CS 520: Training HMMs

16:198:520

Instructor: Wes Cowan

A Hidden Markov Model is defined by a state space S, a set of output symbols O, an initial distribution π ($\pi(x) = \mathbb{P}(X_0 = x)$), a transition matrix A ($A_{x',x} = \mathbb{P}(X_{t+1} = x | X_t = x')$), and an output matrix B ($B_{x,y} = \mathbb{P}(Y_t = y | X_t = x)$). A key assumption of the model is that **these probabilities are independent of time**. This model gives rise to a sequence of states $X_0, X_1, X_2, \ldots$, the hidden or latent variables, which in turn gives rise to a sequence of observed values or measurements $Y_1, Y_2, \ldots$. Consider a sequence of output values $\underline{y}_T = \{y_1, y_2, \ldots, y_T\}$. It is convenient to denote the likelihood of a given output sequence for a specified model in the following way:

$$\begin{aligned}
&\mathbb{P}\left(\underline{Y}_T = \underline{y}_T | \pi, A, B\right) \\
&= \mathbb{P}\left(Y_1 = y_1, Y_2 = y_2, \ldots, Y_T = y_t | \pi, A, B\right) \\
&= \sum_{x_0, x_1, \ldots, x_T} \mathbb{P}\left(X_0 = x_0, X_1 = x_1, Y_1 = y_1, X_2 = x_2, Y_2 = y_2, \ldots, X_T = x_T, Y_T = y_t | \pi, A, B\right) \\
&= \sum_{x_0, x_1, \ldots, x_T} \mathbb{P}\left(X_0 = x_0 | \pi, A, B\right) \prod_{t=1}^{T} \mathbb{P}\left(X_t = x_t | X_{t-1} = x_{t-1}, \pi, A, B\right) \prod_{t=1}^{T} \mathbb{P}\left(Y_t = y_t | X_t = x_t, \pi, A, B\right) \\
&= \sum_{x_0, x_1, \ldots, x_T} \pi(x_0) \prod_{t=1}^{T} A_{x_{t-1}, x_t} \prod_{t=1}^{T} B_{x_t, y_t}
\end{aligned} \tag{1}$$

If the model is known, various questions can be systematically put to it as seen with **filtering**, **prediction**, **smoothing**, and **maximum likelihood estimation**. But where did this model come from? Given a set of output data, how can we determine a good model to fit to it? How can we determine goodness of fit? One common approach to model fitting is to consider *the model most likely to have given rise to the data* i.e., given observations $\underline{y}_T = \{y_1, \ldots, y_T\}$, find the model that maximizes the likelihood of this output sequence occuring:

$$\begin{aligned}
(\pi^*, A^*, B^*) &= \text{argmax}_{\pi, A, B}\ \mathbb{P}\left(\underline{Y}_T = \underline{y}_T | \pi, A, B\right) \\
&= \text{argmax}_{\pi, A, B} \sum_{x_0, x_1, \ldots, x_T} \pi(x_0) \prod_{t=1}^{T} A_{x_{t-1}, x_t} \prod_{t=1}^{T} B_{x_t, y_t}.
\end{aligned} \tag{2}$$

This problem is well-defined - we can consider the constrained maximization problem of optimizing the above sum for π, A, B, and the parameters that resulted would give us our optimal model. The problem becomes that while the problem is well-defined, it is largely intractable due to the exponentially large number of turns in the summation, looking at all possible sequences of states that might've given rise to a sequence of observations. So what can be done?

The **Baum-Welch** algorithm proceeds by starting with an initial, potentially arbitrary model (π, A, B) and iteratively updating it to try to improve its goodness of fit. Given some (π, A, B), it is convenient, in the following analysis, to define the following functions, which were seen previously as the backwards and forwards message passing algorithms for *filtering* and *smoothing*:

$$\alpha_t(x) = \mathbb{P}\left(X_t = x | Y_1 = y_1, \ldots, Y_t = y_t, \pi, A, B\right), \tag{3}$$

where

$$\begin{aligned}
\alpha_0(x) &= \mathbb{P}\left(X_0 = x | \pi, A, B\right) = \pi(x) \\
\alpha_{t+1}(x) &\propto \sum_{x_t} \alpha_t(x_t) A_{x_t, x} B_{x, y_{t+1}},
\end{aligned} \tag{4}$$

utilizing the filtering equations, or

$$\begin{aligned}\alpha_0(x) &= \pi(x) \\ \alpha_{t+1}(x) &= \frac{\sum_{x_t} \alpha_t(x_t) A_{x_t,x} B_{x,y_{t+1}}}{\sum_{x_{t+1}} \sum_{x_t} \alpha_t(x_t) A_{x_t,x_{t+1}} B_{x_{t+1},y_{t+1}}}.\end{aligned} \tag{5}$$

