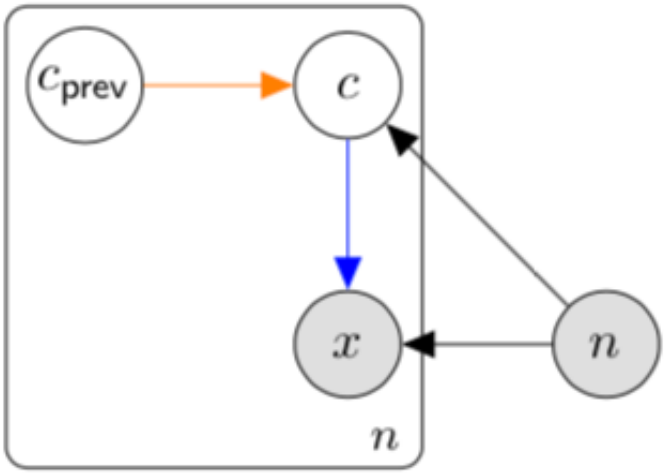
Table of contents

Hidden Markov Models

The Hidden Markov Model **HMM** models the joint probability of a sequence of words and their corresponding tags via a first-order Markov model over the tags where words are conditionally independent given their corresponding tags.



We consider two generative processes:

- Transition: we move from one "state" to another "state" where our state is the POS tag
- Emission: with a certain "state" in mind, we generate a certain word

Let us give names to things, let's model the current class with a random variable C and let's use the random variable C_{prev} to model the previous category. For the word we will use the random variable X. Both C and C_{prev} take on values in the enumeration of a tagset containing t tags, that is, $\{1, \ldots, t\}$. X takes on values in the enumeration of a vocabulary containing v words, that is, $\{1, \ldots, v\}$.

The **transition** distribution captures how our beliefs in a class vary as a function of the previous class. We will use Categorical distributions for that. In fact, for each possible previous class we get a Categorical distribution over the complete set of classes.

$$(1) \qquad C \mid C_{\text{prev}} = p \sim \text{Cat}(\lambda_1^{(p)}, \ldots, \lambda_t^{(p)})$$

The **emission** distribution captures how our beliefs in a word vary as a function of the word's class. We will again use Categorical distributions for that. In fact, for each possible class, we get a Categorical distribution over the complete vocabulary.

$$X \mid C = c \sim \text{Cat}(\theta_1^{(c)}, \dots, \theta_v^{(c)}) \tag{2}$$

The HMM assigns a joint distribution over a sentence x_1^n and its tag sequence c_1^n.

To get to this joint distribution, let's first express $P_{CX|C_{\text{prev}}}$, a joint distribution over classes and words, where we focus on a single step. Then the model factorises as follows:

$$P_{XC|C_{\text{prev}}}(x, c|c_{\text{prev}}) = P_{C|C_{\text{prev}}}(c|c_{\text{prev}})P_{X|C}(x|c) \tag{3}$$

We can then simply iterate over the steps in a sequence pair generating both observations:

$$\begin{aligned} P_{X_1^n C_1^n|N}(x_1^n, c_1^n|n) &= P_{XC|C_{\text{prev}}}(x_1, c_1|c_0)P_{XC|C_{\text{prev}}}(x_2, c_2|c_1) \cdots P_{XC|C_{\text{prev}}}(x_n, c_n|c_{n-1}) \\ &= P_{C|C_{\text{prev}}}(c_1|c_0)P_{X|C}(x_1|c_1)P_{C|C_{\text{prev}}}(c_2|c_1)P_{X|C}(x_2|c_1) \cdots P_{C|C_{\text{prev}}}(c_n|c_{n-1} \\ &= \prod_{i=1}^{n} P_{C|C_{\text{prev}}}(c_i|c_{i-1})P_{X|C}(x_i|c_i) \end{aligned} \tag{4}$$

Marginal probability

Errata There was an error in how we exposed the marginal probability previously. It will be corrected in class, and here you will find a step by step view of the correct solution

---

In case we use an HMM as a language model, our ultimate goal is to assing a probability to a sentence x_1^n, regardless of its tag sequence. For that we need to marginalise away all possible assignments to C_1^n, where every C_i may take on any of the t available tags.

