TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Noise Contrastive Estimation

Noise Contrastive Estimation

Gutmann and Hyvärinen, 2010

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \; -\ln P_\Psi(i|y_1,\ldots,y_N)$$

$$p_\Phi \text{ is fixed "noise"}$$

Assume p_Φ is both samplable and computable — we can sample from p_Φ and for any given y we can compute $p_\Phi(y)$.

Assume $P_\Psi(i|y_1,\ldots,y_N) = \text{softmax}_i\, s_\Psi(y_i)$

Assume Ψ universal

Noise Contrastive Estimation

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \; -\ln P_\Psi(i|y_1,\ldots,y_N)$$

p_Φ is fixed "noise"

Theorem: $\mathrm{pop}(y) = \mathrm{softmax}_y \quad s_{\Psi^*}(y) + \ln p_\Phi(y)$

We then have a computable score function (energy function) for the population. We do not have the partition function Z.

Noise Contrastive Estimation

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_{\Phi}^{(N)}} \; - \ln P_{\Psi}(i|y_1,\ldots,y_N)$$

p_Φ is fixed "noise"

Lemma: $P_{\Psi^*}(i|y_1,\ldots,y_N) = \mathrm{softmax}_i \; \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}$

Lemma Proof

$$\begin{aligned}
\tilde{p}_\Phi^{(N)}(i \text{ and } y_1, \ldots, y_N) &= \frac{1}{N}\,\mathrm{pop}(y_i) \prod_{j \neq i} p_\Phi(y_j) \\
&= \alpha \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}, \quad \alpha = \frac{1}{N} \prod_i p_\Phi(y_i)
\end{aligned}$$

$$\begin{aligned}
\tilde{p}_\Phi^{(N)}(i \mid y_1, \ldots y_N) &= \frac{\tilde{p}_\Phi^{(N)}(i \text{ and } y_1, \ldots, y_N)}{\sum_i \tilde{p}_\Phi^{(N)}(i \text{ and } y_1, \ldots, y_N)} = \frac{1}{Z} \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} \\
&= \operatorname*{softmax}_i \left(\ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} \right)
\end{aligned}$$

Theorem Proof

$$\operatorname*{softmax}_i s_{\Psi^*}(y_i) = \operatorname*{softmax}_i \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}$$

is solved by

$$s_{\Psi^*}(y) = \ln \frac{\mathrm{pop}(y)}{p_\Phi(y)} + \ln Z$$

giving

$$\mathrm{pop}(y) = \frac{1}{Z} \exp(s_\Psi(y) + \ln p_\Phi(y))$$

Another Theorem

$$E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \; - \ln p^{(N)}_{\Psi^*}(i|y_1,\ldots,y_N)$$

$$\geq \ln N - \frac{N-1}{N}(KL(\mathrm{pop}, p_\Phi) + KL(p_\Phi, \mathrm{pop}))$$

Note that the bound holds with equality for $p_\Phi = \mathrm{pop}$.

This is analogous to the JSD expression for the optimal discriminator.

Proof Part A.

$$
\begin{aligned}
& E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \;\; \ln p_{\Psi^*}(i|y_1,\ldots,y_N) \\
= \;& E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \;\; \ln\left(\operatorname*{softmax}_i \;\ln\frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}\right)[i] \\
= \;& E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \;\; \ln\frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln\left(\sum_j \frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)}\right) \\
= \;& E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}} \;\; \ln\frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln\left(\frac{1}{N}\sum_j \frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)}\right) - \ln\; N
\end{aligned}
$$

Proof Part B.

$$
\begin{aligned}
& E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}}\ \ln\frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \ln\left(\frac{1}{N}\sum_j \frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)}\right) - \ln\ N \\
\leq\ & E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}}\ \ln\frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} - \frac{1}{N}\sum_j \ln\frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)} - \ln\ N \\
=\ & E_{y\sim\mathrm{pop}}\ln\frac{\mathrm{pop}(y)}{p_\Phi(y)} - E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_\Phi^{(N)}}\frac{1}{N}\sum_j \ln\frac{\mathrm{pop}(y_j)}{p_\Phi(y_j)} - \ln N \\
=\ & \frac{N-1}{N}(KL(\mathrm{pop},p_\Phi) + KL(p_\Phi,\mathrm{pop})) - \ln\ N
\end{aligned}
$$

Noise Descriminiative Estimation

As in noise contrastive estimation we assume a noise distribution p_Φ that is both samplable and computable.

For noise descriminative estimation, and for for $i \in \{-1, 1\}$, we define a probability distribution over pairs $\langle i, y \rangle$ as in GANs.

$$\begin{aligned}
\tilde{p}_\Phi(i = 1) &= 1/2 \\
\tilde{p}_\Phi(y|i = 1) &= \mathrm{pop}(y) \\
\tilde{p}_\Phi(y|i = -1) &= p_\Phi(y)
\end{aligned}$$

Noise Discriminative Estimation

$$\Psi^* = \operatorname*{argmin}_{\Psi} \; E_{\langle i, y\rangle \sim \tilde{p}_\Phi} \; - \ln P_\Psi(i|y)$$

Assume $P_\Psi(i|y) = \mathrm{softmax}_i \; is_\Psi(y) = \frac{1}{1+e^{-2is_\Psi(y)}}$

Assume Ψ universal.

Noise Discriminative Estimation

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y)\sim\tilde{p}_\Phi} \; - \ln P_\Psi(i|y)$$

Theorem: $\mathrm{pop}(y) = \mathrm{softmax}_y \quad s_{\Psi^*}(y) + \ln p_\Phi(y)$

As with noise contrastive estimation, we have a computable score function (energy function) for the population. We do not have the partition function Z.

Noise Discriminative Estimation

$$\Psi^* = \operatorname*{argmin}_{\Psi} E_{(i,y_1,\ldots,y_N)\sim\tilde{p}_{\Phi}^{(N)}} \; -\ln P_{\Psi}(i|y_1,\ldots,y_N)$$

Lemma: $P_{\Psi^*}(i|y) = \text{softmax}_i \; is_{\Psi^*}(y)$

Lemma Proof

$$
\begin{aligned}
\tilde{p}_\Phi^{(N)}(i \text{ and } y_1, \ldots, y_N) &= \frac{1}{N}\,\mathrm{pop}(y_i) \prod_{j \neq i} p_\Phi(y_j) \\
&= \alpha \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}, \quad \alpha = \frac{1}{N} \prod_i p_\Phi(y_i)
\end{aligned}
$$

$$
\begin{aligned}
\tilde{p}_\Phi^{(N)}(i \mid y_1, \ldots y_N) &= \frac{\tilde{p}_\Phi^{(N)}(i \text{ and } y_1, \ldots, y_N)}{\sum_i \tilde{p}_\Phi^{(N)}(i \text{ and } y_1, \ldots, y_N)} = \frac{1}{Z}\,\frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} \\
&= \operatorname*{softmax}_i \left(\ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)} \right)
\end{aligned}
$$

Theorem Proof

$$\operatorname*{softmax}_i s_{\Psi^*}(y_i) = \operatorname*{softmax}_i \ln \frac{\mathrm{pop}(y_i)}{p_\Phi(y_i)}$$

is solved by

$$s_{\Psi^*}(y) = \ln \frac{\mathrm{pop}(y)}{p_\Phi(y)} + \ln Z$$

giving

$$\mathrm{pop}(y) = \frac{1}{Z} \exp(s_\Psi(y) + \ln p_\Phi(y))$$

END