TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2021

Parels of Differential Entropy

Differential Entropy

In the case of a continuous density (as opposed to a discrete probability) we have the notion of differential entropy.

For a density $p(x)$ on a real value x we have

$$
\begin{aligned}
H(p) &= E_{x \sim p}\ [-\ln p(x)] \\
&= \int_{-\infty}^{\infty} p(x)\ (-\ln p(x)) dx
\end{aligned}
$$

Differential Entropy can Diverge to $-\infty$

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

$$\begin{aligned} H &= E_{x \sim p}\left[- \ln p(x)\right] \\ &= E_{x \sim p}\left[- \ln \frac{1}{\Delta}\right] \\ &= \ln \Delta \end{aligned}$$

As $\Delta \rightarrow 0$ we have $H \rightarrow -\infty$

Differential Entropy can Diverge to $-\infty$

$$
\begin{aligned}
H(\mathcal{N}(0,\sigma^2)) &= E_{x\sim\mathcal{N}(0,\sigma^2)}\left[-\ln\left(\frac{1}{\sqrt{\pi}\sigma}\exp\frac{-x^2}{2\sigma^2}\right)\right] \\
&= E_{x\sim\mathcal{N}(0,\sigma^2)}\left[\ln(\sqrt{\pi}\sigma) + \frac{-x^2}{2\sigma^2}\right] \\
&= (\ln\sigma) + \ln(\sqrt{\pi}) + E_x\left[\frac{x^2}{2\sigma^2}\right] \\
&= (\ln\sigma) + \ln(\sqrt{\pi}) + \frac{1}{2}
\end{aligned}
$$

As $\sigma \rightarrow 0$ we have $H \rightarrow -\infty$

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 objective.

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

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 Can Be Considered 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.

$$
\begin{aligned}
I(x, y) &\doteq KL(p(x, y), p(x)p(y)) \\
&= E_{x,y} \ln \frac{p(x, y)}{p(x)p(y)}
\end{aligned}
$$

Mutual Information

$I(x, y)$ is the reduction in the number of bits we need to name y as a result of observing x (on average).

$$\begin{aligned} I(x,y) &= E\ \ln \frac{P(x,y)}{P(x)P(y)} \\ &= E\ \ln \frac{P(x,y)}{P(x)} - \ln P(y) \\ &= H(y) - H(y|x) \end{aligned}$$

Intuitively, how much does x know about 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

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

Continuous Cross-Entropy as Distortion

For a Gaussian VAE on images we typically have

$$\Phi^*, \Psi^* = \underset{\Phi,\Psi}{\text{argmin}}\; E_{y\sim \text{pop},z\sim \hat{p}_\Psi(z|y)}\;\; \ln \frac{\hat{p}_\Psi(z|y)}{p_\Phi(z)} - \ln p_\Phi(y|z)$$

$$p_\Phi(y|z) \propto \exp(||y - \hat{y}_\Phi(z)||^2/2\sigma^2)$$

$$\Phi^*, \Psi^* = \underset{\Phi,\Psi}{\text{argmin}}\; E_{y\sim \text{pop},z\sim \hat{p}_\Psi(z|y)}\;\; \ln \frac{\hat{p}_\Psi(z|y)}{p_\Phi(z)} + \frac{1}{2\sigma^2}||y - \hat{y}_\Phi(z)||^2$$

The KL divergence term is not problematic. The problematic differential cross-entropy term can just be thought of as a weighted L_2 distortion.

Continuous Cross-Entropy as Distortion

$$\Phi^*, \Psi^* \;=\; \underset{\Phi,\Psi}{\mathrm{argmin}}\; E_{y\sim\mathrm{pop},z\sim\hat{p}_\Psi(z|y)}\;\; \ln\frac{\hat{p}_\Psi(z|y)}{p_\Phi(z)} + \lambda\;\mathrm{Dist}(y, \hat{y}_\Phi(z))$$

Various choices for distortion are possible including L_2 and L_1 distortion measures.

$$\mathrm{Dist}(y, \hat{y}) \;=\; ||y - \hat{y}||^2 \qquad\qquad (L_2)$$

$$\text{or}\;\; \mathrm{Dist}(y, \hat{y}) \;=\; ||y - \hat{y}||_1 = \sum_i \;|y[i] - \hat{y}[i]| \qquad\qquad (L_1)$$

END