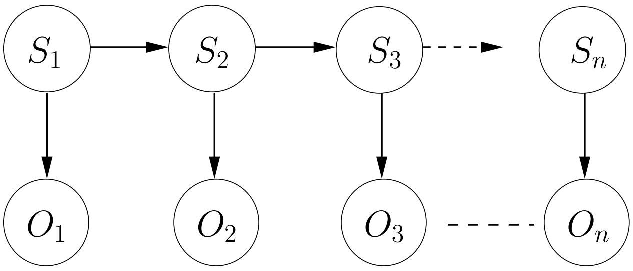
Probabilistic Models: Spring 2014
Gene Finding Example

We are given the following hidden Markov model describing the (simplified) behavior of DNA.



S_0	Θ_{S_0}
Genic	.5
Intergenic	.5

S_{t-1}	S_t	$\Theta_{S_t\|\ S_{t-1}}$
Genic	Genic	.7
Genic	Intergenic	.3
Intergenic	Genic	.3
Intergenic	Intergenic	.7

S_i	O_i	$\Theta_{O_i\|S_i}$
Genic	AT	.9
Genic	CG	.1
Intergenic	AT	.2
Intergenic	CG	.8

1. Use the forward algorithm to calculate the predictive posterior distribution over $S_1 \ldots S_4$ given the following observations: AT, AT, CG, AT.

Table 1: The forward messages.

Time t	$P(S_t = \text{genic} \mid O_1 \ldots O_t)$	$P(S_t = \text{intergenic} \mid O_1 \ldots O_t)$
1	0.8182	0.1818
2	0.8834	0.1166
3	0.1907	0.8093
4	0.7308	0.2692

2. Use the backward algorithm to calculate the smoothed posterior distribution over S_1 given the observations: AT, AT.

Note that we can reuse the forward message from the previous question. So we just need to compute the backward message.

$$P(O_2 = \text{AT} \mid S_1 = \text{genic}) = \sum_{S_2} P(O_2 = \text{AT} \mid S_2) P(S_2 \mid S_1 = \text{genic}) P(O_3 \ldots O_t \mid S_2)$$

We do not have any more observations, so we can drop the third term.

$$\begin{aligned}
&= P(O_2 = \text{AT} \mid S_2 = \text{genic}) P(S_2 = \text{genic} \mid S_1 = \text{genic}) + \\
&\qquad P(O_2 = \text{AT} \mid S_2 = \text{intergenic}) P(S_2 = \text{intergenic} \mid S_1 = \text{genic}) \\
&= .9(.7) + .2(.3) \\
&= .69
\end{aligned}$$

Similarly, $P(O_2 = \text{AT} \mid S_1 = \text{intergenic}) = .41$. Combining this with the forward message, we get that

$$\begin{aligned}
P(S_1 = \text{genic} \mid O_1 = \text{AT}, O_2 = \text{AT}) &\propto 0.8182(0.69), \text{ and} \\
P(S_1 = \text{intergenic} \mid O_1 = \text{AT}, O_2 = \text{AT}) &\propto 0.1818(0.41).
\end{aligned}$$

Multiplying and then normalizing gives that

$$\begin{aligned} P(S_1 = \text{genic}|O_1 = \text{AT}, O_2 = \text{AT}) &\approx 0.883, \text{and} \\ P(S_1 = \text{intergenic}|O_1 = \text{AT}, O_2 = \text{AT}) &\approx 0.117. \end{aligned}$$

3. Use the Viterbi algorithm to find the most likely instantiation of $S_1 \dots S_4$ given the observations: AT, AT, CG, CG.

   To find the most likely instantiation of S_1, we need to find the value of S_0 which maximizes $P(S_1 = s_1, S_0|O_1 = \text{AT})$ for each value of S_1.

   First, we consider when $S_1 = \text{genic}$.

$$\begin{aligned} P(S_1 = s_1|O_1 = \text{AT}) &\propto P(O_1 = \text{AT}|S_0, S_1 = \text{genic})P(S_1 = \text{genic}|S_0) \\ &= P(O_1 = \text{AT}|\ S_1 = \text{genic}) \max_{S_0} P(S_1 = \text{genic}|S_0)P(S_0) \\ &\text{We now consider } S_0 = \text{genic} \\ &= P(O_1 = \text{AT}|S_1 = \text{genic})P(S_1 = \text{genic}|S_0 = \text{genic})P(S_0 = \text{genic}) \\ &= .9(.7)(.5) \\ &= 0.315 \\ &\text{We now consider } S_0 = \text{intergenic} \\ &= P(O_1 = \text{AT}|S_1 = \text{genic})P(S_1 = \text{genic}|S_0 = \text{intergenic})P(S_0 = \text{intergenic}) \\ &= .9(.3)(.5) \\ &= 0.135 \\ &\text{We take the max over } S_0 \text{ and find that } P(S_1 = s_1|O_1 = \text{AT}) \propto 0.315. \end{aligned}$$

   A similar set of calculations shows that $P(S_1 = s_1|O_1 = \text{AT}) \propto 0.070$. Since this is the first state on the path, we normalize these values to find that $P(S_1|O_1 = \text{AT}) \approx< 0.8182, 0.1818 >$.