Similarly, we define

$$\beta_t(x) = \mathbb{P}\left(Y_t = y_t, Y_{t+1} = y_{t+1}, \ldots, Y_T = y_T | X_t = x, \pi, A, B\right), \tag{6}$$

which gives us via the backwards recursion part of smoothing:

$$\begin{aligned}\beta_T(x) &= B_{x,y_T} \\ \beta_t(x) &= \sum_{x_{t+1}} B_{x,y_t} A_{x,x_{t+1}} \beta_{t+1}(x_{t+1}).\end{aligned} \tag{7}$$

Given π, A, B, we want to use the data collected $(y_1, \ldots, y_T)$ to construct a new model $\pi^{\text{new}}, A^{\text{new}}, B^{\text{new}}$ that is better, more informed by the collected data. To begin, consider the following question: from a Bayesian perspective, what has this new data told us about the initial distribution of states? i.e., for a given state x, what can we say about $\mathbb{P}(X_0 = x | \underline{Y}_T = \underline{y}_T, \pi, A, B)$? We have

$$\begin{aligned}&\mathbb{P}(X_0 = x | \underline{Y}_T = \underline{y}_T, \pi, A, B) \\ &\propto \mathbb{P}(X_0 = x, \underline{Y}_T = \underline{y}_T, \pi, A, B) \\ &\propto \sum_{x_1} \mathbb{P}(X_0 = x, X_1 = x_1, \underline{Y}_T = \underline{y}_T, \pi, A, B) \\ &\propto \sum_{x_1} \mathbb{P}\left(X_0 = x, \pi, A, B\right) \mathbb{P}\left(X_1 = x_1 | X_0 = x, \pi, A, B\right) \mathbb{P}\left(Y_1 = y_1, \ldots, Y_T = y_T | X_1 = x_1, \pi, A, B\right) \\ &\propto \sum_{x_1} \pi(x) A_{x,x_1} \beta_1(x_1).\end{aligned} \tag{8}$$

This, in essence, tells us *given our current model, and the sequence of observations, what should we believe about the initial state at time* 0? But we can then use this as the basis for a new initial distribution. Factoring in the necessary normalization, this results in:

$$\pi^{\text{new}}(x) = \frac{\sum_{x_1} \pi(x) A_{x,x_1} \beta_1(x_1)}{\sum_{x_0} \sum_{x_1} \pi(x_0) A_{x_0,x_1} \beta_1(x_1)}. \tag{9}$$

We now want to similarly construct new transition and observation matrices, ideally in a similar way. However, here we run into a problem. Ideally, we'd like to say:

$$A^{\text{new}}_{x',x} = \mathbb{P}\left(X_{t+1} = x | X_t = x', \underline{Y}_T = \underline{y}_T, \pi, A, B\right). \tag{10}$$

The problem is that the above probability might actually depend on t - this would violate the whole premise of the Hidden Markov Model, which is that the transition probabilities do not depend on t. To adjust for this, we effectively average over the entire time interval, look at *the expected number of* $x' \to x$ *transitions* divided by *the expected number of* x' *occurrences*. We can take these expectations, conditioned on the indicated data, giving us an

effectively smoothed 'average' probability of transition, informed by the data. We can write this as

$$\begin{aligned} A_{x',x}^{\text{new}} &= \frac{\mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}\left\{X_t = x', X_{t+1} = x\right\} | \underline{Y}_T = \underline{y}_T, \pi, A, B\right]}{\mathbb{E}\left[\sum_{t=0}^{T-1} \mathbb{1}\left\{X_t = x'\right\} | \underline{Y}_T = \underline{y}_T, \pi, A, B\right]} \\ &= \frac{\sum_{t=0}^{T-1} \mathbb{P}\left(X_t = x', X_{t+1} = x | \underline{Y}_T = \underline{y}_T, \pi, A, B\right)}{\sum_{t=0}^{T-1} \mathbb{P}\left(X_t = x' | \underline{Y}_T = \underline{y}_T, \pi, A, B\right)} \\ &= \frac{\sum_{t=0}^{T-1} \mathbb{P}\left(X_t = x' | \underline{Y}_T = \underline{y}_T, \pi, A, B\right) \mathbb{P}\left(X_{t+1} = x | X_t = x', \underline{Y}_T = \underline{y}_T, \pi, A, B\right)}{\sum_{t=0}^{T-1} \mathbb{P}\left(X_t = x' | \underline{Y}_T = \underline{y}_T, \pi, A, B\right)}, \end{aligned} \tag{11}$$

