CSE 5523: Lecture Notes 21
Markov Random Fields

Contents

Like backprop and Bayes nets, matrix chains can also do inference in structured prediction models.

These include **Markov random fields (MRFs)**, **conditional random fields (CRFs)**, and others.

Generally, these are like Bayes nets but are undirected, defined on overlapping joints over variables.

They assign weights, called **potentials**, to overlapping tuples of variables, called **cliques**.

Here cliques are functions $f(y, y', ...)_{[i]} \in \{0, 1\}$, potentials are weights $\mathbf{w}_{[i]}$ for patterns i of values.

21.1 (Conditional) Markov Random Fields

A common structured prediction models is the linear chain conditional random field.

It consists of a sequence of hidden values y_t and observed values x_t, for time steps $t \in \{1, ..., T\}$:

$$\mathsf{P}(y_{1..T} \mid x_{1..T}) = \frac{e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y_t, y_{t-1})}}{\sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})}}$$

The parameters are set like logistic regression:

$$\begin{aligned}
\frac{\partial C}{\partial \mathbf{w}_{[i]}} &= \frac{\partial}{\partial \mathbf{w}_{[i]}} \frac{1}{N} \sum_{\langle y_{1..T}, x_{1..T} \rangle \in \mathcal{D}} -\ln \frac{e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y_t, y_{t-1})}}{\sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})}} \\
&= \frac{1}{N} \sum_{\langle y_{1..T}, x_{1..T} \rangle \in \mathcal{D}} \frac{\partial}{\partial \mathbf{w}_{[i]}} -\ln \frac{e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y_t, y_{t-1})}}{\sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})}} && \text{sum, prod. rule} \\
&= \frac{1}{N} \sum_{\langle y_{1..T}, x_{1..T} \rangle \in \mathcal{D}} \frac{\partial}{\partial \mathbf{w}_{[i]}} -\ln e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y_t, y_{t-1})} + \ln \sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})} && \text{log of fract.} \\
&= \frac{1}{N} \sum_{\langle y_{1..T}, x_{1..T} \rangle \in \mathcal{D}} \frac{\partial}{\partial \mathbf{w}_{[i]}} -\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y_t, y_{t-1}) + \ln \sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})} && \text{log of exp.} \\
&= \frac{1}{N} \sum_{\langle y_{1..T}, x_{1..T} \rangle \in \mathcal{D}} -\sum_{t=1}^{T} f(x_t, y_t, y_{t-1})_{[i]} + \frac{\partial}{\partial \mathbf{w}_{[i]}} \ln \sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})} && \text{sum, prod. rule}
\end{aligned}$$

$$= \frac{1}{N} \sum_{\langle y_{1..T}, x_{1..T} \rangle \in \mathcal{D}} - \sum_{t=1}^{T} f(x_t, y_t, y_{t-1})_{[i]} + \frac{1}{\sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})}} \frac{\partial}{\partial \mathbf{w}_{[i]}} \sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})} \quad \text{deriv. log}$$

$$= \frac{1}{N} \sum_{\langle y_{1..T}, x_{1..T} \rangle \in \mathcal{D}} - \sum_{t=1}^{T} f(x_t, y_t, y_{t-1})_{[i]} + \frac{1}{\sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})}} \sum_{y'_{1..T} \in Y^T} \frac{\partial}{\partial \mathbf{w}_{[i]}} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})} \quad \text{sum rule}$$

$$= \frac{1}{N} \sum_{\langle y_{1..T}, x_{1..T} \rangle \in \mathcal{D}} - \sum_{t=1}^{T} f(x_t, y_t, y_{t-1})_{[i]} + \frac{\sum_{y'_{1..T} \in Y^T} \overbrace{e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})}}^{\text{pattern weight}} \overbrace{\sum_{t=1}^{T} f(x_t, y'_t, y'_{t-1})_{[i]}}^{\text{number of occurrences of pattern}}}{\sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})}} \quad \text{sum, prod. rule}$$

21.2 Message passing for random fields

The denominator above sums over many hidden sequences, but is tractable via message passing:

$$\sum_{y_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y_t, y_{t-1})} = \sum_{y_{1..T} \in Y^T} \prod_{t=1}^{T} e^{\mathbf{w}^\top f(x_t, y_t, y_{t-1})} \quad \text{exponentiation of sum}$$

$$= \sum_{y_T \in Y} \left(\sum_{y_{1..T-1} \in Y^{Y-1}} \prod_{t=1}^{T} e^{\mathbf{w}^\top f(x_t, y_t, y_{t-1})} \right) \quad \text{associative axiom}$$

$$= \underbrace{\sum_{y_T \in Y} \left(\sum_{y_{1..T-1} \in Y^{Y-1}} \prod_{t=1}^{T} e^{\mathbf{w}^\top f(x_t, y_t, y_{t-1})} \right) \delta_{y_T}^\top \mathbf{1}}_{\text{(un-normalized) forward message } \mathbf{f}_T^\top} \quad \text{def. inner product}$$

$$= \sum_{y_T \in Y} \sum_{y_{T-1} \in Y} \left(\sum_{y_{1..T-2} \in Y^{T-2}} \prod_{t=1}^{T-1} e^{\mathbf{w}^\top f(x_t, y_t, y_{t-1})} \right) e^{\mathbf{w}^\top f(x_T, y_T, y_{T-1})} \delta_{y_T}^\top \mathbf{1} \quad \text{distributive axiom}$$

$$= \underbrace{\sum_{y_{T-1} \in Y} \left(\sum_{y_{1..T-2} \in Y^{T-2}} \prod_{t=1}^{T-1} e^{\mathbf{w}^\top f(x_t, y_t, y_{t-1})} \right) \delta_{y_{T-1}}^\top \mathbf{M}_T}_{\text{(un-normalized) forward message } \mathbf{f}_{T-1}^\top} \mathbf{1} \quad \text{def. inner product}$$

where $\mathbf{M}_t = \sum_{y_{t-1}, y_t} \delta_{y_{t-1}} e^{\mathbf{w}^\top f(x_t, y_t, y_{t-1})} \delta_{y_t}^\top$.

These last two steps can be repeated to define a matrix chain.

The numerator $\sum_{y'_{1..T} \in Y^T} e^{\sum_{t=1}^{T} \mathbf{w}^\top f(x_t, y'_t, y'_{t-1})} \sum_{t=1}^{T} f(x_t, y'_t, y'_{t-1})_{[i]}$ is expected count of feature i in chain.

This can be estimated by multiplying each matrix in the chain by the filter $\delta_{y'_{t-1}} f(x_t, y'_t, y'_{t-1})_{[i]} \delta_{y'_t}^\top$.

(It's efficient using forward $\mathbf{f}_t^\top = \mathbf{1}^\top \mathbf{M}_1 \ldots \mathbf{M}_t$ and backward messages $\mathbf{b}_t = \mathbf{M}_t \ldots \mathbf{M}_T \mathbf{1}$.)

The result is the expected count of feature i at each time step, which we sum to get the numerator.

This generalizes to other (singly-connected) topologies as message passing.