$$
\begin{aligned}
(5) \qquad P_{S|N}(x_1^n|n) &= \sum_{c_1=1}^{t} \cdots \sum_{c_n=1}^{t} P_{X_1^n C_1^n|N}(x_1^n, c_1^n|n) \\
&= \sum_{c_1=1}^{t} \cdots \sum_{c_n=1}^{t} \underbrace{\prod_{i=1}^{n} P(c_i|c_{i-1})P(x_i|c_i)}_{\text{From Eq (4)}} \\
&= \sum_{c_1=1}^{t} \cdots \sum_{c_n=1}^{t} P(c_1|c_0)P(x_1|c_1)P(c_2|c_1)P(x_2|c_1) \cdots P(c_n|c_{n-1})P(x_n|c_n)
\end{aligned}
$$

This looks pretty bad! If we have to enumerate all possible tag sequences, there would be just too many of them. That is, in the first sum, c_1 takes 1 of t values, then for each of those values c_2 will take 1 of t values, and so on. This leads to t^n different tag sequences. An exponential number of them!!! We will never manage to enumerate them, compute their joint probabilities and then sum them up.

Let's see if we can simplify this task!

Let's start easy with a short sequence where $n = 3$ and let's consider marginalising the first tag out. First recall the joint distribution:

$$
\begin{aligned}
(6) \qquad P(\langle x_1, x_2, x_3 \rangle, \langle c_1, c_2, c_3 \rangle) &= P_{C|C_{\text{prev}}}(c_1|c_0 = \text{BoS})P_{X|C}(x_1|c_1) \\
&\times P_{C|C_{\text{prev}}}(c_2|c_1)P_{X|C}(x_2|c_2) \\
&\times P_{C|C_{\text{prev}}}(c_3|c_2)P_{X|C}(x_3|c_3) \\
&\times P_{C|C_{\text{prev}}}(\text{EoS}|c_3)P_{X|C}(\text{</s>}|\text{EoS})
\end{aligned}
$$

Recall that transition to `-EOS-` tag and emission of `</s>` token are factored in for convenience. To avoid clutter we hide the conditioning on length.

Recall that C_0 is always deterministically set to `-BOS-`, thus there's no uncertainty as to which tag precedes c_1.

Let's call $\alpha_1(c)$ the total probability that we generate $X = x_1, C_1 = c$, namely,

$$
(7) \qquad \alpha_1(c_1) = P_{X|C}(x_1|c_1)P_{C|C_{\text{prev}}}(c_1|c_0 = \text{BoS})
$$

Now let's marginalise assignments to C_1:

$$
\begin{aligned}
(8) \qquad P(\langle x_1, x_2, x_3 \rangle, \langle \cdot, c_2, c_3 \rangle) &= \sum_{c_1=1}^{t} P(\langle x_1, x_2, x_3 \rangle, \langle c_1, c_2, c_3 \rangle) \\
&= \sum_{c_1=1}^{t} \underbrace{\underbrace{P(c_1|c_0)P(x_1|c_1)}_{\alpha_1(c_1)} P(c_2|c_1)P(x_2|c_2)P(c_3|c_2)P(x_3|c_3)P_{XC|C_{\text{prev}}}(\text{</s>}, \text{EoS}|c_3)}_{\text{From Eq (3)}} \\
&= \sum_{c_1=1}^{t} \alpha_1(c_1) \times P(c_2|c_1)P(x_2|c_2)P(c_3|c_2)P(x_3|c_3)P_{XC|C_{\text{prev}}}(\text{</s>}, \text{EoS}|c_3) \\
&= \underbrace{P(x_2|c_2) \times \left(\sum_{c_1=1}^{t} \alpha_1(c_1) \times P(c_2|c_1) \right)}_{\alpha_2(c_2)} \times P(c_3|c_2)P(x_3|c_3)P_{XC|C_{\text{prev}}}(\text{</s>}, \text{EoS}|c_3)
\end{aligned}
$$

Note that:

