Lecture 8

Gidon Rosalki

2025-11-11

1 Reminder

Last lecture we discussed HMMs (Hidden Markov Models):

$$\mathbb{P}\left[X_1 \dots X_n, H_1 \dots H_n\right] = \mathbb{P}_0\left[H_0\right] \cdot \prod_i \mathbb{P}\left[H_{i+1} | H_i\right] \cdot \prod_i \mathbb{P}\left[X_i | H_i\right]$$
$$F_i\left[s\right] = \mathbb{P}\left[H_i = s, X_1 \dots X_i\right]$$
$$B_i\left[s\right] = \mathbb{P}\left[X_{i+1} \dots X_n | H_i = s\right]$$

We also discussed likelihood, with the probability of a chain being

$$\begin{aligned}\mathbb{P}\left[X_1 \dots X_n\right] &= \sum_s F_n\left[s\right] \\ &= \sum_s B_1\left[s\right] \mathbb{P}_1\left[s\right] \\ \mathbb{P}\left[H_i = s | X_1 \dots X_n\right] &\propto F_i\left[s\right] \cdot B_i\left[s\right] \\ \mathbb{P}\left[H_i = s | H_{i+1} = t | X_1 \dots X_n\right] &\propto \text{ something, we will discuss soon}\end{aligned}$$

2 Maximum probability reconstruction

We want to find the most likely Markov Chain

$$arg \max_{h_i \dots h_n} \left\{\mathbb{P}\left[H_1 = h_1 \dots H_n = h_n, X_1 \dots X_n\right]\right\}$$

How do we solve this problem? Dynamic Programming! I'm shocked! We will construct the following expression, to help us reach a problem solvable with DP:

$$\max_{h_1 \dots h_i} \left\{\mathbb{P}\left[H_1 = h_1 \dots H_i = h_i, X_1 \dots X_i\right]\right\}$$
$$\max_{h_1 \dots h_{i-1}} \left\{\max_{h_i} \left\{\mathbb{P}\left[H_1 = h_1 \dots H_{i-1} = h_{i-1}, X_1 \dots X_{i-1}\right] \cdot \tau\left[h_i, h_{i-1}\right] \cdot \pi\left[x_i, h_i\right]\right\}\right\}$$
$$= \max_{h_1 \dots h_{i-1}} \left\{\mathbb{P}\left[H_1 = h_1, \dots, H_{i-1} = h_{i-1}, X_1 \ddots X_{i-1}\right] \left(\max_{h_i} \left\{\tau\left[h_i, h_{i-1}\right] \pi\left[x_i, h_i\right]\right\}\right)\right\}$$

Since

$$\begin{aligned}\mathbb{P}\left[H_1 \dots H_i, X_1 \dots X_i\right] = &\ \mathbb{P}\left[H_1 \dots H_{i-1}, X_1 \dots X_{i-1}\right] \\ &\cdot \mathbb{P}\left[H_i | H_1 \dots H_{i-1}, X_1 \dots X_{i-1}\right] \\ &\cdot \mathbb{P}\left[X_i | H_i, H_1 \dots H_{i-1}, X_1 \dots X_{i-1}\right]\end{aligned}$$

The final solution for the optimum comes out to

$$V_i\left[s\right] = \max_t \left\{V_{i-1}\left[t\right] \cdot \tau\left[s, t\right] \pi\left[x_i, s\right]\right\}$$

So, the optimal solution, of length i, finishing at the point $H_i = s$, is the optimal solution until $H_{i-1} = t$, of length $i - 1$, with the optimum transition between s and t.
Why are we using V? Due to the name of the algorithm being **Viterbi**, named after its creator Andwer Viterbi.

Consider if you have a sequence $X_1 \dots X_n$, upon which you run Viterbi, and get $\hat{h}_1 \dots \hat{h}_n$. So, if we consider what happens at i:

$$\mathbb{P}\left[H_i = \hat{h}_i | X_1 \dots X_n\right]$$

What can we say about this?

3 Learning HMMs

Let us consider the full data

$$\begin{bmatrix} h_1 & h_2 & \dots & h_{n_1} \\ x_1 & x_2 & \dots & x_{n_2} \end{bmatrix}$$

So, firstly remember that each state is i.i.d. Then the probability of the data:

$$\begin{aligned}
\mathbb{P}\left[\text{Data}\right] &= \prod_m \mathbb{P}\left[h_1\left[m\right]\dots h_{n_m}\left[m\right], X_1\left[m\right]\dots X_{n_m}\left[m\right]\right] \\
&= \prod_m \left[p_0\left[h_1\left[m\right]\right] \prod_{i=1}^{n_m-1} \mathbb{P}\left[H_i = h_i\left[m\right] | H_{i-1} = h_{i-1}\left[m\right]\right]\right] \\
&= \prod_m \left[\mathbb{P}_0\left[h_1\left[m\right]\right] \prod_{i=1}^{h_m-1} \tau\left[h_i\left[m\right], h_{i-1}\left[m\right]\right] \cdot \prod_i \pi\left[X_i\left[m\right], h_i\left[m\right]\right]\right]
\end{aligned}$$

Let us define a statistic that counts how many times we had s

$$N_{0,s} = \sum_m \mathbb{1}_{n_1[m]=s}$$

Where

$$\begin{aligned}
N_{t,s} &= \sum_m \sum_i \mathbb{1}_{h_{i-1}[m]=t, h_i[m]=s} \\
N_{s,x} &= \sum_r \sum_i \mathbb{1}_{h_i[r]=s, X_i[r]=x}
\end{aligned}$$

So:

$$\begin{aligned}
\mathbb{P}\left[\text{Data}\right] &= \prod_m \mathbb{P}\left[h_1\left[m\right]\dots h_{n_m}\left[m\right], X_1\left[m\right]\dots X_{n_m}\left[m\right]\right] \\
&= \left[\prod_s \mathbb{P}_0\left[s\right]^{N_{0,s}}\right] \cdot \left[\prod_{t\in S}\prod_{s\in S} \tau\left[s,t\right]^{N_{t,s}}\right] \cdot \left[\prod_{s\in S}\prod_{x\in\mathcal{X}} \pi\left[x,s\right]^{N_{s,x}}\right]
\end{aligned}$$

Where $\mathcal{X}$ is all the observations (ie, all the options that can happen in π).

We can now also discuss for our MLE:

$$\begin{aligned}
\hat{p}_0\left[s\right] &= \frac{N_0, s}{\sum_t N_{0,t}} \\
\hat{\tau}\left[s,t\right] &= \frac{N_{t,s}}{\sum_u N_{t,u}} \\
\hat{\pi}\left[s,x\right] &= \frac{N_{s,x}}{\sum_y N_{s,y}}
\end{aligned}$$

Let us now consider the case where we have many sequences $\left[X_1\left[1\right], \dots, X_{n_1}\left[1\right]\right]$, and $\left[X_1\left[2\right], \dots, X_{n_2}\left[2\right]\right]$. Our hidden state is now hidden. So

$$\begin{aligned}
\mathbb{P}\left[\text{Data}\right] &\overset{iid}{=} \prod_m \mathbb{P}\left[X_1\left[m\right]\dots X_{n_m}\left[m\right]\right] \\
&= \prod_m \sum_{n_i} \sum_{h_{n_m}} \mathbb{P}\left[\text{same contents}\right]
\end{aligned}$$

So this is hard, because combining together sums and multiplications is difficult. We cannot simply extract the sums from each other. We also do not know *anything* about the hidden states, there could be many between every other state.