TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Perils of Differential Entropy

Differential Entropy and Cross Entropy

For a probability density function $p(y)$ on continuous y it is standard practice to define differential entropy and cross entropy:

$$\begin{aligned} H(p) &= E_{y\sim p(y)} \quad -\ln p(y) \\ &= \int -\ln p(x) \;\; p(x)dx \end{aligned}$$

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

Difficulties with Differential Cross Entropy

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

Differential cross-entropy naturally occurs in modeling sounds and images.

But differential cross-entropy is problematic, especially for structured continuous values (images and sounds).

Note that GANs avoid differential cross entropy by defining $p_\Phi(y)$ as the distribution of $y_\Phi(z)$ for random noise z.

Perils of Differential Entropy

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 $H \to -\infty$.
Similarly, for a Gaussian $\mathcal{N}(0, \sigma))$ we have that as $\sigma \to 0$ we get$H(\mathcal{N}(0, \sigma)) \to -\infty$.

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 Cross Entropy can Diverge to $-\infty$

Consider the unsupervised training object.

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

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 $-\infty$.

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

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

This integral can be computed by dividing the real numbers into bins and computing the KL divergence between the distributions on bins.

The KL divergence between the bin distribution typically approaches a finite limit as the bin size goes to zero.

Unlike entropy, differential KL divergence is always non-negative. But as in the discrete case, it can be infinite.

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 meaningful.

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

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

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