TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Perils of Differential Entropy

Differential Entropy and Cross-Entropy

Cross entropy is a challenging objective for continuous structured values y such as images and sounds.

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

$$\Phi^* = \operatorname*{argmin}_{\Phi} \; E_{\langle x, y \rangle \sim \text{pop}} \; - \ln p_\Phi(y|x)$$

GANs replace the cross-entropy loss with an adversarial discrimination loss.

Rate-Distortion Auto-Encoders (RDAs) and Variational Auto-Encoders (VAEs) use the cross-entropy objective more directly.

Differential Cross-Entropy can Diverge to $-\infty$

For a uniform distribution over an interval on the real line of width Δ we have

$$H = E_{x\sim p} - \ln p(x)$$

$$= \ln \Delta$$

As $\Delta \to 0$ we have $H \to -\infty$.

This also happens for a Gaussian $\mathcal{N}(0, \sigma)$ as $\sigma \to 0$.

Differential Cross-Entropy can Diverge to $-\infty$

Consider the unsupervised training objective.

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

The training set is finite (discrete).

For each y the density $p_\Phi(y)$ can go to infinity.

This will drive the cross-entropy training loss to $-\infty$.

Sensitivity to the Choice of Units

$$H(N(0,\sigma)) = C + \ln \sigma$$

Differential entropy depends on the choice of units — a distributions on lengths will have a different entropy when measuring in inches than when measuring in feet.

Differential Entropy is Actually Infinite

An actual real number carries an infinite number of bits.

Consider quantizing the real numbers into bins.

A continuous probability densisty p assigns a probability $p(B)$ to each bin.

As the bin size decreases toward zero the entropy of the bin distribution increases toward ∞.

A meaningful convention is that $H(p) = +\infty$ for any continuous density p.

Differential KL-divergence is Meaningful

$$KL(p,q) = \int \left(\ln \frac{p(x)}{q(x)}\right) p(x)dx$$

Unlike differential entropy, differential KL divergence is always non-negative (but can be infinite).

Note that $KL(p,p) = 0$ independent of $H(p)$.

Mutual Information

For two random variables x and y there is a distribution on pairs (x, y) determined by the population distribution.

Mutual information is a KL divergence and hence differential mutual information is always non-negative.

$$I(x, y) \doteq KL(p(x, y), p(x)p(y))$$

$$= E_{x,y} \ln \frac{p(x, y)}{p(x)p(y)}$$

The Data Processing Inequality

For continuous y and z with $z = f(y)$ we get that $H(z)$ can be either larger or smaller than $H(y)$ (consider $z = ay$ for $a > 1$ vs. $a < 1$).

However, mutual information is a KL divergence and is more meaningful than entropy and for $z = f(y)$ we do have

$$I(x, z) \leq I(x, y)$$

END