CS691GM: Graphical Models - Spring 2012

Supplemental Note on Converting CRFs from Log-Linear to Standard Form

1 CRF Factor Reduction

The probabilistic model for the CRF is given below. The CRF model contains one feature parameter W^F_{cfv} for each of the C class labels, F features and feature values v (note that different features f have different numbers of values as given by the set V_f). The CRF also contains one transition parameter $W^T_{cc'}$ for each pair of labels c and c'.

$$P_W(\mathbf{y}_i|\mathbf{x}_i) = \frac{\exp\left(\sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F}\sum_{v\in V_f} W^F_{cfv}[y_{ij}=c][x_{ijf}=v] + \sum_{j=1}^{L_i-1}\sum_{c=1}^{C}\sum_{c'=1}^{C} W^T_{cc'}[y_{ij}=c][y_{ij+1}=c']\right)}{\sum_{\mathbf{y}'_i}\exp\left(\sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F}\sum_{v\in V_f} W^F_{cfv}[y'_{ij}=c][x_{ijf}=v] + \sum_{j=1}^{L_i-1}\sum_{c=1}^{C}\sum_{c'=1}^{C} W^T_{cc'}[y'_{ij}=c][y'_{ij+1}=c']\right)}$$

This form of the CRF model represents the conditional probability of the label variables given the feature variables and includes the explicit partition function in the denominator. This form is used to derive the conditional log likelihood of a data set and the derivatives of the conditional log likelihood needed for learning.

However, as we saw with the Ising model, learning requires marginal probabilities that must be computed using the sum-product algorithm. To apply the sum product algorithm, we need to first convert this form of the CRF model into the standard factor representation. To convert the above form of the CRF model to standard form, we first remove the explicit partition function. We can do this because the sum-product algorithm operates over the un-normalized product of factors. We denote the unnormalized distribution by $\tilde{P}_W(\mathbf{y}_i|\mathbf{x}_i)$.

$$\tilde{P}_W(\mathbf{y}_i|\mathbf{x}_i) = \exp\left(\sum_{j=1}^{L_i}\sum_{c=1}^{C}\sum_{f=1}^{F}\sum_{v\in V_f} W^F_{cfv}[y_{ij}=c][x_{ijf}=v] + \sum_{j=1}^{L_i-1}\sum_{c=1}^{C}\sum_{c'=1}^{C} W^T_{cc'}[y_{ij}=c][y_{ij+1}=c']\right)$$

Now, with the feature variables x_{ijf} fixed specific feature values, the CRF model has two types of reduced factors. Node factors involve a single label variable Y_{ij}, and pairwise factors involve pairs of label variables Y_{ij}, Y_{ij+1}. To derive the factor tables for the two types of factors, we first use the fact that the $\exp$ of a sum is a product of $\exp$'s.

$$\tilde{P}_W(\mathbf{y}_i|\mathbf{x}_i) = \prod_{j=1}^{L_i}\prod_{c=1}^{C}\exp\left(\sum_{f=1}^{F}\sum_{v\in V_f} W^F_{cfv}[y_{ij}=c][x_{ijf}=v]\right) \cdot \prod_{j=1}^{L_i-1}\prod_{c=1}^{C}\prod_{c'=1}^{C}\exp\left(W^T_{cc'}[y_{ij}=c][y_{ij+1}=c']\right)$$

Now we can use the fact that $\exp(W\cdot 0) = \exp(W)^0$ and $\exp(W\cdot 1) = \exp(W)^1$ to move the indicator functions over the Y variables outside of the $\exp$ terms. This is valid because the indicator functions (or product of indicator functions) always evaluate to either 0 or 1.

$$\tilde{P}_W(\mathbf{y}_i|\mathbf{x}_i) = \prod_{j=1}^{L_i}\prod_{c=1}^{C}\exp\left(\sum_{f=1}^{F}\sum_{v\in V_f} W^F_{cfv}[x_{ijf}=v]\right)^{[y_{ij}=c]} \cdot \prod_{j=1}^{L_i-1}\prod_{c=1}^{C}\prod_{c'=1}^{C}\left(W^T_{cc'}\right)^{[y_{ij}=c][y_{ij+1}=c']}$$

Now we can identify the values for each of each of the factor tables, as shown below.

$$\phi_{ijj+1}(c,c') = \exp(W^T_{cc'})$$

$$\begin{aligned}\phi_{ij}(c) &= \exp\left(\sum_{f=1}^{F}\sum_{v\in V_f} W^F_{cfv}[x_{ijf}=v]\right) = \exp\left(\sum_{f=1}^{F} W^F_{cfx_{ijf}}\right)\\ &= \exp\left(W^F_{c1x_{ij1}} + W^F_{c2x_{ij2}} + W^F_{c3x_{ij3}} + W^F_{c4x_{ij4}} + W^F_{c5x_{ij5}}\right)\end{aligned}$$

Each pairwise factor table value $\phi_{ijj+1}(c,c')$ is clearly a single non-negative real value. With the feature variables fixed to specific values, each node factor table value $\phi_{ij}(c)$ is also clearly a single non-negative real value. Lastly, we can re-represent the unnormalized model distribution in standard form using these factors:

$$\begin{aligned}\tilde{P}_W(\mathbf{y}_i|\mathbf{x}_i) &= \prod_{j=1}^{L_i}\prod_{c=1}^{C}\phi_{ij}(c)^{[y_{ij}=c]} \cdot \prod_{j=1}^{L_i-1}\prod_{c=1}^{C}\prod_{c'=1}^{C}\phi_{ijj+1}(c,c')^{[y_{ij}=c][y_{ij+1}=c']}\\ &= \prod_{j=1}^{L_i}\phi_{ij}(y_{ij}) \cdot \prod_{j=1}^{L_i-1}\phi_{ijj+1}(y_{ij},y_{ij+1})\end{aligned}$$

Note that this gives the standard-form of the reduced Markov network for a **single sentence** i **only**. Once the factor reduction and conversion to standard form has been performed as above, we can find the clique tree for sentence i (exactly as in the warm-up section of the assignment) and then run the sum-product algorithm over the resulting clique tree. This will compute the marginals for sentence i only. However,

now that we have the reduced factors represented in standard form, it's easy to see how to obtain the reduced factors in standard form for any sentence i. We simply plug the feature values for each token j in sentence i into the general form for the node factor table as given above. The pairwise factor is always the same for each pair of adjacent label variables in any sentence i. Also note that during learning, the parameters of the model W will change on each iteration. Thus, we must re-compute the values of the factor tables using the new W values, re-compute the clique-tree factors, and re-run the sum-product algorithm for each sentence to obtain the marginals corresponding to the updated parameters for each sentence.