Daiki Shirafuji
daikish@kth.se
Jiameng Bian
jiamengb@kth.se

DD2380 ARTIFICIAL INTELLIGENCE

Hidden Markov Models

Question 1: This problem can be formulated in matrix form. Please specify the initial probability vector π, the transition probability matrix A and the observation probability matrix B.

We want to formulate $P(O|\lambda)$, where λ is (A, B, π) ($t = [1, T]$). Now, **O is** $\{\mathbf{o}_1, \mathbf{o}_2, ..., \mathbf{o}_K\}$, K is the number of Possible Observations, **X** is $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$, and N is the number of Possible Hidden States.

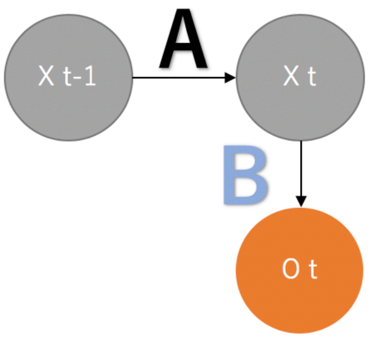

Figure 1: An Example of states

As a preparation, let's prove $P(O, X \mid \lambda) = P(O \mid X, \lambda) * P(X|\lambda)$. From

$$P(\boldsymbol{O}, \boldsymbol{X} \mid \lambda) = \frac{P(\boldsymbol{O} \cap \boldsymbol{X} \cap \lambda)}{P(\lambda)}$$

and

$$\frac{P(\boldsymbol{O} \cap \boldsymbol{X} \cap \lambda)}{P(\lambda)} = \frac{P(\boldsymbol{O} \cap \boldsymbol{X} \cap \lambda)}{P(\boldsymbol{X} \cap \lambda)} * \frac{P(\boldsymbol{X} \cap \lambda)}{P(\lambda)} = P(\boldsymbol{O} \mid \boldsymbol{X}, \lambda) * P(\boldsymbol{X} \mid \lambda)$$

we proved $P(O, X \mid \lambda) = P(O \mid X, \lambda) * P(X|\lambda)$.

Now, we define the forward path as $\alpha_t(j)= P(\mathbf{o_1},\mathbf{o_2},\ldots,\mathbf{o_t}, x_t= q_j \mid \lambda)$, where q_j is the state at t. "The t state in the sequence of states is state q_j" is be represented by $x_t=q_j$.

So, when t = 1,

$$\alpha_1(j) = P(\boldsymbol{o}_1, x_1 = qj \mid \lambda)$$
$$= P(\boldsymbol{O} \mid x_1 = qj, \lambda) * P(x_1 = qj \mid \lambda) = \pi_j * B_j(\boldsymbol{o_0})$$

and when $1 < t \leq T$,

$$\alpha_t(j) = \mathrm{P}(\boldsymbol{o}_1, \boldsymbol{o}_2, \ldots, \boldsymbol{o}_t, x_t = \mathrm{q_j} \mid \lambda)$$
$$= P(\boldsymbol{O} \mid \boldsymbol{X}, \lambda) * P(\boldsymbol{X} \mid \lambda) = \sum_{i=1}^{N} \alpha_{t-1} * A_{i,j} * B_j(\boldsymbol{O}t).$$

The figure, which describes α concretely, is shown in the following Figure 2.

