1 Questions

1. Suppose:

the possible states are \{one, two, three, four, five\}
the possible observation symbols are \{a, x, z, y, b\}

let the parameters of particular HMM with these states and observation symbols be defined by \(A, B\) and \(P_I\) as follows:

\(A\) \((A[i,j] \text{ is prob of } j \text{ given } i)\)

<table>
<thead>
<tr>
<th></th>
<th>one</th>
<th>two</th>
<th>three</th>
<th>four</th>
<th>five</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>0.0</td>
<td>0.0</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>two</td>
<td>0.0</td>
<td>0.0</td>
<td>0.00</td>
<td>1.00</td>
<td>0.00</td>
</tr>
<tr>
<td>three</td>
<td>0.0</td>
<td>0.25</td>
<td>0.25</td>
<td>0.25</td>
<td>0.5</td>
</tr>
<tr>
<td>four</td>
<td>0.5</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.5</td>
</tr>
<tr>
<td>five</td>
<td>0.0</td>
<td>0.00</td>
<td>0.00</td>
<td>0.00</td>
<td>1.0</td>
</tr>
</tbody>
</table>

\(B\) \((B[i,k] \text{ prob of } k \text{ given } i)\)

<table>
<thead>
<tr>
<th></th>
<th>a</th>
<th>x</th>
<th>z</th>
<th>y</th>
<th>b</th>
</tr>
</thead>
<tbody>
<tr>
<td>one</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>two</td>
<td>0.0</td>
<td>1.0</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>three</td>
<td>0.0</td>
<td>0.5</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>four</td>
<td>0.5</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>five</td>
<td>0.0</td>
<td>0.5</td>
<td>0.0</td>
<td>0.0</td>
<td>1.0</td>
</tr>
</tbody>
</table>

\(P_I\) \((P_I[i] \text{ prob of starting in } i)\)

<table>
<thead>
<tr>
<th></th>
<th>one</th>
<th>two</th>
<th>three</th>
<th>four</th>
<th>five</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.6666667</td>
<td>0.3333333</td>
<td>0.0000000</td>
<td>0.0000000</td>
<td>0.0000000</td>
</tr>
</tbody>
</table>

Let \((o^{0}, s^{0})\) be the following pairing of observed and state sequence

one/a three/x four/z five/.
(a) Give the formula for its probability in terms of entries from $A$, $B$ and $P_I$:
ANSWER see 1
(b) Thereby confirm that its numerical value is 0.02083333
ANSWER: see 2
(c) Considering the observation probabilities ($B$), the states which generate the symbols of $a \times z$, with non-zero probability are:
- $a$ one
- $x$ three or four
- $z$ four
- $s$ five
So considering just the $B$ probs the following is an alternative possible pairing of states with the observations
one/a four/x four/z five/.
Explain what aspects of the transition probabilities ($A$) rule out this pairing
ANSWER: see 3
(d) Explain why therefore the conditional probability of $s^0$ given $o^0$, $P(s^0|o^0)$, is 1.
ANSWER: see 4

2. (a) the so-called ‘forward probability’ $\alpha_t(i)$ represents a particular probability involving the state being $i$ at time $t$. Say what that probability is
ANSWER: see 5
(b) If the length of the entire observation sequence is $T$, and the forward algorithm has already calculated $\alpha_T(i)$, for all $i$, explain how to derive $P(o_1:T)$ from this – that is, the probability of the entire observation sequence.
ANSWER: see 6
(c) Whereas the forward probability only concerns the state at $t$ and the observations up to and including $t$, the product $\alpha_t(i)\beta_t(i)$ can be shown to give a probability involving also the remaining observations, $o_{t+1}\ldots o_T$, namely
$$\alpha_t(i)\beta_t(i) = P(o_1\ldots o_t, s_t = i, o_{t+1}\ldots o_T)$$
If for some arbitrary $t$, $\alpha_t(i)\beta_t(i)$ were known for all $i$, explain how one could derive from this $P(o_1\ldots o_T)$
ANSWER: see 7
(d) In the efficient Baum-Welch training algorithm, for each $t$ and $i$ it is necessary to calculate the quantity
\[ \gamma_t(i) = P(s_t = i | o_{1:T}) \]

that is, the conditional probability of state \( i \) at \( t \) given the entire observation sequence \( o_{1:T} \). Give a formula using \( \alpha \) and \( \beta \) which will give this conditional probability.

ANSWER: see 8

(e) The crucial property of \( \alpha_t(i) \) and \( \beta_t(i) \) is that when multiplied they give \( P(o_1 \ldots o_t, s_t = i, o_{t+1} \ldots o_T) \).

\( \beta_t(i) \) is defined to be \( P(o_{t+1} \ldots o_T | s_t = i) \). Due to independence assumptions of HMMs, \( \beta_t(i) \) is also equal to

\[ P(o_{t+1} \ldots o_T | o_1 \ldots o_t, s_t = i) \]  (1)

which adds \( o_1 \ldots o_t \) as a conditioning factor\(^1\).

Use (1) to show the crucial \( \alpha_t(i)\beta_t(i) = P(o_1 \ldots o_t, s_t = i, o_{t+1} \ldots o_T) \) property

ANSWER see 9

3. (a) Give a pseudo-code outline of the algorithm which calculates a table of \( \alpha_t(i) \) values, that is, an outline of the forward algorithm

ANSWER: see 10

(b) '\( P(o_{1:t}) \) stands for a summation of \( N^t \) terms'. Explain this statement.

