CS 181 Section: Hidden Markov Models

Vincent Nguyen

Week of April 11, 2016

1 Introduction

Last week, we introduced mixture models. This unsupervised probabilistic clustering method operates under the hypothesis that all data points are generated from a finite mixture of distributions with unknown parameters. This week, we introduce Hidden Markov Models (HMM). We will see how a single time slice of an HMM can be interpreted as an extension of a mixture model where the mixture depends on the *previous* observation. But first, let's briefly go through the modeling of *sequence* data.

Since an HMM deals with sequential data, let us represent a particular sequence, X, as

$$X = (x_1, x_2, \ldots, x_N) \tag{1}$$

These are known as the *observed* states.

Each observed state has a corresponding *latent* state which we denote as

$$Z = (z_1, z_2, \ldots, z_N) \tag{2}$$

where N is the length of the sequence. This is the *hidden* part of the HMM. Note: we had previously used N to denote *all* training examples but in dealing with sequential data, we use N to denote the sequence length. The HMM is an example of a *directed graph* which we can represent with Figure 1.

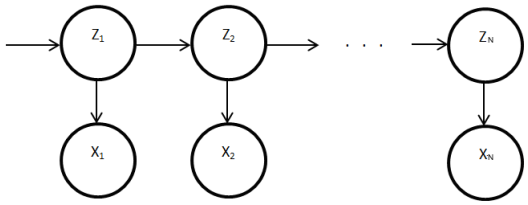

Figure 1: Observed and Hidden States of an HMM

2 Mixture Model Tie-In to HMM

In our regular mixture model, we had the familiar joint probability for one single example x_n:

$$p(x_n, z_{nk}|\theta_k) = \prod_{k=1}^{K} \Big(\pi_k p(x_n|\theta_k)\Big)^{z_{nk}} \tag{3}$$

If we consider the hidden states of the HMM, then we might want a way of representing the probability of some state z_n at position n given the previous state z_{n-1} at the previous position $n-1$. Let us represent the probability distribution of the latent variable z_n to be *conditioned* on the previous latent variable, so $p(z_n|z_{n-1})$. We can present all of these probabilities in a table, A, that denotes *transition probabilities* between latent variables:

$$A = \begin{pmatrix} A_{11} & A_{12} & \dots & A_{1K} \\ A_{21} & A_{22} & \dots & A_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ A_{K1} & A_{K2} & \dots & A_{KK} \end{pmatrix} \tag{4}$$

where

- A_{12} is the probability of going from z_1 to z_2, $A_{12} \equiv p(z_{n2} = 1|z_{n-1,1} = 1)$
- A_{22} is the probability of staying in z_2, $A_{22} \equiv p(z_{n2} = 1|z_{n-1,2} = 1)$
- $0 \leq A_{jk} \leq 1$ (entries are probabilities) and $\sum_k A_{jk} = 1$ (rows sum to 1)

We can now show the conditional distribution of z_n given the previous latent state and A as

$$p(z_n|z_{n-1}, A) = \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j}\ z_{nk}} \tag{5}$$

This notation can get confusing, but remember that n refers to the inputs in sequence while j and k explicitly refer to which one-hot encoding $k \in K$. All that Equation 5 is showing is the entry in A for the row corresponding to the state of z_{n-1} and the column corresponding to the state of z_n.

Unfortunately, we cannot generate Equation 5 for all states in the sequence. Starting out at the first observation in the sequence where $n = 1$, we have no information about the previous latent state, so instead, we represent the first state as

$$p(z_1) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \tag{6}$$

Alright, so we've managed to define Markovian conditional distributions for latent variables, so let's now incorporate our inputs x and unknown distribution parameters θ_k. For one single input x_n, this is represented as a vector of K entries:

$$p(x_n|z_n, \theta) = \prod_{k=1}^{K} p(x_n|\theta_k)^{z_{nk}} = \begin{pmatrix} p(x_n|\theta_1)^{z_{n1}} \\ p(x_n|\theta_2)^{z_{n2}} \\ \vdots \\ p(x_n|\theta_K)^{z_{nK}} \end{pmatrix} \tag{7}$$

Here, θ_k represents model parameter(s). If our underlying model is a Gaussian, then we would have μ_k and $\sum_k$. However, if we have a Bernoulli, then we would just have μ_k (see last week's section notes). We'll just keep the model general this time around.

