SINGAPORE UNIVERSITY OF TECHNOLOGY AND DESIGN

01.112 Machine Learning, Spring 2018
Lecture Notes for Week 9

15. Hidden Markov Models (II)

Last update: Thursday 15th March, 2018 15:19

1 Decoding

Suppose now that we have the HMM parameters (see the example in the previous lecture) and the problem is to predict the underlying tags $y_1, \ldots, y_n$ corresponding to an observed sequence of words $x_1, \ldots, x_n$. In other words, we wish to find:

$$\arg\max_{y_1,\ldots,y_n} p(y_0, y_1, \ldots, y_{n+1} | x_1, \ldots, x_n; \theta) \tag{1}$$

where we have:

$$p(x_1, \ldots, x_n, y_0, y_1, \ldots, y_{n+1}; \theta) = \prod_{i=1}^{n+1} a_{y_{i-1}, y_i} \prod_{i=1}^{n} b_{y_i}(x_i) \tag{2}$$

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

Discussion We have the joint distribution here, but we are interested in a conditional distribution above. What's the connection here?

$$\arg\max_{y_1,\ldots,y_n} p(y_0, y_1, \ldots, y_{n+1} | x_1, \ldots, x_n; \theta) = \arg\max_{y_1,\ldots,y_n} \frac{p(x_1, \ldots, x_n, y_0, y_1, \ldots, y_{n+1}; \theta)}{p(x_1, \ldots, x_n; \theta)} \tag{3}$$

Since the term $p(x_1, \ldots, x_n)$ is a constant that is independent of ys once the parameters are fixed, we can drop it when taking the $\arg\max$:

$$\arg\max_{y_1,\ldots,y_n} \frac{p(x_1, \ldots, x_n, y_0, y_1, \ldots, y_{n+1}; \theta)}{p(x_1, \ldots, x_n; \theta)} = \arg\max_{y_1,\ldots,y_n} p(x_1, \ldots, x_n, y_0, y_1, \ldots, y_{n+1}; \theta) \tag{4}$$

This leads to:

$$\arg\max_{y_1,\ldots,y_n} p(y_0, y_1, \ldots, y_{n+1} | x_1, \ldots, x_n; \theta) = \arg\max_{y_1,\ldots,y_n} p(x_1, \ldots, x_n, y_0, y_1, \ldots, y_{n+1}; \theta) \tag{5}$$

One possible solution for finding the most likely sequence of tags is to do brute force enumeration. Consider the example: {the, dog}, x = "the the the dog". The possible state sequences include:

START 1 1 1 2 STOP
START 1 1 2 2 STOP
START 1 2 2 2 STOP
⋮

But there are $|\mathcal{T}|^n$ possible sequences in total! Solving the tagging problem by enumerating the tag sequences will be prohibitively expensive.

Viterbi Algorithm

The HMM has a simple dependence structure (recall, tags form a Markov sequence, observations only depend on the underlying tag). We can exploit this structure in a dynamic programming algorithm.

Input: $\mathbf{x} = x_1, \ldots, x_n$ and model parameters θ.

Output: $\arg\max_{y_1,\ldots,y_n} p(x_1, \ldots, x_n, y_0, y_1, \ldots, y_{n+1}; \theta)$.

Now, let's look at a truncated version of the joint probability, focusing on the first k tags for any $k \in \{1, \ldots, n\}$ In other words, we define

$$r(y_1, \ldots, y_k) = \prod_{i=1}^{k} a_{y_{i-1},y_i} \prod_{i=1}^{k} b_{y_i}(x_i) \tag{6}$$

where $k \neq n+1$. Note that our notation $r(y_1, \ldots, y_k)$ suppresses any dependence on the observation sequence. This is because we view $x_1, \ldots, x_n$ as given (fixed). Note that, according to our definition,

$$p(x_1, \ldots, x_n, y_0, y_1, \ldots, y_{n+1}) = r(y_1, \ldots, y_n) \cdot a_{y_n,y_{n+1}} = r(y_1, \ldots, y_n) \cdot a_{y_n,\texttt{STOP}} \tag{7}$$

Let $S(k, v)$ be the set of tag sequences $y_1, \ldots, y_k$ such that $y_k = v$. In other words, $S(k, v)$ is a set of all sequences of length k whose last tag is v. The dynamic programming algorithm will calculate

$$\pi(k, v) = \max_{(y_1,\ldots,y_k) \in S(k,v)} r(y_1, \ldots, y_k) \tag{8}$$

recursively in the forward direction.