   To find the most likely instantiation of S_2, we need to find the value of S_1 which maximizes $P(S_2 = s_2, S_1|O_1 = \text{AT}, O_2 = \text{AT}$ for each value of S_2.

First, we consider when $S_2 = \text{genic}$.

$$\begin{aligned} P(S_2 = s_2, S_1 | O_1 = \text{AT}, O_2 = \text{AT}) &\propto P(O_2 = \text{AT} | S_2 = \text{genic}) P(S_2 = \text{genic} | S_1) \\ &= P(O_2 = \text{AT} |\ S_2 = \text{genic}) \max_{S_1} P(S_2 = \text{genic} | S_1) \max_{s_0} P(s_0, s_1 | O_1 = \text{AT}) \end{aligned}$$

The second "max" is simply the numbers we calculated in the previous step.

We now consider $S_1 = \text{genic}$

$= P(O_2 = \text{AT} | S_2 = \text{genic}) P(S_2 = \text{genic} | S_1 = \text{genic}) \max_{s_0} P(s_0, s_1 | O_1 = \text{AT})$

$= .9(.7)(.8182)$

$= 0.5155$

We now consider $S_1 = \text{intergenic}$

$= P(O_2 = \text{AT} | S_2 = \text{genic}) P(S_2 = \text{genic} | S_0 = \text{intergenic}) \max_{s_0} P(s_0, s_1 | O_1 = \text{AT})$

$= .9(.3)(.1818)$

$= 0.0491$

We take the max over S_1 and find that $P(S_2 = \text{genic}, S_1 = s_1 | O_1 = \text{AT}, O_2 = \text{AT}) \propto 0.5155$.

A similar set of calculations shows that $P(S_2 = \text{intergenic}, S_1 = s_1 | O_1 = \text{AT}, O_2 = \text{AT}) \propto 0.0491$. Since this is not the first state on the path, these values give us the probabilities of the paths and do not need to be normalized.

Similar calculations show that the values for all of the states are as follows.

Table 2: The path probabilities.

Time t	$P(S_1 \dots S_t = \text{genic} \mid O_1 \dots O_{t-1})$	$P(S_1 \dots S_t = \text{intergenic} \mid O_1 \dots O_{t-1})$
1	0.8182	0.1818
2	0.5155	0.0491
3	0.0361	0.1237
4	0.0037	0.0693

Some useful equations

The forward algorithm

$$P(\text{next state}|\text{observations so far, next observation}) \propto P(\text{next observation}|\text{next state}) \sum_{\text{current state}} P(\text{next state}|\text{ current state}) P(\text{current state}|\text{observations so far})$$

$$P(S_{t+1}|O_1, O_2, \dots, O_{t+1}) \propto P(O_{t+1}|S_{t+1}) \sum_{S_t = s_t} P(S_{t+1} \ |S_t) P(S_t = s_t|O_1, \dots, O_t)$$

The backward algorithm

$$P(S_k|O_1, \dots, O_t) \propto \texttt{forward}(k) P(O_{k+1}, \dots, O_t| \ S_k)$$

$$P(\text{remaining observations}|\text{ current state}) = \sum_{\text{next state}} P(\text{next state}|\text{current state}) P(\text{next observation}|\text{ next state}) P(\text{further observations}|\text{next state})$$

$$P(O_{k+1}, \dots, O_t| \ S_k) = \sum_{S_{k+1} = s_{k+1}} P(S_{k+1} = s_{k+1}|S_k) P(O_{k+1}|S_{k+1}) P(O_{k+2}, \dots, O_t| \ S_{k+1} = s_{k+1})$$

The Viterbi algorithm

$$\max_{\text{path so far}} P(\text{path so far}, \text{next state in path}|\ \text{observations so far}, \text{next observation})$$

$$\propto P(\text{next observation } | \text{ next state}) \left\{ \max_{\text{current state}} P(\text{next state}|\ \text{current state}) \left\{ \max_{\text{previous states}} P(\text{previous states}, \text{current state}|\text{observations so far} \right\}\right\}$$

$$\max_{s_1 \ldots s_t} P(s_1 \ldots s_t, S_{t+1}|O_1 \ldots O_{t+1}) \propto P(O_{t+1}|S_{t+1}) \left\{ \max_{s_t} P(S_{t+1}|s_t) \left\{ \max_{s_1 \ldots s_{t-1}} P(s_1 \ldots s_{t-1}, s_t|\ O_1 \ldots O_t) \right\}\right\}$$