Now, taking Equation 5, Equation 6, and Equation 7, and generalizing for the entire sequence, then we have the joint probability

$$p(X, Z|\theta) = p(z_1) \left[\prod_{n=2}^{N} p(z_n|z_{n-1}, A) \right] \prod_{n'=1}^{N} p(x_{n'}|z_{n'}, \theta) \tag{8}$$

$$= \prod_{k=1}^{K} \pi_k^{z_{1k}} \left[\prod_{n=2}^{N} \prod_{k=1}^{K} \prod_{j=1}^{K} A_{jk}^{z_{n-1,j}\ z_{nk}} \right] \prod_{n'=1}^{N} \prod_{k=1}^{K} p(x_{n'}|\theta_k)^{z_{n'k}} \tag{9}$$

We will eventually want to maximize to obtain the optimal parameters π_k, A_{jk}, and θ_k so the log-likelihood is

$$L(\theta) = \sum_{k=1}^{K} z_{1k} \ln \pi_k + \sum_{n=2}^{N} \sum_{k=1}^{K} \sum_{j=1}^{K} (z_{n-1,j}\ z_{nk}) \ln A_{jk} + \sum_{n'=1}^{N} \sum_{k=1}^{K} z_{n'k} \ln p(x_{n'}|\theta_k) \tag{10}$$

We might be tempted to simply use this log-likelihood and do maximum likelihood estimation (MLE) to obtain our parameters. However, we cannot directly maximize because we currently are dealing with an intractable expression. If we take a step back, what we are actually dealing with is

$$p(X|\theta) = \sum_{Z} p(X, Z|\theta) = \sum_{Z} p(Z|X, \theta) p(X|\theta) \tag{11}$$

If this is unfamiliar, think back to last week's section. We introduce a latent variable and want to marginalize it out. This can be done with integration with respect to dZ or summing over all Z.

If we take the $\log$ of $p(X|\theta)$, then we have a $\log \sum$ expression which is not fun to work with. Since the likelihood function currently is a generalization of a mixture distribution, we can use Expectation Maximization to find our parameters!

3 EM for HMM

The general Expectation Maximization (EM) idea says that for X, θ, θ^{old},

$$\sum_{Z} p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) = \mathbb{E}_Z[\ln p(X, Z|\theta)] \tag{12}$$

The proof is complicated and a bit beyond the scope of this section, so I wouldn't worry about understanding the fine details.[1] Essentially, we can now just plug in Equation 10 and are no longer dealing with a $\log \sum$ expression! What Equation 12 says is that we can maximize $L(\theta)$ by doing two steps. First, we do the E-step and then the M-step. This is much more tractable to deal with than direct MLE of Equation 11. Going forwards, we introduce the expression

$$Q(\theta, \theta^{\text{old}}) = \sum_{Z} p(Z|X, \theta^{\text{old}}) \ln p(X, Z|\theta) \tag{13}$$

As a refresher, the EM algorithm consists of two steps:

1. **E-Step**: Take the old model parameters or some initialization (if first starting out) denoted as θ^{old} and evaluate the posterior $p(Z|X, \theta^{\text{old}})$. Calculate the Expectation of Q using the calculated posterior distribution.
2. **M-Step**: Maximize the parameters of π_k, A_{jk}, and θ_k. These are now the parameters that we will plug into the next **E-Step**. This step is just MLE.

In the mixture model, we only have one formulation of z but here, we have two which are located in the first & second and then the third additive parts of Equation 10. Let us define the notation for the expectation with respect to those latent variables:

$$\gamma(z_n) = p(z_n|X, \theta^{\text{old}}) \tag{14}$$

$$\xi(z_{n-1}, z_n) = p(z_{n-1}, z_n|X, \theta^{\text{old}}) \tag{15}$$

where the expectation of a one-hot vector is just the probability that its entry k is 1:

$$\gamma(z_{nk}) = \mathbb{E}[z_{nk}] = \sum_{z} \gamma(z) z_{nk} \tag{16}$$

$$\xi(z_{n-1,j}\ z_{nk}) = \mathbb{E}[z_{n-1,j}\ z_{nk}] = \sum_{z} \gamma(z) z_{n-1,j}\ z_{nk} \tag{17}$$