which can be seen as a weighted average of the quantity we were originally interested in, where the concerning dependence on time has now been averaged out. It is useful then to define the following 'intermediary' quantities, which can be seen as related to smoothing:

$$\begin{aligned} \gamma_t(x') &= \mathbb{P}\left(X_t = x' | \underline{Y}_T = \underline{y}_T, \pi, A, B\right) \\ \eta_t(x', x) &= \mathbb{P}\left(X_{t+1} = x | X_t = x', \underline{Y}_T = \underline{y}_T, \pi, A, B\right) \end{aligned} \tag{12}$$

Unpacking these the same way as we do generally (HMMs are essentially just the same three tricks, over and over and over again): For $0 \leq t \leq T-1$, we have

$$\begin{aligned} \gamma_t(x') &\propto \mathbb{P}\left(X_t = x', \underline{Y}_T = \underline{y}_T, \pi, A, B\right) \\ &\propto \sum_{x_{t+1}} \mathbb{P}\left(X_t = x', X_{t+1} = x_{t+1} \underline{Y}_T = \underline{y}_T, \pi, A, B\right) \\ &\propto \sum_{x_{t+1}} \mathbb{P}\left(X_t = x' | \underline{Y}_t = \underline{y}_t, \pi, A, B\right) \mathbb{P}\left(X_{t+1} = x_{t+1} | X_t = x', \pi, A, B\right) \mathbb{P}\left(Y_{t+1} = y_{t+1}, \ldots, Y_T = y_T | X_{t+1} = x_{t+1}, \pi, A, B\right) \\ &\propto \sum_{x_{t+1}} \alpha_t(x') A_{x',x} \beta_{t+1}(x_{t+1}), \end{aligned} \tag{13}$$

or, adjusting for normalization,

$$\gamma_t(x') = \frac{\sum_{x_{t+1}} \alpha_t(x') A_{x',x} \beta_{t+1}(x_{t+1})}{\sum_{x_t} \sum_{x_{t+1}} \alpha_t(x_t) A_{x_t,x} \beta_{t+1}(x_{t+1})} \tag{14}$$

It is worth noting, for completeness, that

$$\gamma_T(x') = \mathbb{P}\left(X_T = x', \underline{Y}_T = \underline{y}_T, \pi, A, B\right) = \alpha_T(x'). \tag{15}$$

Similarly, for $0 \leq t \leq T-1$,

$$\begin{aligned} \eta_t(x', x) &\propto \mathbb{P}\left(X_{t+1} = x, X_t = x', \underline{Y}_T = \underline{y}_T, \pi, A, B\right) \\ &\propto \alpha_t(x') A_{x',x} \beta_{t+1}(x), \end{aligned} \tag{16}$$

or

$$\eta_t(x', x) = \frac{\alpha_t(x') A_{x',x} \beta_{t+1}(x)}{\sum_{x_t} \sum_{x_{t+1}} \alpha_t(x_t) A_{x_t,x_{t+1}} \beta_{t+1}(x_{t+1})}. \tag{17}$$

With these two intermediary functions, we now have:

$$A^{\text{new}}_{x',x} = \frac{\sum_{t=0}^{T-1} \gamma_t(x')\eta_t(x',x)}{\sum_{t=0}^{T-1} \gamma_t(x')}, \tag{18}$$

