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

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

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

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.

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

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

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

We will show

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

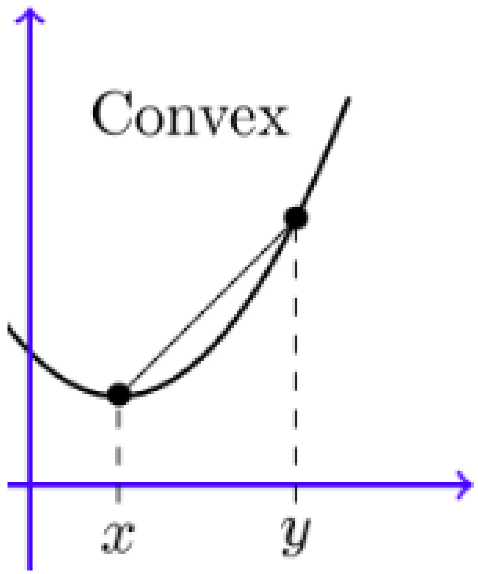
KL Divergence

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

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

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

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

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

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

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

$$\begin{aligned}
\Phi^* &= \operatorname*{argmin}_\Phi \; KL(Q_\Phi, \mathrm{Pop}) \\
&= \operatorname*{argmin}_\Phi H(Q_\Phi, \mathrm{Pop}) - H(Q_\Phi) \qquad (2)
\end{aligned}$$

For a synthetic population where $\mathrm{Pop}(y)$ 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.

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.

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

$$H(x) = E_{x \sim P(x)} \left[- \ln P(x) \right]$$

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 Cross-Entropy can Diverge to $-\infty$

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

For a Gaussian mixture model in which some mixtures are focused on a single point the training loss goes to $-\infty$.

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

The Gaussian Noise Trick (L_2 Distortion)

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

Linear regression invokes the Gaussian noise trick.

Define $P_{\Phi,\sigma}(y|x)$ by $(\hat{y}_\Phi(x) + \epsilon)$, $\quad \epsilon \sim \mathcal{N}(0, I)$.

$$\begin{aligned} \Phi^* &= \operatorname*{argmin}_\Phi \; E_{(x,y)\sim \mathrm{Train}} \left[- \ln P_{\Phi,\sigma}(y|x)\right] \\ &= \operatorname*{argmin}_\Phi E_{(x,y)\sim \mathrm{Train}} \left[\frac{||y - \hat{y}(x)||^2}{2\sigma^2}\right] \end{aligned}$$

L_2 distortion is non-negative but goes to infinity as $\sigma \to 0$.

The Gaussian Noise Trick (L_2 Distortion)

For sufficiently small σ we have that L_2 distortion is simply accounting for numerical precision.

Intuitively this corresponds to using a discrete distribution defined by finite precision arithmetic with rounding error σ.

We should **not** think of the Gaussian noise trick as introducing a Bayesian assumption.

We can prove PAC-Bayesian generalization guarantees for this trick (no Bayesian assumptions).

The Laplacian Noise Trick (L_1 Distortion)

define $P_{\Phi,\lambda}(y|x)$ by $(\hat{y}_\Phi(x) + \epsilon)$, $\quad \epsilon \sim \mathrm{softmax}_\epsilon \;\lambda|\epsilon|_1$

$$||\epsilon||_1 = \sum_i |\epsilon_i|$$

$$\begin{aligned} \Phi^* &= \operatorname*{argmin}_\Phi \; E_{(x,y)\sim\mathrm{Train}}\left[-\ln P_{\Phi,\lambda}(y|x)\right] \\ &= \operatorname*{argmin}_\Phi E_{(x,y)\sim\mathrm{Train}}\left[\frac{||y - \hat{y}(x)||_1}{\lambda}\right] \end{aligned}$$

The Gaussian Noise Trick

"Finite Precision" Differential Entropy

Define $P_\sigma(\tilde{y}|y)$ by

$$\tilde{y} = y + \epsilon, \ \epsilon \sim \mathcal{N}(0, I)$$

Define $H_\sigma(y)$ by

$$H_\sigma(y) \ = \ E_{y \sim P(y), \tilde{y} \sim P_\sigma(\tilde{y}|y)} \left[-\ln \frac{P(y)}{P(y|\tilde{y})} \right]$$

$$= \ I(y, \tilde{y}) \geq 0$$

The Gaussian Noise Trick

"Finite Precision" Differential Entropy

If $P(y)$ is smooth at the scale of σ we have $P(y|\tilde{y})$ is a Gaussian centered at $\tilde{y}$.

$$E_{y,\tilde{y}}[-\ln P(y|\tilde{y})] \approx d(\ln \sigma + \ln \sqrt{2\pi} + 1/2)$$

$$H_\sigma(y) \approx E_{y \sim P(y)}\left[-\ln P(y)\right] - d(\ln \sigma + \ln \sqrt{2\pi} + 1/2)$$

For P smooth at scale σ this approximates mutual information and is non-negative.

But $H_\sigma(y)$ goes to infinity (slowly) as $\sigma \to 0$.

Summary

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

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

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

$$\text{Mutual Information}: \quad I(x,y) = H(y) - H(y|x)$$

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

END