The expectation of the log-likelihood with respect to both these latent variable formulations is

$$Q(\theta, \theta^{\text{old}}) = \sum_{k=1}^{K} \gamma(z_{1k}) \ln \pi_k + \sum_{n=2}^{N} \sum_{k=1}^{K} \sum_{j=1}^{K} \xi(z_{n-1}, j\ z_{nk}) \ln A_{jk} + \sum_{n'=1}^{N} \sum_{k=1}^{K} \gamma(z_{n'k}) \ln p(x_{n'}|\theta_k) \tag{18}$$

[1] http://www.cs.nyu.edu/~mohri/asr12/lecture_8.pdf

3.1 M-Step

We are just starting with the M-Step first because it is easier. We won't go into detail about the specifics of these derivations. They are very similar to the M-Step derivations for a general mixture model. See the previous section notes on the Canvas site for in-depth derivations. The optimal parameters are

$$\hat{\pi}_k = \frac{\gamma(z_{1k})}{\sum_{j=1}^{K} \gamma(z_{1j})} \tag{19}$$

$$\hat{A}_{jk} = \frac{\sum_{n=2}^{N} \xi(z_{n-1}, j\ z_{nk})}{\sum_{k'=1}^{K} \sum_{n=2}^{N} \xi(z_{n-1}, j\ z_{nk'})} \tag{20}$$

We did not include the optimal θ_k because that is dependent on the distribution type of the model.

3.2 E-Step

This is much harder so we've included it after the M-Step (but the algorithm goes E-Step then M-Step). We want to be able to evaluate Equation 14 and Equation 15. The details of this aren't so important for the midterm versus knowing what EM is so don't be too alarmed. We'll be using d-separation to work out the recursion relations as I've not shown every step of the way.

For $\gamma(z_n)$, we can use Bayes Theorem to show that

$$\gamma(z_n) = p(z_n|X) = \frac{p(X|z_n)p(z_n)}{p(X)} \tag{21}$$

$$= \frac{p(x_1, \ldots, x_n, z_n)p(x_{n+1}, \ldots, x_N|z_n)}{p(X)} \tag{22}$$

$$= \frac{\alpha(z_n)\beta(z_n)}{p(X)} \tag{23}$$

where have a way to calculate $\gamma(z_n)$. We can make use of a recursion relation to show that

$$\alpha(z_n) = p(x_n|z_n) \sum_{z_{n-1}} \alpha(z_{n-1})p(z_n|z_{n-1}) \tag{24}$$

as z_1 has no parent node, we have the initial condition that

$$\alpha(z_1) = \prod_{k=1}^{K} \Big(\pi_k p(x_1|\theta_k)\Big)^{z_{1k}} \tag{25}$$

This α computation is known as the forward pass is we are using previous states to calculate the current one.

Likewise, we can evaluate $\beta(z_n)$ via a recursion relation

$$\beta(z_n) = \sum_{z_{n+1}} p(x_{n+2}, \ldots, x_N|z_{n+1})p(x_{n+1}|z_{n+1})p(z_{n+1}|z_n) \tag{26}$$

$$= \sum_{z_{n+1}} \beta(z_{n+1}) p(x_{n+1}|z_{n+1}) p(z_{n+1}|z_n) \tag{27}$$

This is known as the backward passing algorithm since we are evaluating $\beta(z_n)$ in terms of $\beta(z_{n+1})$. We set $\beta(z_N) = 1$ as an initialization.

Now that we have the α and β to calculate γ, we want a way to calculate ξ.

$$\xi(z_{n-1}, z_n) = p(z_{n-1}, z_n|X) = \frac{p(X|z_{n-}, z_n) p(z_{n-1}, z_n)}{p(X)} \tag{28}$$

$$= \frac{p(x_1, \ldots, x_{n-1}|z_{n-1}) p(x_n|z_n) p(x_{n+1}, \ldots, x_N|z_n) p(z_n|z_{n-1}) p(z_{n-1})}{p(X)} \tag{29}$$

$$= \frac{\alpha(z_{n-1}) p(x_n|z_n) p(z_n|z_{n-1}) \beta(z_n)}{p(X)} \tag{30}$$

This is quite laborious and is more than you need to know for exam pruposes. We included this because it is nice to be able to understand how to evaluate the EM algorithm for a HMM.