TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Some Information Theory

Entropy of a Distribution

The entropy of a distribution P is defined by

$$H(P) = E_{y \sim \mathrm{Pop}} \; -\ln P(y) \;\; \text{in units of "nats"}$$

$$H_2(P) = E_{y \sim \mathrm{Pop}} \; -\log_2 P(y) \; \text{in units of bits}$$

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

$$E_{y \sim Q} \; -\log_2 Q(y) = -\log_2 \frac{1}{256} = \log_2 256 = 8 \text{ bits} = 1 \text{ byte}$$

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

Shannon's Source Coding Theorem

We can interpret $H_2(Q)$ as the number of bits required an average to represent items drawn from distribution Q.

We want to use fewer bits for common items.

For any probability distribution P on a (discrete) set $\mathcal{Y}$ there exist a code mapping $\mathcal{Y}$ to bit strings such that the number of bits used to represent y is no larger than $-\log_2 P(y) + 1$.

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) 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.

$$\begin{array}{lrcl}
\text{Entropy :} & H(P) & = & E_{y\sim P}\; -\ln\; P(y) \\[2ex]
\text{CrossEntropy :} & H(P,Q) & = & E_{y\sim P}\; -\ln\; Q(y) \\[2ex]
\text{KL Divergence :} & KL(P,Q) & = & H(P,Q) - H(P) \\[2ex]
 & & = & E_{y\sim P}\quad \ln\; \frac{P(y)}{Q(y)}
\end{array}$$

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

The Universality Assumption

$$\Phi^* = \underset{\Phi}{\mathrm{argmin}}\; H(\mathrm{Pop}, P_\Phi) = \underset{\Phi}{\mathrm{argmin}}\; H(\mathrm{Pop}) + KL(\mathrm{Pop}, P_\Phi)$$

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:

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

This is the motivatation for the fundamental equation.

Asymmetry of Cross Entropy

Consider

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

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

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

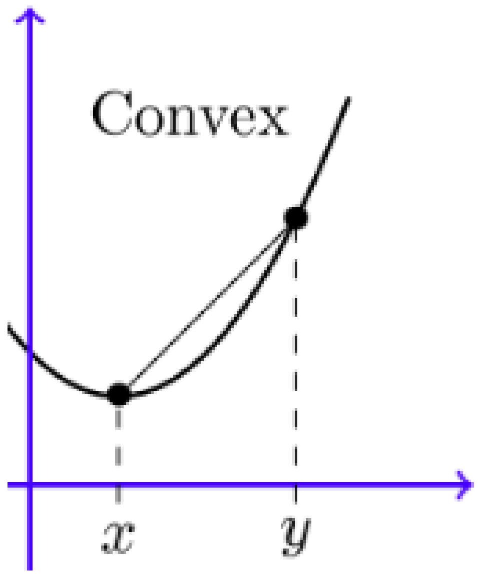
Consider

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

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

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 $KL(P,Q) \geq 0$

$$\begin{aligned}
KL(P,Q) &= E_{y\sim P}\; -\log\frac{Q(y)}{P(y)} \\
&\geq -\log E_{y\sim P}\;\frac{Q(y)}{P(y)} \\
&= -\log\sum_y\; P(y)\frac{Q(y)}{P(y)} \\
&= -\log\sum_y Q(y) \\
&= 0
\end{aligned}$$

Summary

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

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

$$\begin{array}{lrcl}
\text{Entropy :} & H(P) & = & E_{y\sim P}\; -\ln\; P(y) \\[2ex]
\text{CrossEntropy :} & H(P,Q) & = & E_{y\sim P}\; -\ln\; Q(y) \\[2ex]
\text{KL Divergence :} & KL(P,Q) & = & H(P,Q) - H(P) \\[2ex]
 & & = & E_{y\sim P}\;\; \ln\; \frac{P(y)}{Q(y)}
\end{array}$$

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

Appendix: The Rearrangement Trick

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

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

Appendix: The Rearrangement Trick

$$\text{ELBO} = E_{z\sim P_\Psi(z|y)} \ln \frac{P_\Phi(z,y)}{P_\Psi(z|y)}$$

$$= E_{z\sim P_\Psi(z|y)} \ln \frac{P_\Phi(z)P_\Phi(y|z)}{P_\Psi(z|y)}$$

$$= E_{z\sim P_\Psi(z|y)} \ln \frac{P_\Phi(y)P_\Phi(z|y)}{P_\Psi(z|y)}$$

Each of the last two expressions can be grouped three different ways leading to six ways of writing the ELBO.

END