ANSWER: see 11

4. Outline the roles of the Baum-Welch and Forward algorithms in the creation and use of an isolated word recognition system.

ANSWER: see 12

\(^1\)In terms of the notion of 'conditional independence', you would say \( o_{t+1} \ldots o_T \) are conditionally independent of \( o_1 \ldots o_t \) given \( s_t = i \)
2 Answers

1. answer to 1a

\[ PI[one]*B[one,a]*A[one,three]*B[three,x]*A[three,four]*B[four,z]*A[four,five]*B[five,.] \]

2. answer to 1b

\[ 0.6666667 \times 1 \times 1 \times 0.5 \times 0.25 \times 0.5 \times 0.5 \times 1 = 0.02083333 \]

3. answer to 1c. The \( A \) probabilities rule out the transition from one to four.

4. answer to 1d. \( s^0 \) (ie. one three four five) is the only state sequence whose joint probability with \( o^0 \) (ie. a x z .) is non-zero.

By definition the conditional probability of \( s^0 \) given \( o^0 \), \( P(s^0|o^0) \), is

\[ \frac{P(s^0, o^0)}{P(o^0)} \]

The denominator \( P(o^0) \) is by definition \( \sum_s P(s, o^0) \), but in view of the preceding parts only \( s^0 \) contributes anything to this sum, that is \( \sum_s P(s, o^0) = P(s^0, o^0) \). Hence the conditional probability is 1.

5. answer to 2a

By definition, \( \alpha_t(i) = P(o_{1:t}, s_t = i) \). This sums over all sequences which could generate the observations \( o_{1:t} \), with the proviso that the sequence must have \( i \) at \( t \).

6. answer to 2b

By definition, \( \alpha_T(i) = P(o_{1:T}, s_t = i) \), so summing over all possible states, \( \sum_i \alpha_T(i) \) will give \( P(o_{1:T}) \).

7. answer to 2c

\( \alpha_t(i)\beta_t(i) \) represents \( P(o_{1:T}, s_t = i) \), so summing over all states, \( \sum_i \alpha_t(i)\beta_t(i) \) will give \( P(o_{1:T}) \).

Although it would be possible to get \( P(o_{1:T}) \) this way, note from 6 that it can be got by summing just \( \alpha \) at \( T \).

8. answer to 2d

Applying the definition of cond. prob.

\[ P(s_t = i|o) = \frac{P(o, s_t = i)}{P(o)} \]

As noted in Answer 7, the numerator of this \( \alpha_t(i)\beta_t(i) \). As noted in Answer 6, the denominator is \( \sum_{i'} \alpha_T(i') \).
So the formula is

\[ \gamma_t(i) = \frac{\alpha_t(i)\beta_t(i)}{\sum_{i'}\alpha_T(i')} \]

9. answer to 2e

If \( \beta_t(i) = P(o_{t+1} \ldots o_T|o_1 \ldots o_t, s_t = i) \), then applying def cond. prob,

\[ \beta_t(i) = \frac{P(o_1 \ldots o_t, o_{t+1} \ldots o_T, s_t = i)}{P(o_1 \ldots o_t, s_t = i)} \]

bearing in mind that \( \alpha_t(i) = P(o_1 \ldots o_t, s_t = i) \) multiplying these two together must give \( P(o_1 \ldots o_t, o_{t+1} \ldots o_T, s_t = i) \)

10. answer to 3a

// basically for j at t, for each previous i sum together // alpha(t-1,i) times 'next step', ie. A[i,j] * B[j,o_t]

//initialize
for (i = 1 to N) {
    alpha[1,i] = pi[i] * B[i,o_1];
}

//iteration
for (t = 2 to T) {
    for(j = 1 to N) { // each state at present time
        total = 0;
        for(i = 1 to N) { // each state at previous time
            total += alpha[t-1,i] * A[i,j] * B[j,o_t];
        }
        alpha[t,j] = total;
    }
}

11. answer to 3b \( P(o_{1:t}) \) represents the sum of many \( P(s_{1:t}, o_{1:t}) \) terms, in each of which \( s_{1:t} \) is a length \( t \) sequence of states. At each position of \( s_{1:t} \) one of \( N \) states is possible, so there are \( N^t \) different values of \( s \) to consider.

Possibly this way of writing it makes it clearer:

\[ P(o_{1:t}) = \sum_{s_1=1}^{N} \ldots \sum_{s_t=1}^{N} P(o_1, s_1, \ldots, o_t, s_t) \]
12. answer to 4

Suppose the words to be recognised are spoken versions of the digits 1, 2 and 3.

Three HMMs would be created, $M_1$, $M_2$ and $M_3$, each with its own parameters.

These can be trained if for each digit there is a set of recordings of that digit being spoken.

If $O_1$ is the set of observation sequences for the digit 1, the **Baum-Welch** algorithm would be used to train the parameters of $M_1$, setting its parameters to maximise the probability of these sequences. Similarly training sets $O_2$ and $O_3$ would be used to train $M_2$ and $M_3$. This is shown in the following picture:

Once trained, then to do recognition on a recording $o_{test}$, for each model $M_i$ the **Forward Algorithm** would be used to find its value of $P(o; M_i)$ – the probability of those observations according to that model. The HMM which which gives maximum value to this defines the category ie. if it is the model of digit 1, $o_{test}$ is categorised as 1, etc. This is shown in the following picture: