Barry Tee Wei Cong 1001549 01.112 Machine Learning

HW4.

1. Parameter 1: Transition Probabilities, $a_{u,v} = \frac{\text{count}(u;v)}{\text{count}(u)}$

u\v	X	Y	Z	STOP
START	2/4	0/4	2/4	0/4
X	0/5	2/5	2/5	1/5
Y	1/5	0/5	1/5	3/5
Z	2/5	3/5	0/5	0/5

Parameter 2: Emission Probabilities, $b_u(o) = \frac{\text{count}(u \to o)}{\text{count}(u)}$

u\o	a	b	c
X	1/5	3/5	1/5
Y	2/5	0/5	3/5
Z	1/5	2/5	2/5

2. · Base Case

$\pi(0, START) = 1$

· Recursive Case

$\pi(1,X) = \pi(0,START) \cdot a_{START,X} \cdot b_X(a) = 1 \cdot \frac{2}{4} \cdot \frac{1}{5} = \frac{1}{10}$

$\pi(1,Y) = \pi(0,START) \cdot a_{START,Y} \cdot b_Y(a) = 1 \cdot \frac{0}{4} \cdot \frac{2}{5} = 0$

$\pi(1,Z) = \pi(0,START) \cdot a_{START,Z} \cdot b_Z(a) = 1 \cdot \frac{2}{4} \cdot \frac{1}{5} = \frac{1}{10}$

$\pi(2,X) = \max_{u \in T} \{\pi(1,u) \cdot a_{u,X} \cdot b_X(a)\}$

$= \max\{\frac{1}{10} \cdot \frac{0}{5} \cdot \frac{1}{5}, \ 0 \cdot \frac{1}{5} \cdot \frac{1}{5}, \ \frac{1}{10} \cdot \frac{2}{5} \cdot \frac{1}{5}\} = \frac{2}{250}$

$\pi(2,Y) = \max_{u \in T} \{\pi(1,u) \cdot a_{u,Y} \cdot b_Y(a)\}$

$= \max\{\frac{1}{10} \cdot \frac{2}{5} \cdot \frac{2}{5}, \ 0 \cdot \frac{0}{5} \cdot \frac{2}{5}, \ \underline{\frac{1}{10} \cdot \frac{3}{5} \cdot \frac{2}{5}}^{Z}\} = \frac{6}{250} = \frac{3}{125}$

$\pi(2,Z) = \max_{u \in T} \{\pi(1,u) \cdot a_{u,Z} \cdot b_Z(a)\}$

$= \max\{\frac{1}{10} \cdot \frac{2}{5} \cdot \frac{1}{5}, \ 0 \cdot \frac{1}{5} \cdot \frac{1}{5}, \ \frac{1}{10} \cdot \frac{0}{5} \cdot \frac{1}{5}\} = \frac{2}{250}$

$\pi(3,STOP) = \max_{v \in T} \{\pi(2,v) \cdot a_{v,STOP}\}$

$= \max\{\frac{2}{250} \cdot \frac{1}{5}, \ \underline{\frac{3}{125} \cdot \frac{3}{5}}^{Y}, \ \frac{2}{250} \cdot \frac{0}{5}\}$

$= \frac{9}{625}$

Backtracking: $y_2^* = Y$, $y_1^* = Z$ → optimal sequence is Z, Y.

Scanned by CamScanner

Date No.

3. $y_j \neq u$

Modified Viterbi Algorithm

1. $\pi(0, START) = 1$
2. If $j \neq i$ or $i+1$,

$\pi(j, u) = \max_v \pi(j-1, v) \cdot a_{v,u} \cdot b_u(x_j)$

→ move forward recursively where $j = 2, \ldots, n$ for all u.

else

$\pi(i, u) = \max_u \pi(i-1, u) \cdot a_{u,u} \cdot b_u(x_i)$

→ move forward recursively where $i = 2, \ldots, n$ for all u except for u = verb

$\pi(i+1, u) = \max_u \pi(i, u) \cdot a_{u,u} \cdot b_u(x_{i+1})$

→ move forward recursively where $i = 2, \ldots, n$ for all u except for u = verb

3. $\pi(n+1, STOP) = \max_u \pi(n, u) \cdot a_{u,STOP}$

4. Backtracking

$y_n^* = \arg\max_u \{ \pi(n, u) \cdot a_{u,STOP} \}$

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

$\vdots$

$y_i^* = \arg\max_u \{ \pi(i, u) \cdot a_{u,v} \}$ for all u except for u = verb

Scanned by CamScanner

Date No.

