TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2018

Deep Graphical Models III

Expectation Maximization (EM)

Expected Gradient (EG)

CTC

Latent Variable Models

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{(x,y)\sim\mathrm{Pop}} - \ln Q_\Phi(y|x)$$

y ranges over a structured set such as sentences or images.

$$Q_\Phi(y|x) = \sum_{\hat{z}} Q_\Phi(\hat{z}, y \mid x)$$

$\hat{z}$ ranges over latent labels such as a word sense for each word or a semantic label for each pixel.

The Expected Gradient (EG) Identity

$$
\begin{aligned}
\nabla_\Phi \ \ln Q(y) &= \frac{\nabla_\Phi \ Q(y)}{Q(y)} \\[2em]
&= \sum_{\hat{z}} \ \frac{\nabla_\Phi \ Q(\hat{z}, y)}{Q(y)} \\[2em]
&= \sum_{\hat{z}} \ \frac{Q(\hat{z}, y) \ \nabla_\Phi \ \ln Q(\hat{z}, y)}{Q(y)} \\
&= E_{\hat{z} \sim Q(\hat{z}|y)} \nabla_\Phi \ln Q(\hat{z}, y)
\end{aligned}
$$

The EG Identity

$$\nabla_\Phi \; \ln Q_\Phi(y|x) = E_{\hat{z} \sim Q_\Phi(\hat{z}|x,y)} \nabla_\Phi \ln Q_\Phi(\hat{z}, y|x)$$

It is important to note that the gradient operation only appears inside the expectation.

This is an **expected gradient**.

Sufficient Statistics

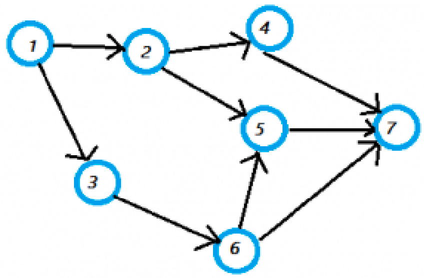
Consider a model $Q_\Phi(\hat{z},\hat{y}|x)$ and a tensor S computed from x, y and $\hat{z}$ such that

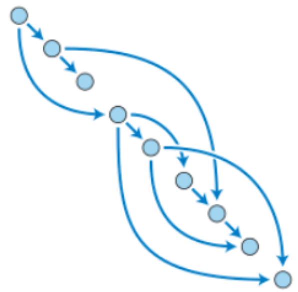
$$E_{\hat{z}\sim Q_\Phi(\hat{z}|x,y)}\;\nabla_\Phi \ln Q_\Phi(\hat{z},y|x) = f\left(E_{\hat{z}\sim Q_\Phi(\hat{z}|x,y)}\;S(x,y,\hat{z})\right)$$

When this equation holds we say that $E_{\hat{z}\sim Q_\Phi(\hat{z}|x,y)}\;S(x,y,\hat{z})$ is a **sufficient statistic** for the model $Q_\Phi(\hat{z},\hat{y}|x)$.

Even when $\hat{z}$ is a structured object (with exponentially many possible values) it is often possible to find a tractable-sized sufficient statistic that can be computed or estimated efficiently.

Example: Latent Variable Directed Graphical Models





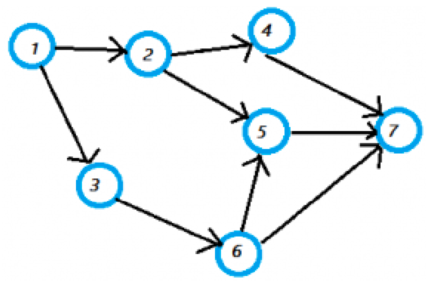
We assume a population of pairs (x, y) and a model $Q_\Phi(y|x)$ to be trained by cross-entropy.

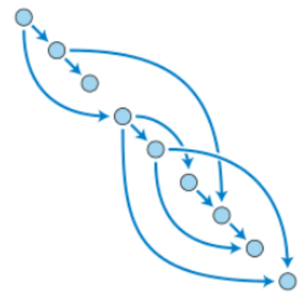
$$\Phi^* = \operatorname*{argmin}_{\Phi} \; -\ln Q_\Phi(y|x)$$

Here y is an assignment values to **observed** nodes.

We must marginalize over assignments to **unobserved** nodes.

Latent Variable Directed Graphical Models





Let $\hat{v}$ denote an assignment to all nodes — both observed and unobserved (latent).

Here we consider only directed models — models satisfying

$$Q_\Phi(\hat{v}|x) = \prod_i Q_\Phi(x)(\hat{v}[i] \;|\hat{v}[\mathrm{Parents}(i)])$$

Here $Q_\Phi(x)$ is a tensor giving the conditional probability tables $Q[\hat{v}[i] \;|\hat{v}[\mathrm{Parents}(i)]]$.

