max_{s_1, s_2, \dots, s_m : s_j = s} p(x_1, x_2, \dots, x_m, s_1, s_2, \dots, s_m)$$

for any $j \in 1 \dots m$, and for any $s \in S$ where S is the set of possible states. That is, the algorithm calculates the maximum probability for any sequence of states under the constraint that s_j is equal to s .

6 Studiousness

- (a) We wish to model the grade of a student in a class that has four quizzes. For $t = 1, 2, 3, 4$, the variable S_t has value 1 if the student studied for quiz t and G_t indicates the student's cumulative grade after quiz t ; it can have values A, B, C, or F. Assume that studying for quiz t only directly affects the student's grade for that quiz. Notice that the cumulative grade through week t does depend on the grade through week $t - 1$. Draw a graphical model relating S_1, \dots, S_4 and G_0, \dots, G_4 (with G_0 denoting the grade before taking any quizzes).
- (b) For the four expressions below, draw a line connecting any pair of expressions that are equal.

$$\Pr(G_1 \mid S_4 = 1) \qquad \Pr(G_1)$$

$$\Pr(G_1 \mid S_4 = 1, G_4 = A) \qquad \Pr(G_1 \mid G_4 = A)$$

- (c) Suppose you want to make a maximum likelihood estimate of $P(G_t \mid G_{t-1}, S_{t-1})$ from data. Can you use EM to estimate it from a sequence S_1, \dots, S_k , but without any samples of G_i ? If so, briefly outline the method you would use. If not, explain why not.
- (d) Let's try to estimate the parameters of this model from a sequence G_1, \dots, G_k using EM. For simplicity, let's consider a reduced model in which there are only two grades, A and F. Our initial model θ_0 is

G _{t-1}	S _t	P(G _t = A)
A	1	.6
A	0	.4
F	1	.4
F	0	.6

P(S _t = 1)
0.7

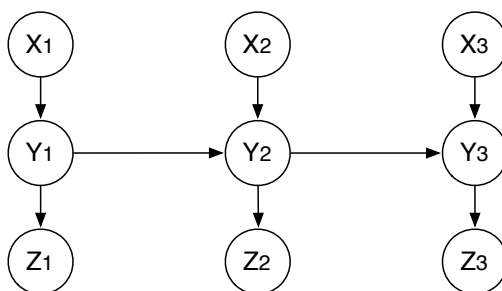
Assume our data (sequences of G 's) is (starting with G_0): A, A, F, F, F, A, A.

What is the numerical value of $P(S_1 \mid G_0, \dots, G_6, \theta_0)$?

- (e) If the values of $P(S_i = 1 \mid D, \theta_0)$ instead were .6, .2, .4, .4, .8, .6, what would the new estimated value for $P(G_t = A \mid G_{t-1} = A, S_t = 1)$ in θ_1 be? (Assume we are using maximum likelihood estimation without any Laplace-type correction).
- (f) If the values of $P(S_i = 1 \mid D, \theta_0)$ instead were .6, .2, .4, .4, .8, .6, what would the new estimated value for $P(S_t = 1)$ in θ_1 be? (Assume we are using maximum likelihood estimation without any Laplace-type correction).
- (g) Jody suggests that, given just data G_1, \dots, G_k , we might as well use a simpler model that does not have any S variables. Describe a situation, if any, in which this simpler model would be better. Describe a situation, if any, in which this simpler model would be worse.

7 IO HMM

Consider a graphical model, which is an HMM, but with “input” nodes X_1, \dots, X_t that are observable. So the model looks like this:



The X_t and the Z_t are observed, but the Y_t are hidden.

- Draw the factor graph corresponding to this model.
- Is this model equivalent to one in which the arrows from X_t to Y_t are reversed (allowing for different parameters)? Explain.

8 Partially Observed Markov Chain

Consider a Markov chain in which only the states on the even time steps are observed. The states are drawn from the set $\{A, B\}$.

- Assume we know the transition model to be:

	$\Pr(X^{t+1} = A \mid X^t)$	$\Pr(X^{t+1} = B \mid X^t)$
$X^t = A$	0.2	0.8
$X^t = B$	0.9	0.1

- What is $\Pr(X^1 \mid X^0 = A)$?
 - What is $\Pr(X^1 \mid X^0 = A, X^2 = A)$?
 - What is $\Pr(X^1 \mid X^0 = A, X^2 = A, X^4 = A)$?
- Now, assume we don't know the transition model, and want to estimate it from the following *observed* data sequence (note that this sequence is $\langle X^0, X^2, X^4, X^6 \rangle$)

$$D = \langle A, A, B, B \rangle$$

using the EM algorithm.

- What are the hidden variables?
- Write an expression for the log likelihood of the data, in terms of θ_{AA}, θ_{AB} , and the hidden variables. (For notational compactness, use θ_{AA} to mean $\Pr(X^{t+1} = A \mid X^t = A)$ and θ_{AB} to mean $\Pr(X^{t+1} = A \mid X^t = B)$ and note that $\theta_{BA} = 1 - \theta_{AA}$ and $\theta_{BB} = 1 - \theta_{AB}$.)
- (E step) Given an initial guess of the model parameters $\theta_{AA}^0, \theta_{AB}^0$, write an expression for the posterior marginal distributions over the hidden variables conditioned on $\theta_{AA}^0, \theta_{AB}^0$ and D .

- iv. (M step) Given D and marginal distributions on the hidden variables, what is the next estimate of the parameters, $\theta_{AA}^1, \theta_{AB}^1$? (You do not need to derive it if you can just write it down).