Note Set 5: Hidden Markov Models

Probabilistic Learning: Theory and Algorithms, CS 274A, Winter 2017

1 Hidden Markov Models (HMMs)

1.1 Introduction

Consider observed data vectors x_t that are d-dimensional and where the subscript t indicates discrete time or position in a sequence, $t = 1, \ldots, T$. We can assume for notational simplicity that the x_t vector is real-valued, but in general x_t could be discrete or could be a mixture of discrete and real-valued.

Our data consists of $D = \{x_1, \ldots, x_T\}$. Unlike the IID assumption we have made in the past we would now like to model the sequential dependence among the x_t's. One approach is to use a *hidden Markov model* where we assume that the x_t's are noisy stochastic functions of an unobserved (hidden) Markov chain denoted by z_t, where z_t is a discrete random variable taking one of K possible values, $z_t \in \{1, \ldots, K\}$. z_t is often referred to as the state variable.

The generative model for a hidden Markov model is simple: at each time step t, a data vector x_t is generated conditioned on the state z_t, and the Markov chain then transitions to a new state z_{t+1} to generate x_{t+1}, and so on. As with standard Markov chains there is an initial distribution π over the K states to initialize the chain.

There are two key assumptions in a hidden Markov model:

1. Observations x_t are conditionally independent of all other variables given z_t, so the observation at time t depends only on the current state z_t.
2. The z_t's form a (first order) Markov chain, i.e., $p(z_t|z_{t-1}, \ldots, z_1) = p(z_t|z_{t-1}), t = 2, \ldots, T$. The chain is also typically assumed to be homogeneous in that the transition probabilities do not depend on t.

The earliest application of hidden Markov models was in speech recognition in the 1970's, but they have since been used for various other problems in bioinformatics, language modeling, economics, and climate modeling.

We will use the shorthand notation $x_{[1,T]}$ for $x_1, \ldots, x_T$, and $z_{[1,T]}$ for $z_1, \ldots, z_T$. Our observed data is $D = \{x_1, \ldots, x_T\}$. From our graphical model we have

$$p(x_{[1,T]}, z_{[1,T]}) = \prod_{t=1}^{T} p(x_t|z_t)p(z_t|z_{t-1})$$

(where $p(z_1|z_0) = \pi$, the initial distribution on states.)

We have two sets of parameters:

- Transition matrix, $K \times K$ matrix A, with $a_{ij} = p(z_t = j|z_{t-1} = i)$, $1 \leq i \leq K$.
- K emission distributions/densities $p(x_t|z_t = j), j = 1 \ldots K$, e.g., multivariate Gaussian for real-valued x_t, and usually assumed to be homogeneous (i.e., does not depend on t). If x_t is very high-dimensional it is common to assume that the components of x are conditionally independent given z_t.

For simplicity we will assume that π the initial distribution on states is known, e.g., set to the uniform distribution, but if we had multiple sequences this could also be learned from the data. We will let θ indicate the full set of parameters, i.e., the transition matrix parameters in A and the parameters of K state-dependent emission distributions $p(x_t|z_t = j), j = 1 \ldots K$.

Note the similarity of the HMM to a finite mixture model with K components. In particular, the HMM can be viewed as adding Markov dependence to the unobserved component indicator variable z in a mixture model.

1.2 Efficient Computation of HMM Likelihood

Below we show how to compute the likelihood $L(\theta)$ (where both the K emission density parameters and the transition matrix A are unknown)?

$$\begin{aligned} L(\theta) &= p(D|\theta) \\ &= p(x_{[1,T]}|\theta) \\ &= \sum_{z_{[1,T]}} p(x_{[1,T]}, z_{[1,T]}|\theta) \end{aligned}$$

But this sum is intractable to compute directly, since it has complexity $O(K^T)$. However, we can use the conditional independence assumptions (or equivalently the graphical model structure in the HMM) to carry out this computation efficiently.

Let $\alpha_t(j) = p(z_t = j, x_{[1,t]}), j = 1, \ldots, K$ (implicitly conditioning on θ). This is the joint probability of (a) the unobserved state at time t being in state j and (b) all of the observed x_t's up to and including time

t.

$$\begin{aligned}
\alpha_t(j) &= p(z_t = j, x_{[1,t]}) \\
&= \sum_{i=1}^{K} p(z_t = j, z_{t-1} = i, x_{[1,t]}) \\
&= \sum_{i} p(x_t | z_t = j, z_{t-1} = i, x_{[1,t-1]})\, p(z_t = j | z_{t-1} = i, x_{[1,t-1]}])\, p(z_{t-1} = i, x_{[1,t-1]}) \\
&= \sum_{i=1}^{K} p(x_t | z_t = j)\, p(z_t = j | z_{t-1} = i)\, p(z_{t-1} = i, x_{[1,t-1]}) \\
&= \sum_{i=1}^{K} p(x_t | z_t = j)\, a_{ij}\, \alpha_{t-1}(i)
\end{aligned}$$

The first term is the "evidence" from the observation x_t at time t, the second term is the transition probability, and the final term is just $\alpha_{t-1}(i)$. This yields a simple recurrence relation for the α's.

We can compute the $\alpha_t(j)$'s recursively, where $\alpha_t(j) = p(z_t = j, x_{[1,t]})$. We can do this in a single forward pass, from $t = 1$ up to $t = T$, initializing the recursion with $\alpha_0(j) = \pi(j)$. This is the forward part of the well-known *forward-backward* algorithm for HMMs. Then given $\alpha_T(j), j = 1, \ldots, K$, the likelihood can be computed as $L(\theta) = \sum_j \alpha_T(j)$, by the LTP.