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

$P(x_1, \ldots, x_{i-1}, y_i, \ldots, y_{i-1}, z_i = u, x_i, \ldots, x_n, y_i, \ldots, y_n ; \theta)$

$= P(x_1, \ldots, x_{i-1}, y_i, \ldots, y_{i-1}, z_i = u ; \theta) \cdot P(x_i, \ldots, x_n, y_i, \ldots, y_n \mid z_i = u ; \theta)$

$= \alpha_u(i) \cdot \beta_u(i)$

Forward Algorithm

- Base Case

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

- Recursive Case for $i = 2, \ldots, n$

$\alpha_u(i+1) = \sum_v \alpha_v(i) \cdot a_{v,u} \cdot b_v(x_i) \times c_v(y_i)$ where $c_v(y) = P(y|v)$

Backward Algorithm

- Base Case

$\beta_u(n) = a_{u,STOP} \cdot b_u(x_n) \times c_u(y_n)$ $\forall u \in \{1, \ldots, N-1\}$

- Recursive Case for $n-1, \ldots, 1$

$\beta_u(n) = \sum_v \beta_v(i+1) \cdot b_u(x_i) \cdot a_{u,v} \times c_u(y_i)$

The length of the sentence is n.

For each position / layer, there are T forward & T backward terms.

Total : $O(n) = O(nT^2)$

01.112 Machine Learning, Spring 2018
Homework 4

Due 4 April 2018, 5pm

This homework will be graded by Allan Jie

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.

Question 1. Assume that we have the following training data available for us to estimate the model parameters:

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

Clearly state what are the parameters associated with the HMM. Under the maximum likelihood estimation (MLE), what would be the values for the optimal model parameters? Clearly show how each parameter is estimated exactly. *(10 points)*

Question 2. 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. *(10 points)*

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

Question 3. The Viterbi algorithm discussed in class is used for finding the optimal $\mathbf{y}$ sequence based on the following:

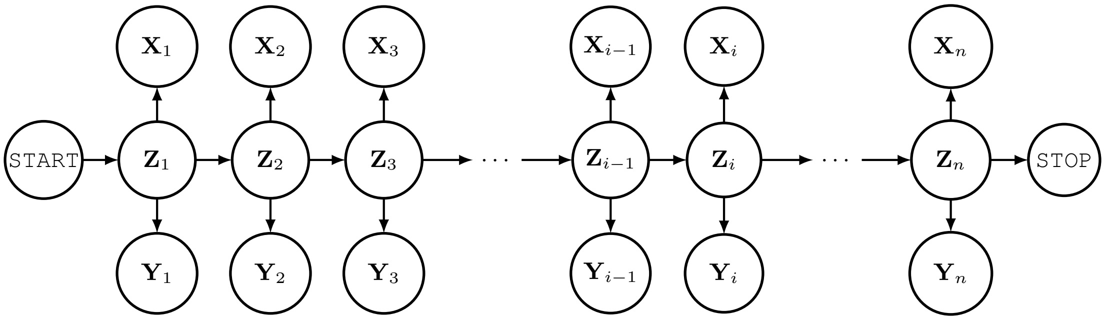
$$y_1^*, \ldots, y_n^* = \arg\max_{y_1, \cdots, y_n} p(x_1, \ldots, x_n, y_1, \ldots, y_n)$$

Now, consider the problem of part-of-speech tagging. Sometimes we have some prior knowledge about certain tags for certain observations. For example, assume the observation x_i="*the*", we are almost certain that it is not a verb (*i.e.*, we believe $y_i \neq \mathbb{V}$). In this case, we would like to do the decoding in the following way, where we would like to incorporate the prior knowledge $y_i \neq \mathbb{V}$ (and find optimal values for all other y_k in the sequence, where $k = 1, \ldots, n, k \neq i$):

$$y_1^*, \ldots, y_{i-1}^*, y_{i+1}^*, \ldots, y_n^* = \arg\max_{y_1, \cdots, y_{i-1}, y_{i+1}, \cdots, y_n} p(x_1, \ldots, x_n, y_1, \ldots, y_{i-1}, y_{i+1} \ldots, y_n | y_i \neq \mathbb{V})$$

Clearly describe how to modify the Viterbi algorithm to perform such a new decoding task. *(10 points)*

Question 4. 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 $\mathbf{X}$ sequence is generated from the $\mathbf{Z}$ sequence and $\mathbf{Y}$ sequence is also generated from the $\mathbf{Z}$ 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 $0 = \textbf{START}$ and $N = \textbf{STOP}$. 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). *(10 points)*