- we reuse the result in (7)
- we factorised $P(x_2|c_2)$ out of the sum, as it does not depend on c_1 ;
- we also factorised $P(c_3|c_2)P(x_3|c_3)P_{XC|C_{\text{prev}}}(\text{</s>}, \text{EoS}|c_3)$ out of the sum, as it also does not depend on c_1
- we identified $\alpha_2(c_2)$: which refers to the marginal probability where we have marginalised C_1, we have just generated $X_2 = x_2, C_2 = c_2$, and we are yet to generate $(\langle x_3, \text{</s>} \rangle, \langle c_3, \text{EoS} \rangle)$

Now let's marginalise assignments to C_2:

$$
\begin{aligned}
(9) \qquad P(\langle \cdot, \cdot, c_3 \rangle, \langle x_1, x_2, x_3 \rangle) &= \sum_{c_2=1}^{t} P(\langle \cdot, c_2, c_3 \rangle, \langle x_1, x_2, x_3 \rangle) \\
&= \sum_{c_2=1}^{t} \underbrace{\alpha_2(c_2) \times P(c_3|c_2) P(x_3|c_3) P_{XC|C_{\text{prev}}}(\text{</s>}, \text{EoS}|c_3)}_{\text{From Eq (8)}} \\
&= \underbrace{P(x_3|c_3) \times \left(\sum_{c_2=1}^{t} \alpha_2(c_2) \times P(c_3|c_2) \right)}_{\alpha_3(c_3)} P_{XC|C_{\text{prev}}}(\text{</s>}, \text{EoS}|c_3)
\end{aligned}
$$

- where we reuse the previous results
- again we factor a term out, namely $P(x_3|c_3)$, as it does not depend on c_2
- this time we factore $P_{XC|C_{\text{prev}}}(\text{</s>}, \text{EoS}|c_3)$ out because it does not depend on c_2 either
- we identify $\alpha_3(c_3)$: which refers to the marginal probability where we have marginalised joint assignments to C_1 and C_2, we have just generated $X_3 = x_3, C_3 = c_3$ and are yet to generate $(\text{</s>}, \text{EoS})$.

Now let's finally marginalise assignments to C_3

$$
\begin{aligned}
(10) \qquad P(\langle \cdot, \cdot, \cdot \rangle, \langle x_1, x_2, x_3 \rangle) &= \sum_{c_3=1}^{t} P(\langle \cdot, \cdot, c_3 \rangle, \langle x_1, x_2, x_3 \rangle) \\
&= \sum_{c_3=1}^{t} \underbrace{\alpha_3(c_3) \times P_{XC|C_{\text{prev}}}(\text{</s>}, \text{EoS}|c_3)}_{\text{From Eq (9)}} \\
&= \underbrace{P_{X|C_{\text{prev}}}(\text{</s>}|\text{EoS}) \times \left(\sum_{c_3=1}^{t} \alpha_3(c_3) \times P_{C|C_{\text{prev}}}(\text{EoS}|c_3) \right)}_{\alpha_4(\text{EoS})} \\
&= P_{X_1^n|N}(x_1^3|n)
\end{aligned}
$$

which yields the marginal of interest, namely, the probability of the sequence of words regardless of tags!

The quantity $\alpha_i(c)$ we identified along the way is called the **forward probability**, for an observation $x_{\leq i}$ (note that we incude the ith word here), it corresponds to the probability of marginalising out the sequence $C_{<i}$ (ith tag not included), and generating the pair $X_i = x_i, C_i = c$.

$$\alpha_i(c) = P_{X|C}(x_i|c) \sum_{c_{i-1}=1}^{t} \alpha_{i-1}(c_{i-1}) \times P_{C|C_{\text{prev}}}(c|c_{i-1}) \tag{11}$$

This recursive formula can be efficient implemented to yield marginal probabilities (an iterative implementation is also possible).

The marginal probability of a sentence x_1^n is therefore

$$P_{S|N}(x_1^n|n) = \alpha_{n+1}(\text{EoS}) \tag{12}$$

that is, the probability of x_1^n where:

- we marginalise joint assignments to C_1^n
- and generate the end of sequence symbols: $X_{n+1} = \text{</s>}, C_{n+1} = \text{EoS}$.