Finally, we consider the problem of constructing an updated emission or observation matrix B^{new}. Similar to A^{new}, we can't simply consider the Bayesian update for the observation probabilities, because this can introduce a dependence on time. We solve this in the same way, smoothing out or averaging over all the relevant timesteps (in this case, $t = 1, \dots T$ to account for no observation at time 0):

$$\begin{aligned} B^{\text{new}}_{x,y} &= \frac{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{X_t = x, Y_t = y\right\} | \underline{Y}_T = \underline{y}_T, \pi, A, B\right]}{\mathbb{E}\left[\sum_{t=1}^{T} \mathbb{1}\left\{X_t = x\right\} | \underline{Y}_T = \underline{y}_T, \pi, A, B\right]} \\ &= \frac{\sum_{t=1}^{T} \mathbb{P}\left(X_t = x, Y_t = y | \underline{Y}_T = \underline{y}_T, \pi, A, B\right)}{\sum_{t=1}^{T} \mathbb{P}\left(X_t = x | \underline{Y}_T = \underline{y}_T, \pi, A, B\right)} \\ &= \frac{\sum_{t=1}^{T} \mathbb{P}\left(X_t = x | \underline{Y}_T = \underline{y}_T, \pi, A, B\right) \mathbb{P}\left(Y_t = y | X_t = x, \underline{Y}_T = \underline{y}_T, \pi, A, B\right)}{\sum_{t=1}^{T} \mathbb{P}\left(X_t = x | \underline{Y}_T = \underline{y}_T, \pi, A, B\right)} \\ &= \frac{\sum_{t=1}^{T} \gamma_t(x) \mathbb{P}\left(Y_t = y | X_t = x, \underline{Y}_T = \underline{y}_T, \pi, A, B\right)}{\sum_{t=1}^{T} \gamma_t(x)} \end{aligned} \tag{19}$$

It remains to simplify the remaining probability in the above summation. However, notice that the conditioning actually includes $Y_t = y_t$. Conditioned on this, the probability is 1 if $y = y_t$, and 0 otherwise. Hence the whole thing reduces to:

$$B^{\text{new}}_{x,y} = \frac{\sum_{t=1}^{T} \gamma_t(x) \mathbb{1}\left\{y_t = y\right\}}{\sum_{t=1}^{T} \gamma_t(x)}. \tag{20}$$

At this point, we have enough to specify the full Baum-Welch algorithm:

The **Baum-Welch Algorithm** begins with an initial Hidden Markov Model (π, A, B) and a sequence of observed data $y_1, \ldots, y_T$, and generates an updated model in the following way:

- Compute the Forward/Backward Recursions:

$$\begin{aligned} \alpha_0(x) &= \pi(x) \\ \alpha_{t+1}(x) &= \frac{\sum_{x_t} \alpha_t(x_t) A_{x_t,x} B_{x,y_{t+1}}}{\sum_{x_{t+1}} \sum_{x_t} \alpha_t(x_t) A_{x_t,x_{t+1}} B_{x_{t+1},y_{t+1}}} \\ \beta_T(x) &= B_{x,y_T} \\ \beta_t(x) &= \sum_{x_{t+1}} B_{x,y_t} A_{x,x_{t+1}} \beta_{t+1}(x_{t+1}). \end{aligned} \tag{21}$$

- Compute the Intermediary Functions:

$$\begin{aligned} \gamma_t(x') &= \frac{\sum_{x_{t+1}} \alpha_t(x') A_{x',x} \beta_{t+1}(x_{t+1})}{\sum_{x_t} \sum_{x_{t+1}} \alpha_t(x_t) A_{x_t,x} \beta_{t+1}(x_{t+1})} \\ \gamma_T(x') &= \alpha_T(x') \\ \eta_t(x', x) &= \frac{\alpha_t(x') A_{x',x} \beta_{t+1}(x)}{\sum_{x_t} \sum_{x_{t+1}} \alpha_t(x_t) A_{x_t,x_{t+1}} \beta_{t+1}(x_{t+1})}. \end{aligned} \tag{22}$$

- Compute the updated model, $\pi^{\text{new}}, A^{\text{new}}, B^{\text{new}}$:

$$\begin{aligned} \pi^{\text{new}}(x) &= \frac{\sum_{x_1} \pi(x) A_{x,x_1} \beta_1(x_1)}{\sum_{x_0} \sum_{x_1} \pi(x_0) A_{x_0,x_1} \beta_1(x_1)} \\ A^{\text{new}}_{x',x} &= \frac{\sum_{t=0}^{T-1} \gamma_t(x') \eta_t(x', x)}{\sum_{t=0}^{T-1} \gamma_t(x')} \\ B^{\text{new}}_{x,y} &= \frac{\sum_{t=1}^{T} \gamma_t(x) \mathbb{1}\{y_t = y\}}{\sum_{t=1}^{T} \gamma_t(x)}. \end{aligned} \tag{23}$$

This process can be iterated, generating a new model from $\pi^{\text{new}}, A^{\text{new}}, B^{\text{new}}$ until convergence is achieved.