Observed and Latent Variables

Let $\hat{v}_y$ be the assignment $\hat{v}$ makes to the observed variables.

Let $\hat{v}_z$ be the assignment $\hat{v}$ makes to the unobserved variables.

Let Q abbreviate the tensor $Q_\Phi(x)$.

$$Q(\hat{v}) = Q(\hat{v}_z, \hat{v}_y)$$

$$Q(\hat{v}_y) = \sum_{\hat{v}_z} Q(\hat{v}_z, \hat{v}_y)$$

The Sufficient Statistics

For each node i the tensor $Q_\Phi(x)$ specifies the conditional probability table $Q[\tilde{v}_i|\tilde{v}_p]$ where $\tilde{v}_i$ ranges over the possible values of node i and $\tilde{v}_p$ ranges over the possible assignments of values to the parents of i.

$$Q.\mathrm{grad} = \nabla_Q \; - \ln \sum_{\tilde{v}_z} Q(\tilde{v}_z, \tilde{v}_y)$$

$$Q.\mathrm{grad}[\tilde{v}_i|\tilde{v}_p] = \frac{-\partial \ln \sum_{\tilde{v}_z} Q(\tilde{v}_z, \tilde{v}_y)}{\partial Q[\tilde{v}_i|\tilde{v}_p]}$$

The Sufficient Statistics

$$\begin{aligned}
\nabla_Q \ \ln \ Q(y) &= E_{\hat{z}\sim Q(\hat{z}|y)} \nabla_Q \ln Q(\hat{z}, y) \\
&= E_{\hat{z}\sim Q(\hat{z}|y)} \sum_i \nabla_Q \ln Q((\hat{z}, y)[i] \mid (\hat{z}, y)[\mathrm{parents}(i)]) \\
Q.\mathrm{grad}[\tilde{v}_i, \tilde{v}_p] &= E_{\hat{z}\sim Q(\hat{z}|y)} \frac{1}{Q[\tilde{v}_i, \tilde{v}_p]} \mathbb{1}[(\hat{z}, y)(i, \mathrm{Parents}(i)) = (\tilde{v}_i, \tilde{v}_p)] \\
&= \frac{1}{Q[\tilde{v}_i|\tilde{v}_p]} E_{\hat{z}\sim Q(\hat{z}|y)} \mathbb{1}[(\hat{z}, y)(i, \mathrm{Parents}(i)) = (\tilde{v}_i, \tilde{v}_p)]
\end{aligned}$$

The quantities $E_{\hat{z}\sim Q(\hat{z}|y)} \mathbb{1}[(\hat{z}, y)(i, \mathrm{Parents}(i)) = (\tilde{v}_i, \tilde{v}_p)]$ are the **sufficient statistics**.

Trees are Tractable

For tree models the sufficient statics can be computed efficiently by message passing (belief propagation).

Loopy BP can be used for non-tree models.

Expected Gradient (EG)
and Expectation Maximization (EM)

EG: $\nabla_Q \ln Q(y) = E_{\hat{z} \sim Q(\hat{z}|y)} \nabla_Q \ln Q(\hat{z}, y).$

EM: $Q^{t+1} = \mathrm{argmax}_Q \; E_{\hat{z} \sim Q^t(\hat{z}|y)} \ln Q(\hat{z}, y).$

EG $= \nabla_Q \ln Q(y)$ equals is the gradient of the EM objective.

EG and EM have **the same sufficient statistics** (E-step).

Connectionist Temporal Classification (CTC)

Phonetic Transcription

A speech signal

$$x = x_1,\ \ldots,\ x_T$$

is labeled with a phone sequence

$$y = y_1, \ldots, y_N$$

with $N << T$ and with $y_n \in \mathcal{Y}$ for a set of phonemes $\mathcal{Y}$.

The length N of y is not determined by x and the alignment between x and y is not given.

CTC

The model defines $Q_\Phi(\hat{z}|x, y)$ where $\hat{z}$ is latent.

$$\hat{z} = \hat{z}_1, \ \ldots, \ \hat{z}_T, \quad \hat{z}_t \in \mathcal{Y} \cup \{\perp\}$$

The sequence

$$y(\hat{z}) = y_1, \ \ldots, \ y_N$$

is the result of removing all the occurrences of $\perp$ from $\hat{z}$.

$$\perp, a_1, \perp, \perp, \perp, a_2, \perp, \perp, a_3, \perp \Rightarrow a_1, a_2, a_3$$

The CTC Model

$$h_1,\ \ldots,\ h_T = \mathrm{RNN}_\Phi(x_1,\ \ldots,\ x_T)$$

$$Q_\Phi(\hat{z}_t|x_1,\ldots,x_T) = \operatorname*{softmax}_{\hat{z}}\ e(\hat{z})^\top h_t$$

This is a locally normalized (directed) graphical model where $\hat{z}_t$ does not have any parent nodes.