In other words, $\pi(k, v)$ can be thought as solving the maximization problem partially, over all the tags $y_1, \ldots, y_{k-1}$ with the constraint that we use tag v for y_k. If we have $\pi(k, v)$, then $\max_v \pi(k, v)$ evaluates $\max_{y_1,\ldots,y_k} r(y_1, \ldots, y_k)$. We leave v in the definition of $\pi(k, v)$ so that we can extend the maximization one step further as we unravel the model in the forward direction. More formally,

- Base case:

$$\pi(0, v) = \begin{cases} 1 & \text{if } v = \texttt{START} \text{ (starting state, no observations)} \\ 0 & \text{otherwise} \end{cases} \tag{9}$$

- Moving forward recursively: for any $k \in \{1, \ldots, n\}$

$$\pi(k, v) = \max_{u \in \mathcal{T}} \{\pi(k-1, u) \cdot a_{u,v} \cdot b_v(x_k)\} \tag{10}$$

  In other words, when extending $\pi(k-1, u), u \in \mathcal{T}$, to $\pi(k, v), v \in \mathcal{T}$, we must transition from $y_{k-1} = u$ to $y_k = v$ (part $a_{u,v}$) and generate the corresponding observation x_k (part $b_v(x_k)$). Then we maximize over the previous tag $y_{k-1} = u$ so that $\pi(k, v)$ only depends on the value of y_k.

- Finally, we must transition from y_n to `STOP` so that

$$\max_{y_1, \ldots, y_n} p(x_1, \ldots, x_n, y_0 = \texttt{START}, y_1, \ldots, y_n, y_{n+1} = \texttt{STOP}) = \max_{v \in \mathcal{T}} \{\pi(n, v) \cdot a_{v,\texttt{STOP}}\} \tag{11}$$

  The whole calculation can be done in time $O(n|\mathcal{T}|^2)$, linear in length, quadratic in the number of tags.

Now, having values $\pi(k, v)$, how do we reconstruct the most likely sequence of tags which we denote as $y_1^*, \ldots, y_n^*$? We can do this via *backtracking*. In other words, at the last step, $\pi(n, v)$ represents maximizations of all $y_1, \ldots, y_n$ such that $y_n = v$. What is the best value for this last tag v, *i.e.*, what is y_n^*? It is:

$$y_n^* = \arg\max_v \{\pi(n, v) \cdot a_{v,\texttt{STOP}}\} \tag{12}$$

Now we can fix y_n^* and work backwards. What is the best value y_{n-1}^* such that we end up with tag y_n^* in position n? It is simply

$$y_{n-1}^* = \arg\max_u \{\pi(n-1, u) \cdot a_{u,y_n^*}\} \tag{13}$$

and so on.

Discussion What is the space complexity of the above algorithm? Is it possible to store the optimal transition information together with the $\pi(k, v)$? If we do so, what is the space complexity? Do we still need to do backtracking in this case? yes ?

Exercise What if we would like to find the top-k ($k > 1$) most optimal sequences instead of finding the single most optimal tag sequence?

Learning Objectives

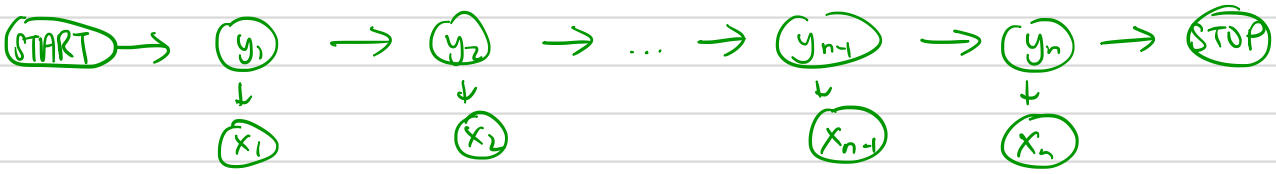
You need to know:

1. What is decoding for an HMM
2. How to perform decoding for an HMM using the Viterbi algorithm
3. What is the guarantee of the Viterbi algorithm and how to implement the Viterbi algorithm
4. What is the space and time complexity of the Viterbi algorithm

W802 Recap - Hidden Markov Models

X : Faith is a fine internship
Y : N V D A N

In green in W802 Notes (Recap)



$$P(x_1, \dots, x_n, y_0, \dots, y_n, y_{n+1}) = \prod_{i=1}^{n+1} P(y_i \mid y_{i-1}) \prod_{i=1}^{n} P(x_i \mid y_i)$$

