

TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Some Information Theory

Why Information Theory?

The fundamental equation 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.

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

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

$H(P, Q)$ also has a data compression interpretation.

$H(P, Q)$ can be interpreted as 1.44 times the number of bits used to code draws from P when using the imperfect code defined by Q .

Entropy, Cross Entropy and KL Divergence

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

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

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

$$\begin{aligned} \text{KL Divergence :} \quad KL(P, Q) &= H(P, Q) - H(P) \\ &= E_{y \sim P} \ln \frac{P(y)}{Q(y)} \end{aligned}$$

We have $H(P, Q) \geq H(P)$ or equivalently $KL(P, Q) \geq 0$.

The Universality Assumption

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

Universality assumption: P_{Φ} can represent any distribution and Φ can be fully optimized.

This is clearly false for deep networks. But it gives important insights like:

$$P_{\Phi^*} = \text{Pop}$$

This is the motivation for the fundamental equation.

Asymmetry of Cross Entropy

Consider

$$\Phi^* = \operatorname{argmin}_{\Phi} H(P, Q_{\Phi}) \quad (1)$$

$$\Phi^* = \operatorname{argmin}_{\Phi} H(Q_{\Phi}, P) \quad (2)$$

For (1) Q_{Φ} must cover all of the support of P .

For (2) Q_{Φ} concentrates all mass on the point maximizing P .

Asymmetry of KL Divergence

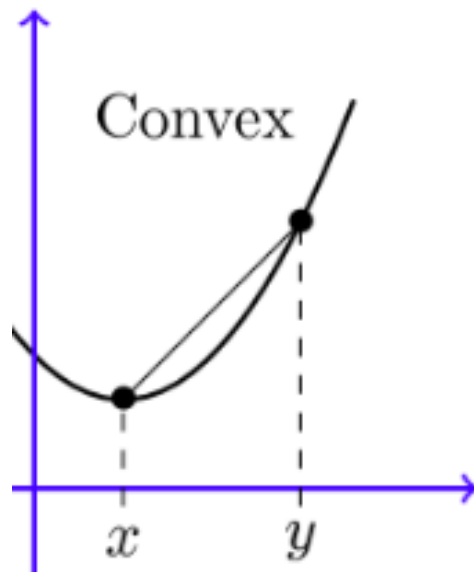
Consider

$$\begin{aligned}\Phi^* &= \operatorname{argmin}_{\Phi} KL(P, Q_{\Phi}) \\ &= \operatorname{argmin}_{\Phi} H(P, Q_{\Phi})\end{aligned}\tag{1}$$

$$\begin{aligned}\Phi^* &= \operatorname{argmin}_{\Phi} KL(Q_{\Phi}, P) \\ &= \operatorname{argmin}_{\Phi} H(Q_{\Phi}, P) - H(Q_{\Phi})\end{aligned}\tag{2}$$

If Q_{Φ} is not universally expressive we have that (1) still forces Q_{Φ} to cover all of P (or else the KL divergence is infinite) while (2) allows Q_{Φ} to be restricted to a single mode of P (a common outcome).

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



For f convex (upward curving) we have

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

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

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

Appendix: The Rearrangement Trick

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

In general $E_{x \sim P} \ln (\prod_i A_i) = E_{x \sim P} \sum_i \ln A_i$

Summary

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

$$\Phi^* = \operatorname{argmin}_{\Phi} E_{x \sim \text{Pop}} H(\text{Pop}(y|x), P_{\Phi}(y|x)) \text{ conditional}$$

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

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

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

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

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

END