The Sufficient Statistics

Since each node has no parents the sufficient statistics are

$$P_{\hat{z} \sim Q(\hat{z}|y)}(z_t = \tilde{z})$$

Dynamic Programming (Forward-Backward)

$$
\begin{aligned}
x &= x_1,\ \ldots,\ x_T \\
\hat{z} &= \hat{z}_1,\ \ldots,\ \hat{z}_T, \quad \hat{z}_t \in \mathcal{Y} \cup \{\perp\} \\
y &= y_1,\ \ldots,\ y_N, \quad\ y_n \in \mathcal{Y}, \quad N << T \\
y &= (\hat{z}_1,\ \ldots,\ \hat{z}_T) - \perp
\end{aligned}
$$

Forward-Backward

$$
\begin{aligned}
\vec{y}_t &= (\hat{z}_1,\ \ldots,\ \hat{z}_t) - \perp \\
F[n,t] &= Q(\vec{y}_t = y_1,\ \ldots,\ y_n) \\
B[n,t] &= Q(y_{n+1},\ \ldots, y_N | \vec{y}_t = y_1,\ \ldots,\ y_n)
\end{aligned}
$$

Dynamic Programming (Forward-Backward)

$$
\begin{aligned}
\vec{y}_t &= (\hat{z}_1,\ \ldots,\ \hat{z}_t) - \perp \\
F[n,t] &= Q(\vec{y}_t = y_1,\ \ldots,\ y_n) \\
B[n,t] &= Q(y_{n+1},\ \ldots, y_N | \vec{y}_t = y_1,\ \ldots,\ y_n)
\end{aligned}
$$

$$
\begin{aligned}
F[0,0] &= 1 \\
F[n,0] &= 0 \quad \text{for } n > 0 \\
F[n+1,t+1] &= Q(\hat{z}_{t+1} = \perp)F[n+1,t] + Q(\hat{z}_{t+1} = y_{n+1})F[n,t]
\end{aligned}
$$

$$
\begin{aligned}
B[N,T] &= 1 \\
B[n,T] &= 0, \quad \text{for } n < N \\
B[n-1,t-1] &= Q(\hat{z}_{t-1} = \perp)B[n-1,t] + Q(\hat{z}_{t-1} = y_{n-1})B[n,t]
\end{aligned}
$$

Latent Variable MRFs

$$Q_f(\hat{z}, \hat{y}) = \frac{1}{Z}\, e^{f(\hat{z},\hat{y})}$$

$$Q_f(\hat{y}) = \sum_{\hat{z}} Q_f(\hat{z}, \hat{y}) \quad = \quad \frac{\sum_{\hat{z}} e^{f(\hat{z},\hat{y})}}{Z}$$

$$Q_f(\hat{y}) = \frac{Z(\hat{y})}{Z}$$

$$\mathrm{loss}(y) = \ln Z - \ln Z(y)$$

The Sufficient Statistics

$$P_{\hat{z}\sim Q(\hat{z}|y)}(\hat{z}_t = \tilde{z})$$

$$= \frac{1}{Q(y)} \sum_n \begin{cases} F[n, t-1]\; Q(\hat{z}_t)\; B[n, t] & \text{for } \hat{z}_t = \perp \\ F[n, t-1]\; Q(\hat{z}_t)\; B[n+1, t] & \text{for } \hat{z}_t = y_{n+1} \\ 0 & \text{otherwise} \end{cases}$$

Undirected Latent Variable MRFs

$$\Phi^* = \operatorname*{argmin}_{\Phi} \;\; E_{(x,y)\sim\mathrm{Pop}} - \ln Q_{f_\Phi(x)}(y)$$

$$Q_f(\hat{z},\hat{y}) = \operatorname*{softmax}_{\hat{z},\hat{y}} \; f(\hat{z},\hat{y})$$

$$f(\hat{z},\hat{y}) = \sum_{\alpha} \; f[\alpha,\hat{z}[\alpha],\hat{y}[\alpha]]$$

$$Q_f(\hat{y}) = \sum_{\hat{z}} Q_f(\hat{z},\hat{y})$$

Latent Variable MRFs

$$\mathrm{loss}(y, f) = \ln Z - \ln Z(y)$$

$$\begin{aligned} f.\mathrm{grad}[\alpha, \tilde{y}, \tilde{z}] = & \; P_{\hat{z}, \hat{y} \sim Q_f}(\hat{y}[\alpha] = \tilde{y}, \hat{z}[\alpha] = \tilde{z}) \\ & - P_{\hat{z} \sim Q_f(\hat{z}|y)}(y[\alpha] = \tilde{y}, \hat{z}[\alpha] = \tilde{z}) \end{aligned}$$

These are the **sufficient statistics** for latent variable MRFs.

END