Announcements

Extra resources for EM and HMM posted on lecture page.

(Pages from Bishop, HMM tutorial by Rabiner)

Data distribution from an HMM

An HMM is specified by: $\theta = \{\pi, A, \phi\}$

$$p(X,Z|\theta) = p(z_1|\pi) \prod_{n=2}^{N} p(z_n|z_{n-1}, A) \prod_{m=1}^{N} p(x_m|z_m, \phi)$$

(complete data, i.e., we can generate from this).

Transition matrix representation

(Not a graphical model)

Data distribution from an HMM

Transition probability to another state is 5%
Classic HMM computational problems

Given data, what is the HMM (learning).

Given an HMM, what is the distribution over the state variables. Also, how likely are the observations, given the model.

Given an HMM, what is the most likely state sequence for some data?

Learning the HMM

If we know the states, we can compute the parameters.

If we know the parameters, we can compute the states (if we know how to solve the second problem).

General EM algorithm

1. Choose initial values for $\theta^{(1)}$
   (can also do assignments, but then jump to M step).
2. E step: Evaluate $\rho(Z|X, \theta^{(t)})$
3. M step: Evaluate $\theta^{(t+1)} = \arg \max_{\theta} \{ Q(\theta^{(t+1)}, \theta^{(t)}) \}$
   where $Q(\theta^{(t+1)}, \theta^{(t)}) = \sum_Z p(Z|X, \theta^{(t)}) \log \{ p(X, Z|\theta^{(t+1)}) \}$
4. Check for convergence; If not done, goto 2.

* At each step, our objective function is increases unless it is at a local maximum. It is important to check this is

Q function for EM for HMM

$$\log \{ p(X, Z|\theta) \} = \sum_{z=1}^K \gamma(z_{nt}) \log(\pi) + \sum_{n=2}^N \sum_{j=1}^K \sum_{z_{n-1}} \sum_{z_n} \gamma(z_{n-1}) \log \{ p(z_n|z_{n-1}, A) \} + \sum_{n=1}^N \sum_{z_n} \log \{ p(x_n|z_n, \phi) \}$$

By analogy with the GMM

$$Q(\theta^{(t+1)}, \theta^{(t)}) = \sum_Z p(Z|\theta^{(t)}) \log \{ X, Z|\theta^{(t+1)} \}$$

$$= \sum_{z=1}^K \gamma(z_{nt}) \log(\pi) + \sum_{n=2}^N \sum_{j=1}^K \sum_{z_{n-1}} \sum_{z_n} \gamma(z_{n-1}) \log \{ p(z_n|z_{n-1}, A) \}$$

$$+ \sum_{n=1}^N \sum_{z_n} \gamma(z_{nt}) \log \{ p(x_n|z_n, \phi) \}$$
EM for HMM

Doing the maximization using Lagrange multipliers

\[
\pi_k = \frac{\gamma(z_{ik})}{\sum_i \gamma(z_{ik})} \text{ (Much like the mixture model case)}
\]

\[
A_{ik} = \frac{\sum_{i'=2}^{n} \zeta(z_{i-1,j}, z_{ik})}{\sum_{i'=2}^{n} \sum_{i''=2}^{n} \zeta(z_{i-1,j}, z_{i''k})}
\]

E step for EM for HMM

Computing the E step is a bit more involved.

Recall that in the mixture case it was easy because we only needed to consider the relative likelihood that each cluster independently explain the observations.

However, here the sequence also must play a role.

Graphical model for the E step

Note that our task is to compute marginal probabilities
Computing marginals in an HMM

Various names, flavors, notations, ...

Forward-Backward algorithm

Alpha-beta algorithm

Sum-product for HMM

(Bishop also says “Baum Welch” but that is a synonym for the EM algorithm as whole).

\[
\gamma(z_n) = p(z_n | X) = \frac{p(X | z_n) p(z_n)}{p(X)} = \frac{p(x_1, \ldots, x_n | z_n) p(z_n)}{p(X)} = \frac{p(x_1, \ldots, x_n, z_n) p(x_{n+1}, \ldots, x_N | z_n)}{p(X)} = \frac{\alpha(z_n) \beta(z_n)}{p(X)} \]

Where we define
\[
\alpha(z_n) = p(x_1, \ldots, x_n, z_n) \quad \beta(z_n) = p(x_{n+1}, \ldots, x_N | z_n)
\]

\[
\gamma(z_n) = p(z_n | X) = \frac{p(X | z_n) p(z_n)}{p(X)} = \frac{p(x_1, \ldots, x_n | z_n) p(z_n)}{p(X)} = \frac{p(x_1, \ldots, x_n, z_n) p(x_{n+1}, \ldots, x_N | z_n)}{p(X)} = \frac{\alpha(z_n) \beta(z_n)}{p(X)} \]

Where we define
\[
\alpha(z_n) = p(x_1, \ldots, x_n, z_n) \quad \beta(z_n) = p(x_{n+1}, \ldots, x_N | z_n)
\]

Note that in EM, \( p(X) = p(X | \theta^{(s)}) \)
Alpha-beta algorithm

\[ \alpha(z_n) = p(x_n | z_n) \sum_{z_{n+1}} \alpha(z_{n+1}) p(z_{n+1} | z_n) \]

This is a recursive evaluation of alpha. So we can compute all of them easily if we know the first one, \( \alpha(z_1) \).

