TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2023

Some Information Theory

Why Information Theory?

The fundamental equation of deep learning involves cross-entropy.

Cross-entropy is an information-theoretic concept.

Information theory arises in many places and many forms in deep learning.

Entropy of a Distribution

The entropy of a distribution P is defined by

$$H(P) = E_{y \sim P}\,[\,-\ln P(y)] \text{ in units of "nats"}$$

$$H_2(P) = E_{y \sim P}\,[\,-\log_2 P(y)] \text{ in units of bits}$$

Why Bits?

Why is $-\log_2 P(y)$ a number of bits?

Example: Let P be a uniform distribution on 256 values.

$$E_{y \sim P} [-\log_2 P(y)] = -\log_2 \frac{1}{256} = \log_2 256 = 8 \text{ bits} = 1 \text{ byte}$$

$$1 \text{ nat} = \frac{1}{\ln 2} \text{ bits} \approx 1.44 \text{ bits}$$

Shannon's Source Coding Theorem

Why is $-\log_2 P(y)$ a number of bits?

A prefix-free code for $\mathcal{Y}$ assigns a bit string $c(y)$ to each $y \in \mathcal{Y}$ such that no code string is a prefix of any other code string.

For a probability distribution P on $\mathcal{Y}$ we consider the average code length $E_{y \sim P}\ [\ |c(y)|]$.

Theorem: For any c we have $E_{y \sim P}\ |c(y)| \geq H_2(P)$.

Theorem: There exists c with $E_{y \sim P}\ |c(y)| \leq H_2(P) + 1$.

Cross Entropy

Let P and Q be two distribution on the same set.

$$H(P,Q) = E_{y\sim P}\ [-\ln\ Q(y)]$$

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}}\ H(\text{Pop}, P_\Phi)$$

We will show $H(P,Q) \geq H(P)$

We have $H(P,P) = H(P)$

Universality implies $H(\text{Pop}, P_{\Phi^*}) = H(\text{Pop})$ or $P_{\Phi^*} = \text{Pop}$

KL Divergence

Let P and Q be two distribution on the same set.

$$\text{Entropy :} \qquad H(P) = E_{y\sim P}\ [-\ln\ P(y)]$$

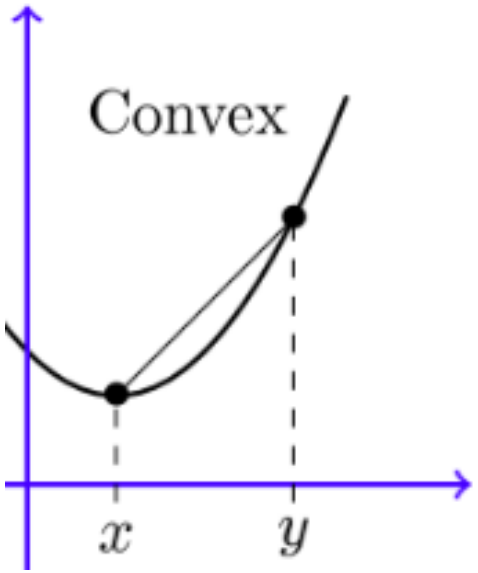
$$\text{CrossEntropy :} \qquad H(P,Q) = E_{y\sim P}\ [-\ln\ Q(y)]$$

$$\text{KL Divergence :} \quad KL(P,Q) = H(P,Q) - H(P)$$

$$= E_{y\sim P} \quad -\ln \frac{Q(y)}{P(y)}$$

We will show $KL(P,Q) \geq 0$ which implies $H(P,Q) \geq H(P)$.

Proving $KL(P, Q) \geq 0$: Jensen's Inequality



For f convex (upward curving) we have

$$E[f(x)] \geq f(E[x])$$

Proving $KL(P, Q) \geq 0$

$$
\begin{aligned}
KL(P,Q) &= E_{y\sim P}\left[-\ln \frac{Q(y)}{P(y)}\right] \\
&\geq -\ln E_{y\sim P}\,\frac{Q(y)}{P(y)} \\
&= -\ln \sum_y P(y)\frac{Q(y)}{P(y)} \\
&= -\ln \sum_y Q(y) \\
&= 0
\end{aligned}
$$

Conditional Entropy and Mutual Information

Assume a joint distribution Q on x and y.

conditional entropy:

$$H(y|x) = E_{(x,y)\sim Q} \; -\ln P(y|x)$$

mutual information:

$$I(x,y) = H(y) - H(y|x) = H(x) - H(y|x)$$

Suppose you dont't know anything about x and y. The mutual information $I(x,y)$ is the expectation over a draw of x of the number of bits you learn about y.

Continuous Densities

Expectations hide the discrete-continuous distinction

$E_{x \sim P(x)}\ f(x)$ is meaningful for both discrete and continuous $P(x)$.

$E_{x \sim P(x)}\ f(x)$ is the limit of the average of $f(x)$ over ever larger samples.

In order to write general equations we will use capital letter notation $P(x)$ for both continuous densities and discrete distributions.

Infinite Precision Real Numbers

An infinite precision real number contains an infinite amount of information (we have an infinite sequence of decimal places).

For a continuous density the probability of a particular infinite precision real number x is zero and $-\ln P(x) = \infty$. This should $H(P) = \infty$

However, for a continuous density P the (unfortunate) convention is to define "differential entropy" $H(P)$ in terms of *probability density* $p(x)$ rather than the probbility $P(x)$ (which is zero).

Differential Entropy: $H_{\text{diff}}(p) = E_{x \sim P(x)} \ [-\ln p(x)]$

Infinite Precision Real Numbers

Since a density can be greater than 1 we have that $-\ln p(x)$ can be negative.

The cross entropy loss as measured by differential entropy can diverge to minus infinity during training.

This will happen in a Gaussian mixture models where one of the Gaussians converges on modelng a single training point.

We do not want to minize an objective that diverges to $-\infty$.

Limiting Numerical Precision

We can avoid the difficulties of infinite precision numbers by just replacing them with floatng point numbers.

Alteratively we can limit precision by adding noise and working with mutual information rather than entropy.

Summary

Entropy : $H(P) = E_{y\sim P}\ [-\ln\ P(y)]$

CrossEntropy : $H(P,Q) = E_{y\sim P}\ [-\ln\ Q(y)]$

KL Divergence : $KL(P,Q) = H(P,Q) - H(P)$

Mutual Information : $I(x,y) = H(y) - H(y|x)$

$H(P,Q) \geq H(P), \quad KL(P,Q) \geq 0, \quad \text{argmin}_Q\ H(P,Q) = P$

END