Lecture 11:

Iterative Proportional Fitting

Sam Roweis

February 10, 2003

Undirected Models

- In directed models, we can think of each node plus its parents as a clique, and the clique potential as $p(x_i|x_{\pi_i})$.
- In an undirected model, we could have the same cliques, but with arbitrary postive potentials $\psi(x_i, x_{\pi_i})$.
- But this flexibility comes at a cost: learning in fully observed directed models is trivial, but learning in fully observed undirected models is not.
- Reason? The normalization factor Z:

$$\mathsf{P}(\mathbf{X}) = \frac{1}{Z} \prod_{\text{cliques } c} \psi_c(\mathbf{x}_c) \qquad Z = \sum_{\mathbf{X}} \prod_{\text{cliques } c} \psi_c(\mathbf{x}_c)$$

Which cliques?

- In general, our undirected model can place potentials on any subset of the cliques of the graph.
- Not all cliques need to be have potentials, and not all potentials need to be over maximal cliques.
- But in general, placing one potential on each *maximal clique* of the graph can represent *any* set of potentials.

Discrete Nodes

- Consider undirected models with all discrete nodes.
- Now there are just a finite number of possible joint settings of $\mathbf{X}$, and our model is defined as:

$$\mathsf{P}(\mathbf{x}|\theta) = \frac{1}{Z(\theta)} \prod_c \psi_c(\mathbf{x}_c) \qquad Z(\theta) = \sum_{\mathbf{x}} \prod_c \psi_c(\mathbf{x}_c)$$

  where $\theta = \{\psi_c(\mathbf{x}_c)\}$.
- For a particular dataset $\mathcal{D} = \{\mathbf{x}^1, \mathbf{x}^2, \ldots, \mathbf{x}^N\}$, we can count the number of times any joint configuration has been observed:

$$n(\mathbf{x}) = \sum_n \delta(\mathbf{x}, \mathbf{x}^n)$$

- We can also count the number of times a clique configuration appears:

$$n(\mathbf{x}_c) = \sum_n \delta(\mathbf{x}_c, \mathbf{x}_c^n) = \sum_{\mathbf{x}_{\tilde{c}}} n([\mathbf{x}_c, \mathbf{x}_{\tilde{c}}])$$

Log Likelihood

- In terms of the counts, the log likelihood is given by:

$$p(\mathcal{D}|\theta) = \prod_n \prod_{\mathbf{x}} p(\mathbf{x}|\theta)^{\delta(\mathbf{x},\mathbf{x}^n)}$$

$$\log p(\mathcal{D}|\theta) = \sum_n \sum_{\mathbf{x}} \delta(\mathbf{x},\mathbf{x}^n) \log p(\mathbf{x}|\theta)$$

$$\ell = \sum_{\mathbf{x}} n(\mathbf{x}) \log \left(\frac{1}{Z} \prod_c \psi_c(\mathbf{x}_c) \right)$$

$$= \sum_c \sum_{\mathbf{x}_c} n(\mathbf{x}_c) \log \psi_c(\mathbf{x}_c) - N \log Z$$

- So the clique counts $n(\mathbf{x}_c)$ are the sufficient statistics for our undirected model.
- But now there is a nasty $\log Z$ in the likelihood.

Iterative Proportional Fitting (IPF)

- Let's go back to the derivative of the likelihood:

$$\frac{\partial \ell}{\partial \psi_c(\mathbf{x}_c)} = \frac{n(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} - N \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$

- From this we can derive another relationship:

$$\frac{n(\mathbf{x}_c)}{N\psi_c(\mathbf{x}_c)} = \frac{q(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} = \frac{p(\mathbf{x}_c|\theta)}{\psi_c(\mathbf{x}_c)}$$

  in which ψ_c appears implicitly in the model marginal $p_c(\mathbf{x}_c|\theta)$.
- To solve for ψ_c is hard, because it appears on both sides of this implicit nonlinear equation.
- The idea of IPF is to hold ψ_c fixed on the right hand side (both in the numberator and denominator) and solve for it on the left hand side. We cycle through all cliques, then iterate:

$$\psi_c^{(t+1)}(\mathbf{x}_c) = \psi_c^{(t)}(\mathbf{x}_c) \frac{q(\mathbf{x}_c)}{p^{(t)}(\mathbf{x}_c)}$$

Conditions on Clique Marginals

- Let's calculate the derivative of the log likelihood with respect to the value of one clique potential at one setting of the clique variables. Next, we set this derivative to zero, trying to find the optimal parameters:

$$\frac{\partial \ell}{\partial \psi_c(\mathbf{x}_c)} = \frac{n(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} - N \frac{p(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)}$$

- Thus, for the maximum likelihood parameters, we know that:

$$p^*_{ML}(\mathbf{x}_c) = \frac{n(\mathbf{x}_c)}{N} = q(\mathbf{x}_c)$$

  In other words, at the maximum likelihood setting of the parameters, for each clique, *the model marginals must be equal to the observed marginals*.
- This doesn't tell us how to get the ML parameters, it just gives us a condition that must be satisfied when we have them.

