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 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.

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

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

$H(P,Q)$ can be interpreted as the number of bits used to code draws from P when using an optimal code for Q.

We will show

$$H(P,Q) \geq H(P)$$

KL Divergence

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

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

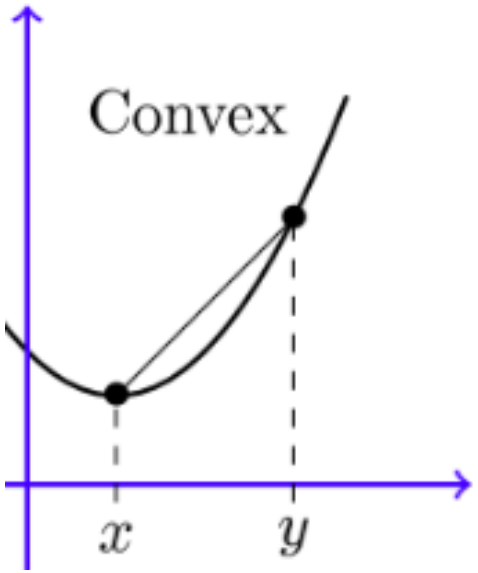
$$\text{CrossEntropy :} \qquad H(P,Q) = E_{y\sim P}\ [-\ln\ Q(y)]$$

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

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

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 $KL(P, Q) \geq 0$

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

Asymmetry of Cross Entropy

Consider

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \; H(\mathrm{Pop}, Q_\Phi) \qquad (1)$$

$$\Phi^* = \underset{\Phi}{\operatorname{argmin}} \; H(Q_\Phi, \mathrm{Pop}) \qquad (2)$$

We cannot use (2) because we cannot calculate $\mathrm{Pop}(y)$.

For a synthetic population where $\mathrm{Pop}(y)$ is computable (2) produces mode collapse — Q_Φ is concentrated on the most likely value of Pop.

Asymmetry of KL Divergence

Consider

$$\begin{aligned}\Phi^* &= \underset{\Phi}{\operatorname{argmin}}\ KL(\text{Pop}, Q_\Phi) \\ &= \underset{\Phi}{\operatorname{argmin}}\ H(\text{Pop}, Q_\Phi) \qquad (1)\end{aligned}$$

$$\begin{aligned}\Phi^* &= \underset{\Phi}{\operatorname{argmin}}\ KL(Q_\Phi, \text{Pop}) \\ &= \underset{\Phi}{\operatorname{argmin}}\, H(Q_\Phi, \text{Pop}) - H(Q_\Phi) \quad (2)\end{aligned}$$

For a synthetic population where $\text{Pop}(y)$ is computable but P_Φ cannot perfectly model Pop, (2) produces mode collapse.

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:

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

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.

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