\[ \alpha(z_t) = p(x_t | z_t) \]
\[ = p(x_t | z_t) p(z_t) \quad (K \text{ dimensional vector}) \]

\[ \alpha(z_t) = \pi_t p(x_t | \phi_t) \]

Similarly, we can derive a recurrence relation for beta

\[ \beta(z_n) = \sum_{z_{n+1}} \beta(z_{n+1}) p(x_{n+1} | z_{n+1}) p(z_{n+1} | z_n) \]

The details for the computation:

\[ \beta(z_n) = p(x_{n+1}, \ldots, x_N | z_n) \]
\[ = \sum_{z_{n+1}} p(x_{n+1}, \ldots, x_N, z_{n+1} | z_n) \]
\[ = \sum_{z_{n+1}} p(x_{n+1}, \ldots, x_N | z_{n+1}, z_n) p(z_{n+1} | z_n) \]
\[ = \sum_{z_{n+1}} p(x_{n+1}, \ldots, x_N | z_{n+1}) p(z_{n+1} | z_n) \]
\[ = \sum_{z_{n+1}} p(x_{n+1}, \ldots, x_N | z_{n+1}) p(x_{n+1} | z_{n+1}) p(z_{n+1} | z_n) \]
\[ = \sum_{z_{n+1}} p(x_{n+1}, \ldots, x_N | z_{n+1}) p(x_{n+1} | z_{n+1}) p(z_{n+1} | z_n) \]

Our recurrence relation for beta

\[ \beta(z_n) = \sum_{z_{n+1}} \beta(z_{n+1}) p(x_{n+1} | z_{n+1}) p(z_{n+1} | z_n) \]

We can compute the betas if we know the last one.

\[ p(z_n | x) = \frac{\alpha(z_n) \beta(z_n)}{p(x)} \]
\[ = p(x, z_n) \beta(z_n) \]

So \( \beta(z_n) = 1 \).
**Alpha-beta algorithm**

Given the alphas and betas, we can compute all the quantities we need for the E step.

\[
\gamma(z_n) = \frac{\alpha(z_n) \beta(z_n)}{p(X)}
\]

Note that \(\sum \gamma(z_n) = 1\)

so \(\sum \frac{\alpha(z_n) \beta(z_n)}{p(X)} = 1\)

and \(p(X) = \sum \alpha(z_n) \beta(z_n)\)

We do not need \(p(X)\) for EM, but it is the likelihood which we want to monitor \(p(X) = p(X|p^{\alpha})\).

---

**Computing marginals, version two**

We can apply sum-product to our E step graph.

Recall the recursive nature of the inference problem when we studied the sum-product algorithm.

We could have done recursion (like we just did now), but message passing is more general.

---

**Recall sum-product**

\[
\xi(z_{n-1}, z_n) = p(z_{n-1}, z_n | X)
\]

\[
= \frac{p(X | z_{n-1}, z_n)p(z_{n-1}, z_n)}{p(X)}
\]

\[
= \frac{p(x_1, \ldots, x_{n-1} | z_{n-1})p(x_n | z_{n-1}, \ldots, x_n)p(x_{n+1} | z_n)p(z_n | z_{n-1})p(z_{n+1} | z_n)}{p(X)}
\]

\[
= \frac{\alpha(z_{n-1})p(x_n | z_{n-1})p(z_n | z_{n-1}) \beta(z_n)}{p(X)}
\]

(13.43)
Computing marginals, version two

We can apply sum-product to our E step graph.

\[
\begin{align*}
\mu_{x_n \rightarrow f_a}(x_1) &= 1 \\
\mu_{x_2 \rightarrow f_b}(x_3) &= \mu_{f_a \rightarrow x_2}(x_3) \mu_{f_b \rightarrow x_2}(x_3) \\
\mu_{x_2 \rightarrow f_c}(x_4) &= \sum_{x_1} f_c(x_2, x_4) \\
\mu_{x_3 \rightarrow f_c}(x_4) &= \sum_{x_2} f_c(x_3, x_4) \\
\mu_{x_4 \rightarrow f_b}(x_3) &= \sum_{x_2} f_b(x_4, x_3)
\end{align*}
\]

Since we condition on all the x’s, we can simplify the graph.
Sum-product for HMM

\[ h = p(z_1) p(x_1 | z_1) \]
\[ f_n = p(z_n | z_{n-1}) p(x_n | z_n) \]

We can ignore the nodes because they just pass through the incoming message on the single in link.

i.e., \( \mu_{f_n \rightarrow z_n} (z_n) = \mu_{f_n \rightarrow f_{n+1}} (z_n) \)

Factor node actions on left to right messages

\[ \mu_{f_n \rightarrow f_{n+1}} (z_n) = \sum_{z_{n-1}} f_n (z_{n-1}, z_n) \mu_{f_{n-1} \rightarrow f_n} (z_{n-1}) \]

The first message is

\[ h = p(z_1) p(x_1 | z_1) = \alpha(x_1, z_1) \]

If we identify \( \mu_{f_n \rightarrow f_{n+1}} (z_n) = \alpha(z_n) \)

\[ \alpha(z_n) = \sum_{z_{n-1}} f_n (z_{n-1}, z_n) \alpha(z_{n-1}) \]
\[ = \sum_{z_{n-1}} p(z_n | z_{n-1}) p(x_n | z_n) \alpha(z_{n-1}) \]
\[ = p(x_n | z_n) \sum_{z_{n-1}} p(z_n | z_{n-1}) \alpha(z_{n-1}) \]