JOINT PROBABILITY

transition param | emission param

$$= \prod_{i=1}^{n+1} a_{y_{i-1}, y_i} \prod_{i=1}^{n} b_{y_i}(x_i)$$

Supervised Learning

In training, given X and Y → find $\theta = \{a, b\}$

Steps : Find joint probability
Add log-likelihood
Take Partial Derivative

Max Likelihood Estimate (similar to Naive Bayes) : $a_{u,v} = \frac{\text{count}(u \to v)}{\text{count}(u)}$

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

Components needed for HMM : states, observations

W901 Eg

- States : { START, A, B, STOP }
- Observations : { "the", "dog" }
- Transition params

u\v	A	B	STOP
START	1.0	0.0	0.0
A	0.5	0.5	0.0
B	0.0	0.8	0.2

u\o	"the"	"dog"	
A	0.9	0.1	→ = 1
B	0.1	0.9	

X : the dog the
Y : A B A

$$P(X,Y) = P(A \mid START) \cdot P(B \mid A) \cdot P(A \mid B) \cdot P(STOP \mid A)$$
$$= a_{START,A} \cdot b_A(\text{"the"}) \cdot a_{A,B} \cdot b_B(\text{"dog"}) \cdot a_{B,A} \cdot b_A(\text{"the"}) \cdot a_{A,STOP}$$
$$= (1.0)(0.9)(0.5)(0.9)(0.0)(0.9)(0.0) = 0$$

X_1 the | X_2 dog | X_3 the

equivalent to P(X|Y)

In testing / Decoding : Given X, $\theta = \{a, b\}$, Find : Y where $Y^* = \underset{Y}{\text{argmax}}\, P(Y \mid X)$

→ ie max P(Y|X) and returns y by trying all Y

↗ conditional probability

drop P(X) w/g concerted effort
↗ as P(X) is a constant (ie common factor)

$$Y^* = \arg\max_{Y} P(Y|X) = \arg\max_{Y} \frac{P(X,Y)}{P(X)} = \arg\max_{Y} P(X,Y)$$

↳ Find y st joint probability is maximised.

TRY ALL DIFF y paths :

AAA	BAA
AAB	BAB
ABA	BBA
ABB	BBB

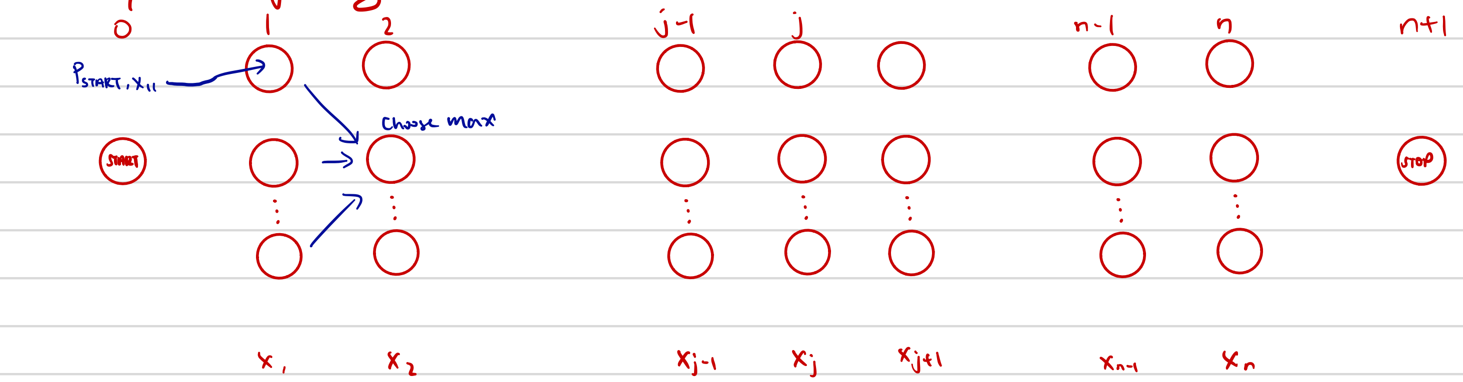
→ T^n (time complexity)

