TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Pseudo-Likelihood and Contrastive Divergence

Notation

x is an input (e.g. an image).

$\hat{\mathcal{Y}}[N]$ is a structured label for x — a vector $\hat{\mathcal{Y}}[0], \ldots, \hat{\mathcal{Y}}[N-1]$. (e.g., n ranges over pixels where $\hat{\mathcal{Y}}[n]$ is a semantic label of pixel n.)

$\hat{\mathcal{Y}}/n$ is the set of labels assigned by $\hat{\mathcal{Y}}$ at indeces (pixels) other than n.

$\hat{\mathcal{Y}}[n = \ell]$ is the structured label identical to $\hat{\mathcal{Y}}$ except that it assigns label ℓ to index (pixel) n.

Intractable Exponential Softmax

We consider a softmax distribution

$$P_s(\hat{\mathcal{Y}}) = \frac{1}{Z} e^{s(\hat{\mathcal{Y}})}$$

$$Z = \sum_{\hat{\mathcal{Y}}} e^{s(\hat{\mathcal{Y}})}$$

Computing Z is intractable.

Psuedo-Likelihood

For any distribution $P(\hat{\mathcal{Y}})$ on structured labels $\hat{\mathcal{Y}}$, we define the pseudo-likelihood $\tilde{P}(\hat{\mathcal{Y}})$ as follows

$$\tilde{P}(\hat{\mathcal{Y}}) = \prod_n P(\hat{\mathcal{Y}}[n] \mid \hat{\mathcal{Y}}/n)$$

$$P(\hat{\mathcal{Y}}[n] \mid \hat{\mathcal{Y}}/n) = \frac{1}{Z_n} e^{s(\hat{\mathcal{Y}})} \qquad Z_n = \sum_{\ell} e^{s(\hat{\mathcal{Y}}[n=\ell])}$$

While computing $P_s(\hat{\mathcal{Y}})$ is intractable, computing $\tilde{P}_s(\hat{\mathcal{Y}})$ involves only local partition functions and is tractable.

Pseudo Cross-entropy Loss

We can then do SGD on pseudo cross-entropy loss.

$$\Phi^* = \operatorname*{argmin}_{\Phi} E_{\langle x, \mathcal{Y}\rangle \sim \mathrm{Pop}} \quad -\ln \tilde{P}_{\Phi,x}(\mathcal{Y})$$

Pseudolikelihood Theorem

$$\operatorname*{argmin}_{Q} \; E_{\mathcal{Y}\sim\mathrm{Pop}} \; -\ln \tilde{Q}(\mathcal{Y}) = \mathrm{Pop}$$

It suffices to show that for any Q we have

$$E_{\mathcal{Y}\sim\mathrm{Pop}} \; -\ln \widetilde{\mathrm{Pop}}(\mathcal{Y}) \leq \; E_{\mathcal{Y}\sim\mathrm{Pop}} \; -\ln \tilde{Q}(\mathcal{Y})$$

Proof II

$$
\begin{aligned}
& \min_{Q}\; E_{\mathcal{Y}\sim\mathrm{Pop}} - \ln \tilde{Q}(\mathcal{Y}) \\
= \;& \min_{Q}\; E_{\mathcal{Y}\sim\mathrm{Pop}} \sum_{n} - \ln Q(\mathcal{Y}[n] \mid \mathcal{Y}/n) \\
\geq \;& \min_{P_1,\ldots,P_N}\; E_{\mathcal{Y}\sim\mathrm{Pop}} \sum_{n} - \ln P_n(\mathcal{Y}[n] \mid \mathcal{Y}/n) \\
= \;& \min_{P_1,\ldots,P_N}\; \sum_{n} E_{\mathcal{Y}\sim\mathrm{Pop}} \;\; - \ln P_n(\mathcal{Y}[n] \mid \mathcal{Y}/n) \\
= \;& \sum_{n} \min_{P_n}\; E_{\mathcal{Y}\sim\mathrm{Pop}} \;\; - \ln P_n(\mathcal{Y}[n] \mid \mathcal{Y}/n) \\
= \;& \sum_{n} E_{\mathcal{Y}\sim\mathrm{Pop}} \;\; - \ln \mathrm{Pop}(\mathcal{Y}[n] \mid \mathcal{Y}/n) = E_{\mathcal{Y}\sim\mathrm{Pop}} - \ln \widetilde{\mathrm{Pop}}(\mathcal{Y})
\end{aligned}
$$

Contrastive Divergence (CDk)

In contrastive divergence we first construct an MCMC process whose stationary distribution is P_s. This could be Metropolis or Gibbs or something else.

