TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2019

Pseudo-Likelihood and Contrastive Divergence

Pseudolikelihood

For any distribution Q on assignments of labels to nodes (segmentations), and any assignment $\hat{y}$, we define $\tilde{Q}(\hat{y})$ as follows.

$$\tilde{Q}(\hat{y}) = \prod_i Q(\hat{y}[i] \mid \hat{y}/i) = \prod_i Q(\hat{y}[i] \mid \hat{y}[N(i)]$$

We then train a graphical model with pseudolikelyhood loss.

$$\Phi^* = \operatorname*{argmin}_\Phi E_{y\sim\mathrm{Pop}} \quad - \ln \tilde{P}_\Phi(y)$$

Pseudolikelihood

$$\mathcal{L}_{\mathrm{PL}} = -\ln \tilde{P}_s(y)$$

We note that by the Markov blanket property for Markov random fields we have

$$\tilde{P}_s(\hat{y}) = \prod_i \; P_s(\hat{y}[i] \mid \hat{y}[N(i)])$$

Since the loss is directly computed we can directly back-propagate on the loss.

Pseudolikelihood Theorem

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

or equivalently

$$\min_{Q} \; E_{y \sim \mathrm{Pop}} \; -\ln \tilde{Q}(y) = E_{y \sim \mathrm{Pop}} \; -\ln \widetilde{\mathrm{Pop}}(y)$$

Proof I

We have

$$\min_{Q}\ E_{y\sim\mathrm{Pop}}\ -\ln\tilde{Q}(y) \quad \leq \quad E_{y\sim\mathrm{Pop}}\ -\ln\widetilde{\mathrm{Pop}}(y)$$

So it suffices to show

$$\min_{Q}\ E_{y\sim\mathrm{Pop}}\ -\ln\tilde{Q}(y) \quad \geq \quad E_{y\sim\mathrm{Pop}}\ -\ln\widetilde{\mathrm{Pop}}(y)$$

Proof II

We will prove the case of two nodes.

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

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

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

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

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

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 y, start the MCMC process from initial state y and run the process for k steps to get $\hat{y}$. Then take the loss to be

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

If $P_s = \mathrm{Pop}$ then the the distribution on $\hat{y}$ is the same as the distribution on 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 y, select a node i at random and draw $c \sim P(y[i] \mid y[N(i)])$. Define $y[i = c]$ to be the assignment (segmentation) which is the same as y except that node i is assigned label c. Take the loss to be

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

Gibbs CD1 Theorem

Gibbs CD1 is equivalent in expectation to pseudolikelihood.

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

Gibbs CD1 Theorem

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

END