The Conditional Random Field

Jason D. M. Rennie
jrennie@csail.mit.edu

January 21, 2005*

1 Preliminaries

Let $\vec{y} \in L^{n+1}$, $L = \{1, \ldots, l\}$, be a sequence of labels. Let $X \in \mathbb{R}^{n \times d}$ be a sequence of observations, each row corresponding to a consecutive pair of labels. Define a weight array, $W \in \mathbb{R}^{d \times l \times l}$, such that each "row" corresponds to a consecutive pair of labels. The Conditional Random Field assigns the following conditional negative log-likelihood,

$$-\log P(\vec{y}|X;W) = \log Z(X,W) - \sum_{i=1}^{n} X_i W_{(\cdot,y_i,y_{i+1})} + \frac{\lambda}{2}\|W\|^2_{\text{Fro}}, \tag{1}$$

where "Fro" denotes the Frobenius norm. The normalization constant is

$$Z(X,W) = \sum_{\vec{s} \in L^{n+1}} \exp\left(\sum_i X_i W_{(\cdot,s_i,s_{i+1})}\right). \tag{2}$$

Note that due to the structure of the problem, we do not need to sum over all possible labelings in order to calculate the normalization constant. We can sum over each position in turn,

$$Z(X,W) = \sum_{s_n} \cdots \left(\sum_{s_2}\left(\sum_{s_1} \exp(X_1 W_{((\cdot,s_1,s_2))}\right) \exp(X_2 W_{(\cdot,s_2,s_3)})\right) \cdots \exp(X_n W_{(\cdot,s_n,s_{n+1})}). \tag{3}$$

2 Learning

The partial derivatives take the usual form, expected feature values minus empirical feature values.

$$-\frac{\partial \log P}{\partial W_{juv}} = \sum_{i=1}^{n} \sum_{\vec{s} \in L^{n+1} | w_i = u, w_{i+1} = v} X_{ij} P(\vec{y}|X,W) - \sum_{i | y_i = u, y_{i+1} = v} X_{ij} + \lambda W_{juv}. \tag{4}$$

*Revised January 25, 2005.

To write this more compactly, define a transition matrix $M = \exp(XW)$, $M \in \mathbb{R}^{n\times l\times l}$. M_i is the (unnormalized) transition probability matrix, from $t = i$ to $t = i + 1$. We calculate forward and backward probabilities. Define

$$\vec{\alpha}_1 = \vec{1}, \qquad \vec{\alpha}_{i+1} = \vec{\alpha}_i^T M_i \quad \text{for } i \in \{2, \dots, n+1\}. \tag{5}$$

Note that $Z(X, W) = \vec{\alpha}_{n+1}^T \vec{1}$. The backwards probabilities are defined similarly,

$$\vec{\beta}_{n+1} = \vec{1}, \qquad \vec{\beta}_i = M_i \vec{\beta}_{i+1} \quad \text{for } i \in \{1, \dots, n\}. \tag{6}$$

Note that $Z(X, W) = \vec{1}^T \vec{\beta}_1$. Define

$$Y \in \{0, 1\}^{n\times l\times l}, \qquad Y_{iuv} = \begin{cases} 1 & \text{if } y_i = u, y_{i+1} = v \\ 0 & \text{othw.} \end{cases}. \tag{7}$$

Define

$$C \in \mathbb{R}^{n\times l\times l}, \qquad C_i = (\vec{\alpha}_i \vec{\beta}_{i+1}^T) * M_i, \tag{8}$$

where $*$ is element-wise product. Note that C_{iuv} is the (unnormalized) model probability assigned to all sequences, $\vec{s}$, with $s_i = u$, $s_{i+1} = v$. Now we can write the derivative more compactly,

$$-\frac{\partial \log P}{\partial W} = X^T C - X^T Y + \lambda W, \tag{9}$$

The gradient can be used in conjunction with Conjugate Gradients or L-BFGS to solve for the optimal weight matrix.

3 Inference

Define the backward max-probabilities:

$$\overline{\beta}_{(n+1,j)} = 1, \forall j \qquad \overline{\beta}_{ij} = \max_k M_{ijk} \overline{\beta}_{(i+1,k)}. \tag{10}$$

$\overline{\beta}_{ij}$ is the (unnormalized) probability of the max-likelihood sequence ($t = i$ to $t = n + 1$) beginning in state j. Then, the max-likelihood sequence is $\vec{s}^*$, where $s_1^* = \max_k \overline{\beta}_{1k}$ and

$$s_{i+1}^* = \max_j M_{(i,s_i^*,j)} \overline{\beta}_{(i+1,j)} \quad \text{for } i \in \{1, \dots, n\}. \tag{11}$$

References

Kakade, S., Teh, Y. W., & Roweis, S. (2002). An alternate objective function for markovian fields. *Proceedings of the Nineteenth International Conference on Machine Learning.*

Lafferty, J., McCallum, A., & Pereira, F. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. *Proceedings of the Eighteenth International Conference on Machine Learning*.

Sha, F., & Pereira, F. (2003). Shallow parsing with conditional random fields. *Proceedings of HLT-NAACL*.

Wallach, H. M. (2003). Efficient training of conditional random fields. *Proceedings of the 6th Annual CLUK Research Colloquium*.