Algorithm CDk: Given a gold segmentation $\mathcal{Y}$, start the MCMC process from initial state $\mathcal{Y}$ and run the process for k steps to get $\mathcal{Y}'$. Then take the loss to be

$$\mathcal{L}_{\mathrm{CD}} = s(\mathcal{Y}') - s(\mathcal{Y})$$

If $P_s = \mathrm{Pop}$ then the the distribution on $\mathcal{Y}'$ is the same as the distribution on $\mathcal{Y}$ and the expected loss gradient is zero.

Gibbs CD1

CD1 for the Gibbs MCMC process is a particularly interesting special case.

Algorithm (Gibbs CD1): Given $\mathcal{Y}$, select a node n at random and draw $\ell \sim P(\mathcal{Y}[n] = \ell \mid \mathcal{Y}/n)$. Define $\mathcal{Y}[n = \ell]$ to be the assignment (segmentation) which is the same as $\mathcal{Y}$ except that node n is assigned label ℓ. Take the loss to be

$$\mathcal{L}_{\mathrm{CD}} = s(\mathcal{Y}[n = \ell]) - s(\mathcal{Y})$$

Gibbs CD1 Theorem

Gibbs CD1 is equivalent in expectation to pseudolikelihood.

$$
\begin{aligned}
\mathcal{L}_{\mathrm{PL}} &= E_{\mathcal{Y}\sim\mathrm{Pop}} \sum_n -\ln P_s(\mathcal{Y} \mid \mathcal{Y}/n) \\
&= E_{\mathcal{Y}\sim\mathrm{Pop}} \sum_n -\ln \frac{e^{s(\mathcal{Y})}}{Z_n} \qquad Z_n = \sum_{\ell'} e^{s(\mathcal{Y}[n=\ell'])} \\
&= E_{\mathcal{Y}\sim\mathrm{Pop}} \sum_n (\ln Z_n - s(\mathcal{Y})) \\
\nabla_\Phi \mathcal{L}_{\mathrm{PL}} &= E_{\mathcal{Y}\sim\mathrm{Pop}} \sum_n \left(\frac{1}{Z_n}\sum_{\ell'} e^{s(\mathcal{Y}[n=\ell'])}\; \nabla_\Phi\; s(\mathcal{Y}[n=\ell'])\right) - \nabla_\Phi s(\mathcal{Y}) \\
&= E_{\mathcal{Y}\sim\mathrm{Pop}} \sum_n \left(\sum_{\ell'} P_s(\mathcal{Y}[n]=\ell' \mid \mathcal{Y}/n)\; \nabla_\Phi\; s(\mathcal{Y}[n=\ell'])\right) - \nabla_\Phi s(\mathcal{Y})
\end{aligned}
$$

Gibbs CD1 Theorem

$$
\begin{aligned}
\nabla_\Phi \; \mathcal{L}_{\mathrm{PL}} &= E_{\mathcal{Y}\sim\mathrm{Pop}} \; \sum_n \left(\sum_{\ell'} P_s(\mathcal{Y}[n] = \ell' \mid \mathcal{Y}/n) \; \nabla_\Phi \; s(\mathcal{Y}[n = \ell']) \right) - \nabla_\Phi s(\mathcal{Y}) \\
&= E_{\mathcal{Y}\sim\mathrm{Pop}} \; \sum_n \big(E_{\ell' \sim P_s(\mathcal{Y}[n]=\ell' \mid \mathcal{Y}/n)} \nabla_\Phi \; s(\mathcal{Y}[n = \ell']) \big) - \nabla_\Phi s(\mathcal{Y}) \\
&\propto E_{\mathcal{Y}\sim\mathrm{Pop}} \; E_n \; E_{\ell' \sim P_s(\mathcal{Y}[n]=\ell' \mid \mathcal{Y}/n)} \;\; (\nabla_\Phi \; s(\mathcal{Y}[n = \ell']) - \nabla_\Phi s(\mathcal{Y})) \\
&= E_{\mathcal{Y}\sim\mathrm{Pop}} \; E_n \; E_{\ell' \sim P_s(\mathcal{Y}[n]=\ell' \mid \mathcal{Y}/n)} \;\; \nabla_\Phi \; \mathcal{L}_{\text{Gibbs CD(1)}}
\end{aligned}
$$

END