0.1 A Cautionary Word

There are two words of warning that might be offered at this point, one practical and one more theoretical. On the practical side, one of the main difficulties in implementing this algorithm is that the relevant probabilities (and function values along the way) may ultimately be quite small, approaching the limit of what can be represented in a given programming language. This is worth being careful about - to some extent, this can be ameliorated by renormalization, rescaling things appropriately to keep them at reasonable levels.

The more theoretical concern is the following: can we be guaranteed that this process converges to the model that maximizes the likelihood of the data (as in our original consideration above)? Over all, we cannot - this represents an iterative, local-search optimization algorithm that risks getting caught in local minima. Random restarts (re-choosing initial π, A, B can potentially overcome this somewhat). But a slightly more interesting question is this - should we care? Obviously the answer is yes to some extent, but recall that the likelihood that the 'maximum likelihood estimator' is the correct one is actually quite small. Getting within spitting distance of it has the potential to be just as effective in most practical settings. However, taking a step back, it is important to remember that for the data

we are collecting, the idea that the underlying process is a Hidden Markov Model to begin with is an assumption, or a simplification. There is no 'true' Hidden Markov Model to discover - but we can hope to discover parameters for our Hidden Markov Model that capture, as accurately as we are able, the data we are given.

0.2 Results on a Small Example

As a small example, a passage of text from ***Dracula*** (1897) was treated as the output of a hidden Markov model, and the Baum-Welch algorithm was utilized to estimate the underlying parameters.

This can be approached in one of two ways: in the first, the individual characters are treated as the output tokens, and each internal state has some associated distribution of what character it will output when visited. This is a very etymological example, as the internal states (ideally) would represent common roots or associations between letters in words. With an output length of 5000 characters, taken form the middle of the book, a randomly initialized HMM on 50 states produces output that looks utterly random:

```
"'i i.vrcvd._g;ose?__'iyq-;r!e-mr:_.xb?t,moa so.nuakilljtc,!.a"'u

aw-oyy!hebicuxvwjc'osn "nu.

g"_,e
```

Even just a few rounds of training produces significantly different results:

```
he

fated ety are ocater, bigastlerew tao, inl focs em t'on cot

livee nen y thy too hor a-co tad tond
```

The model has clearly learned that punctuation marks cannot be thrown in anywhere, and has begun to learn the statistical distribution of word lengths (or, at the very least, 'space' frequency). Recognizable words or letter sequences have begun to appear as well.

The same problem can be treated taking the tokens to be the individual words. This is potentially a much more interesting example; in some sense, the internal states can be thought of as representing underlying concepts, giving rise to words that describe them and leading to other concepts and their related words. Taking a passage from the middle of the book of about 10000 words with approximately 2100 unique words, an initial HMM on 100 generates garbage, albeit legible garbage:

```
walks growl earth disastrous enjoyed followed do.  tigers hereafter application, latched outside presenc
fatal, theory." blissful throat horrors night; eat why suspicious * dad gave above easily specialist,
pallor blood--to town came smile, again.  hands live, grotesque breed.  included.  obscene nerves
whose always had business, night-a-huntin' "tell except wondering sight.  take.  yet way.  those _13
seward's placidly, madmen feebly.  god hall, fust--the 'e said.  lor' waked wan-looking body can't
repeated surer above, with.  garlic." greater me,' holmwood rubbed much terrors.  danger.  point mine
water, waited because pagan half-remembrance formed, a-'owling.  well.

we'll growl burst him september_.--i language make tea-time,
```

Several rounds of training later, basic sentence structure has started to emerge, and while the text is far from sensible by itself, individual phrases and intelligible word groupings have begun to emerge.

```
them swiftly yet treatment off, with a kind and will not tell in the breathing and in, call in the
children and at us; and blest such and she present, with in nostril wolves in past worse.  and my
dear, there bersicker influence in any kind somewheres what deserve towards infinitely yes."
```

```
i felt good, and i might much before which took over his he and then he wife to him length how * you
proceeded none.  look kiss sarcastic-like transfusion in the most memory, in my cause and with the
band.  in my wolf and vein will not amsterdam.
```