TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2021

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

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

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 $-\infty$

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

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

Sensitivity to the Choice of Units

$$H(\mathcal{N}(0, \sigma^2)) = C + \ln \sigma$$

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 $y \in \mathrm{Train}$ 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

$I(x, y)$ 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?

Differential Mutual Information

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

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

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

Assume that Train is a set of pairs (x, y) with $y \in R^d$.

Define $P_{\mathrm{Train},\sigma}$ by

$$(x, y + \sigma\epsilon),\ (x, y) \sim \mathrm{Train}, \epsilon \sim \mathcal{N}(0, I)$$

Define $P_{\Phi,\sigma}(y|x)$ by

$$(\hat{y}_\Phi(x) + \epsilon),\ \ \epsilon \sim \mathcal{N}(0, I)$$

$$\Phi^* = \operatorname*{argmin}_\Phi\ KL(p_{\mathrm{Train}}(z|y), p_{\mathrm{pri}}(z)) + \lambda\ \mathrm{Dist}(y, \hat{y}_{\mathrm{dec}}(z))$$

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

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

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

END