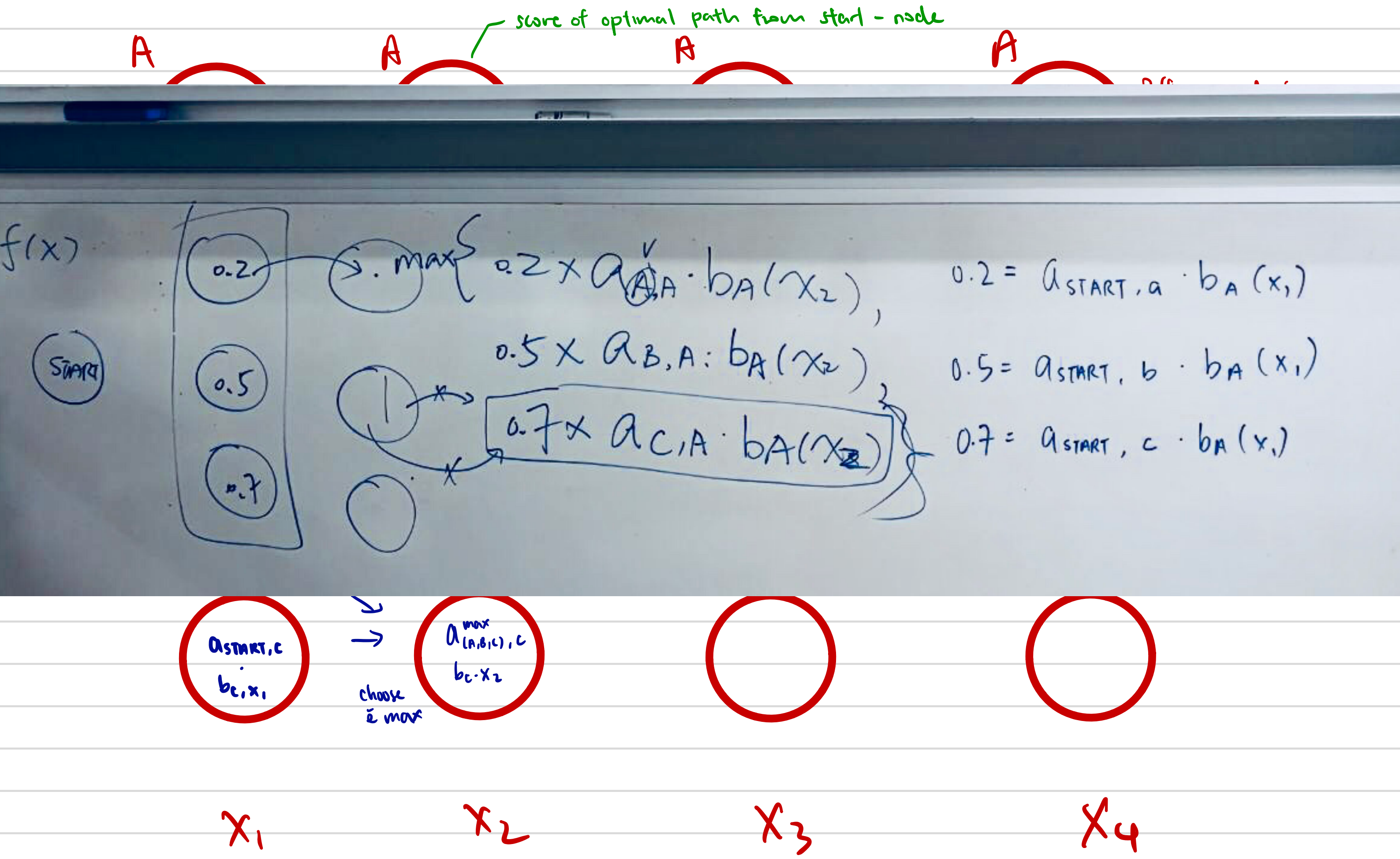
n : no of words (the dog the)

T : no of states. (a, b)

$= 2^3$ (too long) – exponential

Soln : Dynamic Programming (reduce prob by n-1 terms)





T nodes in j layer

1 node has to visit T no of operations in j-1 layer

ea position / j layer has $T \times T = T^2$ no of operations

$\pi(j,u)$: é score of é optimal partial path from START to é node (j,u)

VITERBI ALGO

① $\pi(0, \text{START}) = 1$

$\pi(1, u) = a_{\text{START},u} \cdot b_u(x_1)$

② Move forward recursively : $\pi(j,u) = \max_v \pi(j-1, v) \cdot a_{v,u} \cdot b_u(x_j)$ → for $j = 2, \dots n$ (layer 2 to n), for all u.

③ $\pi(n+1, \text{STOP}) = \max_v \pi(n,v) \cdot a_{v,\text{STOP}}$

no of operations for each node = T

n : no of words (pay layer) → no of nodes = nT

T : no of states / nodes

→ Total no operations = $nT^2 + T + 1 \approx nT^2$ (↗ stop, ↗ start)

↓ quadratic algo

, Space Complexity for whole Algo : $n(0) = O(2nT)$

There is a back-tracking path. (top k-path)

↓ know score → but want to know path.