TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2019

Exponential Softmax Backpropagation:

The Model Marginals

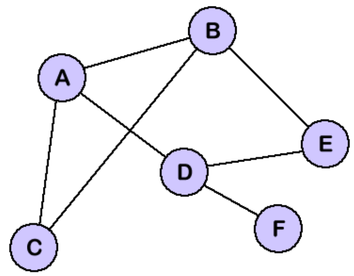
Exponential Softmax

$$\text{for } \hat{y} \quad s(\hat{y}) = \sum_i \; s_n[i, \hat{y}[i]] + \sum_{\langle i,\,j\rangle \in \text{Edges}} \; s_e[\langle i, j\rangle, \hat{y}[i],\; \hat{y}[j]]$$

$$\text{for } \hat{y} \;\; P_s(\hat{y}) = \text{softmax}_{\hat{y}} \; s(\hat{y}) \;\; \text{all possible } \hat{y}$$

$$\mathcal{L} = -\ln P_s(y) \quad \text{gold label (training label) } y$$

Exponential Softmax is Typically Intractable



$\hat{y}$ assigns a label $\hat{y}[i]$ to each node i.

$s(\hat{y})$ is defined by a sum over node and edge tensor scores.

$P_s(\hat{y})$ is defined by an exponential softmax over $s(\hat{y})$.

Computing Z in general is #P hard (there is an easy direct reduction from SAT).

Compactly Representing Scores on Exponentially Many Labels

The tensor $s_n[I, C]$ holds IC scores.

The tensor $s_e[E, C, C]$ holds EC^2 scores where e ranges over edges $\langle i, j \rangle \in \text{Edges}$.

Back-Propagation Through Exponential Softmax

$$\begin{aligned} s_n[I,C] &= f^n_\Phi(x) \\ s_e[E,C,C] &= f^e_\Phi(x) \end{aligned}$$

$$s(\hat{y}) = \sum_i s_n[i,\hat{y}[i]] + \sum_{\langle i,j\rangle \in \mathrm{Edges}} s_e[\langle i,j\rangle, \hat{y}[i],\ \hat{y}[j]]$$

$$P_s(\hat{y}) = \operatorname*{softmax}_{\hat{y}}\ s(\hat{y}) \quad \text{all possible } \hat{y}$$

$$\mathcal{L} = -\ln P_s(y) \quad \text{gold label } y$$

We want the gradients $s_n.\mathrm{grad}[I,C]$ and $s_e.\mathrm{grad}[E,C,C]$.

Model Marginals Theorem

Theorem:

$$s_n.\mathrm{grad}[i,c] \;=\; P_{\hat{y}\sim P_s}(\;\hat{y}[i]=c\;) \\ -\mathbf{1}[\;y[i]=c\;]$$

$$s_e.\mathrm{grad}[\langle i,j\rangle,c,c'] \;=\; P_{\hat{y}\sim P_s}(\;\hat{y}[i]=c\;\wedge\;\hat{y}[j]=c'\;) \\ -\mathbf{1}[\;y[i]=c\;\wedge\;y[j]=c'\;]$$

We need to compute (or approximate) the model marginals.

Proof of Model Marginals Theorem

We consider the case of node marginals, The case of edge marginals is similar.

$$
\begin{aligned}
s_n.\mathrm{grad}[i,c] &= \partial \mathcal{L}(\Phi, x, y) \;/\; \partial s_n[i,c] \\
&= \partial \left(-\ln \frac{1}{Z} \exp(s(y)) \right) \;/\; \partial s_n[i,c] \\
&= \partial (\ln Z - s(y)) \;/\; \partial s_n[i,c] \\
&= \left(\frac{1}{Z} \sum_{\hat{y}} e^{s(\hat{y})} \left(\partial s(\hat{y}) / \partial s_n[i,c] \right) \right) - \left(\partial s(y) / \partial s_b[i,c] \right)
\end{aligned}
$$

Proof of Model Marginals Theorem

$$s_n.\mathrm{grad}[i,c] \;=\; \left(\frac{1}{Z}\sum_{\hat{y}} e^{s(\hat{y})}\left(\partial s(\hat{y})/\partial s_n[i,c]\right)\right) - \left(\partial s(y)/\partial s_b[i,c]\right)$$

$$\;=\; \left(\sum_{\hat{y}} P_s(\hat{y})\left(\partial s(\hat{y})/\partial s_n[i,c]\right)\right) - \left(\partial s(y)/\partial s_n[i,c]\right)$$

$$s(\hat{y}) \;=\; \sum_i \; s_n[i,\hat{y}[i]] \;+\; \sum_{\langle i,j\rangle \in \mathrm{Edges}} \; s_e[\langle i,j\rangle,\hat{y}[i],\;\hat{y}[j]]$$

$$\frac{\partial s(\hat{y})}{\partial s_n[i,c]} \;=\; \mathbf{1}[\hat{y}[i]=c]$$

Proof of Model Marginals Theorem

$$s_n.\mathrm{grad}[i,c] = \left(\frac{1}{Z}\sum_{\hat{y}} e^{s(\hat{y})}\left(\partial s(\hat{y})/\partial s_n[i,c]\right)\right) - \left(\partial s(y)/\partial s_b[i,c]\right)$$

$$\left(\sum_{\hat{y}} P_s(\hat{y})\left(\partial s(\hat{y})/\partial s_n[i,c]\right)\right) - \left(\partial s(y)/\partial s_n[i,c]\right)$$

$$= E_{\hat{y}\sim P_s}\mathbf{1}[\hat{y}[i]=c] - \mathbf{1}[y[i]=c]$$

$$= P_{\hat{y}\sim P_s}(\hat{y}[i]=c) - \mathbf{1}[y[i]=c]$$

Model Marginals Theorem

Theorem:

$$
\begin{aligned}
s_n.\mathrm{grad}[i, c] \;=\; & P_{\hat{y} \sim P_s}(\;\; \hat{y}[i] = c \;\;) \\
& -\mathbf{1}[\;\; y[i] = c \;\;]
\end{aligned}
$$

$$
\begin{aligned}
s_e.\mathrm{grad}[\langle i, j\rangle, c, c'] \;=\; & P_{\hat{y} \sim P_s}(\;\; \hat{y}[i] = c \;\wedge\; \hat{y}[j] = c' \;\;) \\
& -\mathbf{1}[\;\; y[i] = c \;\wedge\; y[j] = c' \;\;]
\end{aligned}
$$

END