If we know $\alpha_{t-1}(i), i = 1, \ldots, K$, we can compute $\alpha_t(j)$ in time $O(K^2 + Kf(d))$. The K^2 is because we have to compute the probability for all i, j pairs, and the function f reflects the complexity of computing the likelihood of the data vector x_t for each possible state, e.g., $f(d) = O(d^2)$ for a Gaussian emission density. The overall complexity of computing all of the α's is $O(TK^2 + TKf(d))$.

1.3 Efficient Computation of State Probabilities

In developing an EM algorithm for HMMs we will want to compute the probability of each possible state at each time t given all of the observed data, i.e., $p(z_t = j | x_{[1,T]})$ (using all of the data, both before and after t). We factor it as follows.

$$\begin{aligned}
p(z_t = j | x_{[1,T]}) &\propto p(z_t = j, x_{[1,t]}, x_{[t+1,T]}) \\
&= p(x_{[t+1,T]} | z_t = j, x_{[1,t]})\, p(z_t = j, x_{[1,t]}) \\
&= p(x_{[t+1,T]} | z_t = j)\, p(z_t = j, x_{[1,t]}) \\
&= p(x_{[t+1,T]} | z_t = j)\, \alpha_t(j)
\end{aligned}$$

Note that given $z_t = j$, the $x_{[1,t]}$ values give us no additional information about $x_{[t+1,T]}$ (which is how we get from the 2nd to the 3rd line above).

Define $\beta_t(j) = p(x_{[t+1,T]} | z_t = j),\ \ t = 1, \ldots, T,\ \ j = 1, \ldots, K$. Then, from above, we have

$$p(z_t = j | x_{[1,T]}) \propto \beta_t(j) \alpha_t(j)$$

Using the same type of recursive decomposition as we used for $\alpha_t(j)$, the $\beta_t(j)$'s can be computed in time $O(TK^2 + TKf(d))$, working backwards from $t = T$ to $t = 1$.

Thus, to compute $p(z_t = j|x_{[1,T]}),\ t = 1, \ldots, T,\ j = 1, \ldots, K$:

- We first recursively compute the $\alpha_t(j)$'s (forward step).
- Next we recursively compute the $\beta_t(j)$'s (backward step).
- Finally, we can compute $p(z_t = j|x_{[1,T]})$ as

$$w_t(j) \;=\; p(z_t = j|x_{[1,T]}) \;=\; \frac{\alpha_t(j)\beta_t(j)}{\sum_k \alpha_t(k)\beta_t(k)}$$

where the denominator is the normalization term. This yields a set of $T \times K$ probabilities $w_t(j)$, playing the same role in EM as the membership weights for mixture models.

2 EM for learning HMM Parameters

The EM algorithm for HMMs follows the same general idea as the EM algorithm for finite mixtures.

In the E-step we compute the probabilities $w_t(j)$ (or membership weights) of the unobserved states, for each state j and each time t, conditioned on all of the data $x_{[1,T]}$ and conditioned on the current parameters θ.

In the M-step we compute point estimates of the parameters given the membership weights from the E-step. There are two different sets of parameters: (1) the emission density parameters $p(x_t|z_t = j)$, and (2) the transition parameters $a_{ij}, 1 \leq i, j, \leq K$.

The estimation of the emission density parameters proceeds in exactly the same manner as for the finite mixture case. For example, if the emission densities are Gaussian, then the membership weights are used to generate "fractional counts" for estimating the mean and covariance for each of the K emission densities.

For the transition probabilities we proceed as follows. We first compute $E[N_j]$, the expected number of times in state j, which is $\sum_{t=1}^{T} w_t(j)$.

Next, we need to compute $E[N_{ij}]$, the expected number of times we transition from state i to state j.

$$\begin{aligned}
E[N_{ij}] &= \sum_{t=1}^{T-1} p(z_t = i, z_{t+1} = j|x_{[1,T]}) \\
&\propto \sum_{t=1}^{T-1} p(z_t = i, z_{t=1} = j, x_{[1,T]})
\end{aligned}$$

Letting $\gamma_t(i,j) = p(z_t = i, z_{t=1} = j, x_{[1,T]})$, we have

$$\begin{aligned}
\gamma_t(i,j) &= p(z_t = i, z_{t=1} = j, x_{[1,T]}) \\
&= p(x_{[t+2,T]}|z_{t+1} = j)\; p(x_{t+1}|z_{t+1} = j)\; p(z_{t+1} = j|z_t = i)\; p(z_t = i, x_{[1,t]}) \\
&= \beta_{t+1}(j)\; p(x_{t+1}|z_{t+1} = j)\; a_{ij}\; \alpha_t(i).
\end{aligned}$$

In going from the first to the second line we have used various conditional independence properties that exist in the model. The final line consists of quantities that can easily be computed, e.g., they be computed directly from the model ($p(x_{t+1}|z_{t+1} = j)$), or are known parameters (a_{ij}), or have been computed during the forward-backward computations of the E-step (the β's and α's).

We then normalize the γ's to get the conditional probabilities we need, i.e.,

$$p(z_t = i, z_{t+1} = j|x_{[1,T]}) = \frac{\gamma_t(i,j)}{\sum_{k_1}\sum_{k_2}\gamma_t(k_1,k_2)}$$

from which we can compute $E[N_{ij}]$ above.

The M step for the transition probabilities is now very simple:

$$\hat{a}_{ij} = \frac{E[N_{ij}]}{E[N_j]},\ 1 \leq i,j \leq K$$