

TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Pseudo-Likelihood and Contrastive Divergence

Notation

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

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

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

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

Intractable Exponential Softmax

We consider a softmax distribution

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

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

Computing Z is intractable.

Pseudo-Likelihood

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

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

$$P(\mathcal{Y}[n] \mid \mathcal{Y}/n) = \frac{1}{Z_n} e^{s(\mathcal{Y})} \quad Z_n = \sum_y e^{s(\mathcal{Y}[n=y])}$$

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

Pseudo Cross-entropy Loss

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

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

Pseudolikelihood Theorem

$$\operatorname{argmin}_Q E_{\mathcal{Y} \sim \text{Pop}} - \ln \tilde{Q}(\mathcal{Y}) = \text{Pop}$$

It suffices to show that for any Q we have

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

Proof II

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

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

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

If $P_s = \text{Pop}$ 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 $y \sim P(\mathcal{Y}[n] = y \mid \mathcal{Y}/n)$. Define $\mathcal{Y}[n = y]$ to be the assignment (segmentation) which is the same as \mathcal{Y} except that node n is assigned label y . Take the loss to be

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

Gibbs CD1 Theorem

Gibbs CD1 is equivalent in expectation to pseudolikelihood.

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

Gibbs CD1 Theorem

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

END