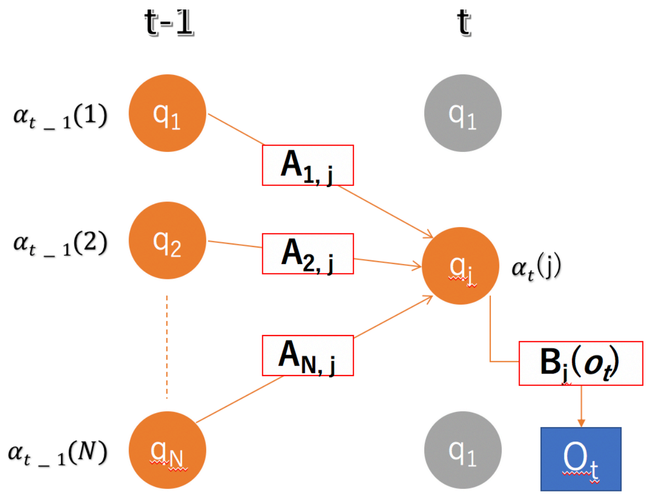

Figure 2: α explanation

From the α definition,

$$P(\boldsymbol{O}|\lambda) = \sum_{i=1}^{\mathrm{N}} \alpha_T(i).$$

So, we can formulate P(O | λ) with A, B, and π.

<u>Question 2 & 3: What is the result of this operation?</u>

Figure 3 shows Pi, A and B in the given problem for this Question 2 and 3.

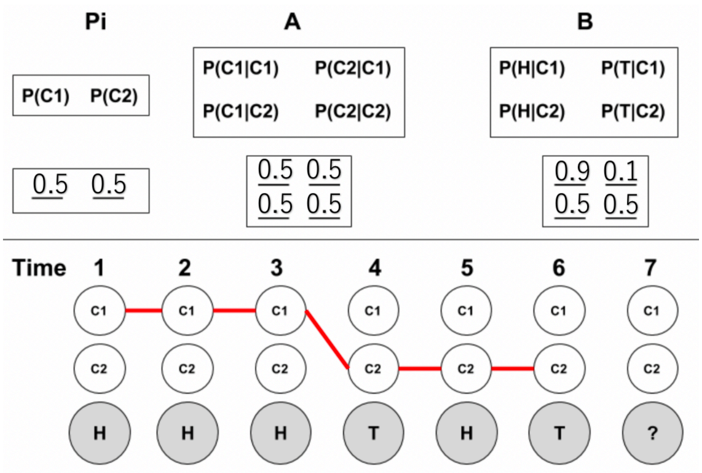

Figure 3: Pi, A and B in the given problem

When Xt-1=C1, P(Xt=C1)=0.5, and when Xt-1=C2, P(Xt=C2)=0.5. Also, when Xt-1=C1, P(Ot=C1)=0.9, and when Xt-1=C2, P(Ot=C2)=0.5. Therefore, we can get the matrix shown in the Figure 3. So, the answer of Question 2 (Pi*A) is following:

$$
\begin{aligned}
&Pi * A \\
&= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix} \\
&= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}.
\end{aligned}
$$

And the answer of Question 3 (Pi*A*B) is following:

$$
\begin{aligned}
&(Pi * A) * B \\
&= \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix} \\
&= \begin{bmatrix} 0.7 & 0.3 \end{bmatrix}.
\end{aligned}
$$

Therefore, O7 is Head with a probability of 0.7, and Tail with a probability of 0.3.

Question 4: Why is it valid to substitute O1:t=o1:t with Ot=ot when we condition on the state Xt =xi?

When Xt=xi is given, we can ignore **O**1:t-1 because there is no arrow from **O**1:t-1 to **O**t since Figure 4 shows. Therefore, we can substitute them with **O**t=**o**t.

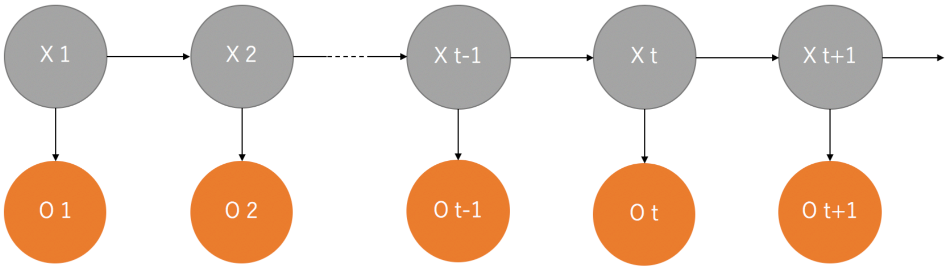

Figure 4: Hidden states and observations

Question 5: How many values are stored in the matrices δ and δi d x respectively?

