

1 EM review

Judging from the last hw, it seems many students have a fundamental misunderstanding of the E-step

Practically speaking, *the E-step is simply computing the the expected complete data log-likelihood given the observed data and current estimate of the parameters*, hence the 'E' in E-step.

The expected complete data log-likelihood is used as a surrogate for the true incomplete data log-likelihood, and you maximize it with respect to the parameters for the M-step.

Most of the confusion about the E-step came from the description of using the averaging distribution q . This is how the ecll relates:

- The description of the E-step in the book is

$$q^{(t+1)} = \arg \max_q \mathcal{L}(q, \theta^{(t)}) = p(z|x, \theta^{(t)})$$

So q is just the conditional probability of the hidden variables given the observed data and current value of the parameters. You don't directly need this conditional probability though.

- The M-step is:

$$\begin{aligned} \theta^{(t+1)} &= \arg \max_{\theta} \mathcal{L}(p(z|x, \theta^{(t)}), \theta) \\ &= \arg \max_{\theta} \langle l_c(\theta; x, z) \rangle_{p(z|x, \theta^{(t)})} - \sum_q q(z|x) \log q(z|x) \\ &= \arg \max_{\theta} \text{ecll}(\theta) \end{aligned}$$

So the quantity of interest is really the ecll and not $p(z|x, \theta^{(t)})$ itself.

The "interpretation" that you're somehow replacing unobserved data with the conditional expectations comes from the fact that the complete data log-likelihood for exponential families can be written as

$$l_c = T(X, Z)^T b(\theta) + \text{something that doesn't depend on } \theta$$

The linearity of expectation means you just replace the sufficient stats with their conditional expectations. If you weren't working with exponential families, you may not get something so nice for doing EM. eg.

$$z \sim \text{unif}(0, \theta)$$

$$x \sim \text{Normal}(z, \sigma^2)$$

2 Factorial HMM

Now we will do some variational inference on the factorial HMM model. (Draw figure)

Suppose we have independent markov chains $S_t^{(1)}, S_t^{(2)}, \dots, S_t^{(m)}$. These chains generate observations $Y_t \sim F(S_t^{(1)}, S_t^{(2)}, \dots, S_t^{(m)})$ ie. the distribution of Y_t depends on only the states of the variables $S_t^{(1)}, S_t^{(2)}, \dots, S_t^{(m)}$.

Why do we want to do this rather than make it one big markov chain? The transition matrix contains $K \times (K - 1) \approx K^2$ parameters ($K - 1$ to enforce the sum to 1 constraint). If you have independent markov chains you can get $k_1 k_2 \dots k_m = K$ states with only $k_1^2 + k_2^2 + \dots k_m^2 \ll k_1^2 k_2^2 \dots k_m^2$ parameters. You need a reasonable number of data points to estimate each parameter. So doing the big markov chain means you probably don't have enough data points. Plus, if the assumed independence relations are correct, you're estimates should be better.

You reduce the number of parameters you need in the model.

If we moralize in order to run sum-product, we get a bunch of m -cliques, and it becomes one big markov chain with many states. If the number of states is too large, then it'll be computationally intractable to run sum-product. In particular, it won't be feasible to do the E-step in EM.

One method is to use MCMC to estimate the E-step. Here we'll use a variational method. Unlike other cases you've seen, we don't assume that the q distribution completely factorizes into independent chunks. However, computation is still easy.

Draw pics:

completely factorized = no edges between any nodes for latent S variables,

structured = independent markov chains for each $S^{(m)}$

Keep the initial distribution and transition matrices the same. However, add an additional multinomial parameter $h_t^{(m)}$ for each node (like in the completely factorized case)

So what we're doing is removing the dependence among the chains which is induced by the observation y . We now need to find the h 's that minimize KL-divergence between the true joint p and our surrogate q .

(I omit the calculations in these notes. The important steps can be obtained from the paper *Factorial Hidden Markov Models* by Ghahramani and Jordan. see Appendix D)

The KL-divergence is

$$\begin{aligned}
 KL &= \sum_i \sum_m \langle S_t^{(m)} \rangle \log h_t^{(m)} + \frac{1}{2} \sum_t \left[Y_t^T C^{-1} Y_t - 2 \sum_m Y_t^T C^{-1} W^{(m)} \langle S_t^{(m)} \rangle \right. \\
 &+ \sum_m \sum_{n \neq m} \text{tr} \left(W^{(m)T} C^{-1} W^{(n)} \langle S_t^{(m)} \rangle \langle S_t^{(n)} \rangle \right) \\
 &\left. + \sum_m \text{tr} \left(W^{(m)T} C^{-1} W^{(m)} \text{diag}(\langle S_t^{(m)} \rangle) \right) \right] - \log Z_q + \log Z
 \end{aligned}$$

What is worth noting in the calculations is that when calculating the KL-divergence, many terms cancel out because we kept the same transition matrices $P^{(m)}$ and initial distributions $\pi^{(m)}$ in the variational approximation. We also get matching terms (1) $h_t^{(m)}$ and (2) something from the exponent of a normal. We compute $\langle S_t^{(m)} \rangle$ using the α, β algorithm.

Taking the partial derivatives w.r.t. $\log h_t^{(m)}$ and setting to 0 gives us the parameters for the q distribution we use as our approximation (again refer to the paper for the exact update). Note that we take advantage of the exponential family form to get the derivative $\frac{\partial \log Z_q}{\partial \log h_t^{(m)}} = \langle S_t^{(m)} \rangle$. $\log Z$ drops out after taking the

derivative.

2.1 q distribution for EM and mean field

- In EM the “best” q distribution is always $p(z|x, \theta)$.
- but this distribution/expectations under this distribution may be intractable to compute
- The bound $l(\theta; x) \geq \mathcal{L}(q, \theta)$ holds for *all* q though.
- Instead of the best surrogate $\mathcal{L}(p(z|x, \theta^{(old)}), \theta)$, we use the “best” *computable* surrogate $\mathcal{L}(q(z|h), \theta)$. In the context of EM, we consider E-step: $q^{(t+1)} = \arg \max_{q \in C} \mathcal{L}(q, \theta^{(t)})$ where C is our computable class of distributions.
- In this case, we choose the best out of the class of independent markov chain models with an additional node potential $h_t^{(m)}$.
- Note EM and variational methods are complementary rather than mutually exclusive methods.
- Note that both the E-step and the step to find best variational approximation use α, β . The E-step will require computing an additional statistic, namely $\langle S_{t-1}^{(m)} S_t^{(m)} \rangle$.

3 Slice sampling

Draw a bunch of figures giving more intuition.

Note that we don’t need to exactly find the set $A = \{x : f_i(x) > w_i \forall i\} = \cap_i \{x : f_i(x) > w_i\} = \cap_i A_i$. We just need to find something bigger $B \supset A$. The sample uniformly from the bigger set B and check if the sample belongs in the smaller set A . We can often shrink the set B whenever we find something outside of A . eg. if f_i is unimodal, then we look at whether $x_{new} >$ or $< x_{old}$ and shrink on the appropriate side.