TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Pseudo-Likelihood and Contrastive Divergence

Some Pseudo-Likelihood Notation

We let $\hat{\mathcal{Y}} \backslash n$ be the assignment of colors given by $\hat{\mathcal{Y}}$ except that no color is assigned to node n.

We let $\hat{\mathcal{Y}}[N(n)]$ be the assignment that $\hat{\mathcal{Y}}$ gives to the nodes (pixels) that are the neighbors of node n (connected to n by an edge.)

Psuedo-Likelihood

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

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

While computing $P_{\Phi,x}(\mathcal{Y})$ is intractable, computing $\tilde{P}_{\Phi,x}(\mathcal{Y})$ is tractable. We then use

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

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

We will prove the case of two nodes.

$$\min_{Q} \; E_{y \sim \mathrm{Pop}} - \ln Q(\mathcal{Y}[1]|\mathcal{Y}[2]) \; Q(\mathcal{Y}[2]|\mathcal{Y}[1])$$

$$\geq \; \min_{P_1,P_2} E_{\mathcal{Y} \sim \mathrm{Pop}} - \ln P_1(\mathcal{Y}[1]|\mathcal{Y}[2]) \; P_2(\mathcal{Y}[2]|\mathcal{Y}[1])$$

$$= \; \min_{P_1} E_{\mathcal{Y} \sim \mathrm{Pop}} - \ln P_1(\mathcal{Y}[1]|\mathcal{Y}[2]) + \min_{P_2} E_{\mathcal{Y} \sim \mathrm{Pop}} - \ln P_2(\mathcal{Y}[2]|\mathcal{Y}[1])$$

$$= \; E_{\mathcal{Y} \sim \mathrm{Pop}} - \ln \mathrm{Pop}(\mathcal{Y}[1]|\mathcal{Y}[2]) + E_{\mathcal{Y} \sim \mathrm{Pop}} - \ln \mathrm{Pop}(\mathcal{Y}[2]|\mathcal{Y}[1])$$

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

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 $\hat{\mathcal{Y}}'$. Then take the loss to be

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

If $P_s = \mathrm{Pop}$ then the the distribution on $\hat{\mathcal{Y}}'$ 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 $y \sim P(\mathcal{Y}[n] \mid \mathcal{Y}[N(n)])$. 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

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

Gibbs CD1 Theorem

Gibbs CD1 is equivalent in expectation to pseudolikelihood.

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

Gibbs CD1 Theorem

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

END