δ is represented with:

$$\delta_1(i) = \pi_i * B_i(\boldsymbol{O}_1) \quad when\ t = 1, and$$

$$\delta_t(i) = \max_{j \subset \{1,\dots,N\}} \delta_{t-1}(j) * A_{j,i} * B_i(\boldsymbol{O}_t) \qquad when\ t = 2, \dots, T$$

δ_idx is represented with:

$$\delta_t^{idx}(i) = log(\pi_i * B_i(\boldsymbol{O}_1)) \quad when\ t = 1, and$$

$$\delta_t^{idx}(i) = \max_{j \subset \{1,\dots,N\}} \{\delta_{t-1}^{idx}(j) + log(A_{j,i}) + \log(Bi(\boldsymbol{O}_t))\} \quad when\ t = 2, \dots, T.$$

Therefore, there are N*T values in δ, where N is the number of states in the model and T is the length of the observation sequence. And, there are N*T values in δ_idx.

Question 6: Why we do we need to divide by the sum over the final α values for the di-gamma function?

The reason is for scaling α. Before proving this reason, we show an intuitive explanation.

First of all, γt (i, j) is the probability in a state qi at t and a state qj at time t+1. And α definition is following:

$$\alpha_t(j) = P(\boldsymbol{o_1}, \boldsymbol{o_2}, \dots, \boldsymbol{o_t}, x_t = q_j \mid \lambda)$$

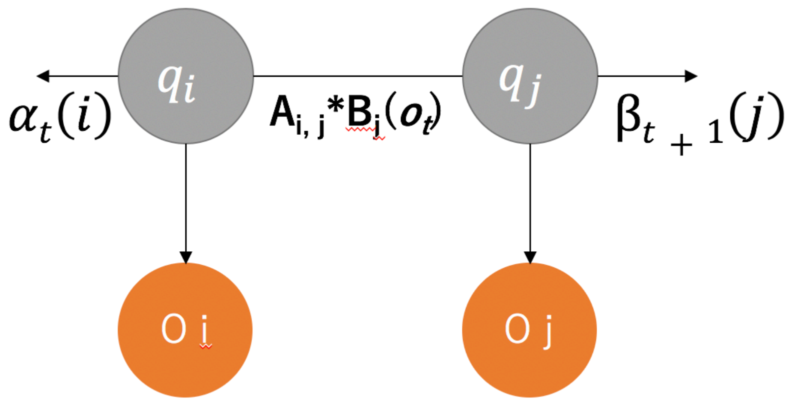

Figure 5: α and β in the model of (A, B, and Pi)

αt(j) (j=[1, N]) will become lower number (approach to 0) as t increases. This is because the number of condition (**o**t) increases. Therefore, the result of γt (i, j) become lower as t increases. For solving this problem, α should be scaled. Therefore, we divide α with the sum over the final α.

Now, we start to prove the need to divide by the sum over the final α values for the di-gamma function.

Before calculating γt (i, j), we have to calculate:

$$P(A|B,C) = \frac{P(A,B|C)}{P(B|C)}$$

and

$$P(x_t = q_i, x_{t+1} = q_j, \boldsymbol{O}|\lambda) = \alpha_t(i)A_{i,j}B_j(\boldsymbol{o}_t)\beta_{t+1}(j).$$

From the definition of γt (i, j),

$$\gamma_t(i,j) = P(x_t = q_i, x_{t+1} = q_j|\boldsymbol{O}, \lambda)$$

Then,

$$\gamma_t(i,j) = \frac{P(x_t = q_i, x_{t+1} = q_j, O|\lambda)}{P(O|\lambda)}$$

$$= \frac{\alpha_t(i)A_{i,jBj}(\boldsymbol{o}_t)\beta_{t+1}(j)}{\sum_{j=1}^{N}\alpha_T(j)}.$$

Therefore, the need to divide is proved.