TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Noisy Channel RDAs

The Fundamental Equation for Continuous y

If y is continuous then the fundamental equation for estimating the distribution on y (cross entropy) involves continuous probability densities.

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \; E_{y \sim \text{pop}} \; - \ln p_\Phi(y)$$

This occurs in unsupervised pretraining for sounds and images.

But differential entropy and differential cross-entropy are conceptually problematic.

Noisy Channel RDAs

In a noisy channel RDA we do not compress y into a finite number of bits.

Instead we add noise to a continuous representation.

As in the image compression RDA, the addition of noise is similar to rounding.

But instead of viewing the addition of noise as a hack to allow differentiation, we can reinterpret "rate" as mutual information and eliminate the discrete representation.

Noisy Channel RDAs

$$z = z_\Phi(y, \epsilon) \quad \epsilon \text{ is fixed (parameter independent) noise}$$

$$\Phi^* = \underset{\Phi}{\text{argmin}} \; I_\Phi(y, z) + \lambda E_{y,\epsilon} \, \text{Dist}(y, y_\Phi(z_\Phi(y, \epsilon)))$$

By the channel capacity theorem $I(y, z)$ is the **rate** of information transfer from y to z.

Noisy Channel RDAs

$$z = z_\Phi(y, \epsilon) \;\; \epsilon \text{ is fixed (parameter independent) noise}$$

$$\Phi^* = \underset{\Phi}{\text{argmin}} \;\; I_\Phi(y, z) + \lambda E_{y,\epsilon} \, \text{Dist}(y, y_\Phi(z_\Phi(y, \epsilon)))$$

Using parameter-independent noise is called the "reparameterization trick" and allows SGD.

$$\nabla_\Phi \; E_{y,\epsilon} \, \text{Dist}(y, y_\Phi(z_\Phi(y, \epsilon)))$$

$$= E_{y,\epsilon} \, \nabla_\Phi \, \text{Dist}(y, y_\Phi(z_\Phi(y, \epsilon)))$$

Mutual Information as a Channel Rate

Typically $z_\Phi(y, \epsilon)$ is simple. For example

$$\epsilon \sim \mathcal{N}(0, I)$$

$$z_\Phi(y, \epsilon) = \mu_\Phi(y) + \sigma_\Phi(y) \odot \epsilon$$

In this example $p_\Phi(z|y)$ is easily computed.

Mutual Information Replaces Rate

$$I_\Phi(y,z) \;=\; E_{y,\epsilon}\;\ln\frac{\mathrm{pop}(y)p_\Phi(z|y)}{\mathrm{pop}(y)p_{\mathrm{pop},\Phi}(z)}$$

$$\;=\; E_{y,\epsilon}\;\ln\frac{p_\Phi(z|y)}{p_{\mathrm{pop},\Phi}(z)}$$

where $\quad p_{\mathrm{pop},\Phi}(z) \;=\; E_{y\sim\mathrm{pop}}\;\; p_\Phi(z|y)$

A Variational Bound

$$p_{\text{pop},\Phi}(z) = E_{y \sim \text{pop}} \; p_\Phi(z|y)$$

We cannot compute $p_{\text{pop},\Phi}(z)$.

Instead we will use a model $\hat{p}_\Phi(z)$ to approximate $p_{\text{pop},\Phi}(z)$.

A Variational Bound

$$
\begin{aligned}
I(y,z) &= E_{y,\epsilon}\ \ln \frac{p_\Phi(z|y)}{p_{\mathrm{pop},\Phi}(z)} \\
&= E_{y,\epsilon}\ \ln \frac{p_\Phi(z|y)}{\hat{p}_\Phi(z)} + E_{y,\epsilon}\ \ln \frac{\hat{p}_\Phi(z)}{p_{\mathrm{pop},\Phi}(z)} \\
&= E_{y,\epsilon}\ \ln \frac{p_\Phi(z|y)}{\hat{p}_\Phi(z)} - KL(p_{\mathrm{pop},\Phi}(z), \hat{p}_\Phi(z)) \\
&\leq E_{y,\epsilon}\ \ln \frac{p_\Phi(z|y)}{\hat{p}_\Phi(z)}
\end{aligned}
$$

The Noisy Channel RDA

RDA: $z_\Phi(y)$ discrete

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\, E_{y\sim \text{pop}} - \ln P_\Phi(z_\Phi(y)) + \lambda \text{Dist}(y, y_\Phi(z_\Phi(y)))$$

Noisy Channel RDA: $z_\Phi(y, \epsilon)$ continuous

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\, E_{y,\epsilon}\ \ln \frac{p_\Phi(z_\Phi(y,\epsilon)|y)}{\hat{p}_\Phi(z_\Phi(y,\epsilon))} + \lambda \text{Dist}(y, y_\Phi(z_\Phi(y,\epsilon)))$$

END