Properties of IPF Updates

- The IPF updates have two amazing properties:
  1. At each iteration the model marginal $p^{(t+1)}(\mathbf{x}_c)$ is equal to the observed marginal $q(\mathbf{x}_c)$.
  2. The partition function Z remains constant across all IPF updates.
- To show this, calculate the marginal:

$$p^{(t+1)}(\mathbf{x}_c) = \sum_{\mathbf{x}_{\tilde{c}}} p^{(t+1)}(\mathbf{x})$$

$$= \frac{Z^{(t)}}{Z^{(t+1)}} q(\mathbf{x}_c)$$

  Now sum both sides over $\mathbf{x}_c$, yielding:

$$1 \cdot Z^{(t+1)} = Z^{(t)} \cdot 1$$

  and thus

$$p^{(t+1)}(\mathbf{x}_c) = q(\mathbf{x}_c)$$

IPF as Coordinate Ascent

- We saw that the IPF iterations always achieve the conditions we require at the ML parameter estimates.
- But how do we know if they will converge, and if so, if they always increase the log likelihood?
- We can show that IPF is actually *coordinate ascent* in the log likelihood, just like EM was.
- At any iteration, the derivative of the likelihood can be written as:

$$\frac{\partial \ell}{\partial \psi_c(\mathbf{x}_c)} = \frac{n(\mathbf{x}_c)}{\psi_c(\mathbf{x}_c)} - N\frac{p^{(t)}(\mathbf{x}_c)}{\psi_c^{(t)}(\mathbf{x}_c)}$$

  and if we evaluate this at $\psi_c^{(t+1)}(\mathbf{x}_c)$ we find that the gradient is in fact equal to zero.

KL Divergence View

- IPF can also be seen to be coordinate ascent in the likelihood using a trick of expressing likelihoods using KL divergences (cross entropies).
- First, we observe that maximizing the log likelihood is equivalent to minimizing the KL divergence from the observed distribution to the model distribution:

$$\max \ell \Leftrightarrow \min KL[q(\mathbf{x})\|p(\mathbf{x}|\theta)] = \sum_{\mathbf{x}} q(\mathbf{x}) \log \frac{q(\mathbf{x})}{p(\mathbf{x}|\theta)}$$

- Next, we use a property of KL divergence based on the conditional chain rule: $p(\mathbf{x}) = p(\mathbf{x}_a)p(\mathbf{x}_b|\mathbf{x}_a)$:

$$KL[q(\mathbf{x}_a, \mathbf{x}_b)\|p(\mathbf{x}_a, \mathbf{x}_b)] = KL[q(\mathbf{x}_a)\|p(\mathbf{x}_a)] + \sum_{\mathbf{x}_a} q(\mathbf{x}_a) KL[q(\mathbf{x}_b|\mathbf{x}_a)\|p(\mathbf{x}_b|\mathbf{x}_a)]$$

IPF minimizes KL divergence

- Putting these two together, we see that:

$$KL[q(\mathbf{x})\|p(\mathbf{x}|\theta)] = KL[q(\mathbf{x}_c)\|p(\mathbf{x}_c|\theta)] + \sum_{\mathbf{x}_c} q(\mathbf{x}_c) KL[q(\mathbf{x}_{\tilde{c}}|\mathbf{x}_c)\|p(\mathbf{x}_{\tilde{c}}|\mathbf{x}_c, \theta)]$$

  But changing the clique potential has no effect on the conditional distribution, so the second term in unaffected. To minimize the first term, we set the marginal to the observed marginal, just as in IPF.
- In fact, we can interpret IPF updates as retaining the "old" conditional probabilities $p^{(t)}(\mathbf{x}_{\tilde{c}}|\mathbf{x}_c)$ while replacing the "old" marginal probability $p^{(t)}(\mathbf{x}_c)$ with the observed marginal $q(\mathbf{x}_c)$.