

Optimization

Mahdi Roozbahani
Georgia Tech

Outline

Motivation

Entropy

Conditional Entropy and Mutual Information

Cross-Entropy and KL-Divergence



Let's work on this subject in our Optimization lecture

Cross Entropy

Cross Entropy: The expected number of bits when a wrong distribution Q is assumed while the data actually follows a distribution P

$$H(p, q) = - \sum_{x \in \mathcal{X}} p(x) \log q(x) = H(P) + KL[P][Q]$$

This is because:

$$H(p, q) = \mathbb{E}_p[l_i] = \mathbb{E}_p \left[\log \frac{1}{q(x_i)} \right]$$

$$H(p, q) = \sum_{x_i} p(x_i) \log \frac{1}{q(x_i)}$$

$$H(p, q) = - \sum_x p(x) \log q(x).$$

Labeling target values

Label encoding (ordinal) and One-hot encoding

Why Cross entropy and not simply use dot product?

Kullback-Leibler Divergence

Another useful information theoretic quantity measures the difference between two distributions.

$$\begin{aligned}\mathbf{KL}[P(S)||Q(S)] &= \sum_s P(s) \log \frac{P(s)}{Q(s)} \\ &= \underbrace{\sum_s P(s) \log \frac{1}{Q(s)}}_{\text{cross entropy}} - \mathbf{H}[P] = H(P, Q) - H(P)\end{aligned}$$

Excess cost in bits paid by encoding according to Q instead of P .

KL Divergence is
a **KIND OF**
distance
measurement

$$-\mathbf{KL}[P||Q] = \sum_s P(s) \log \frac{Q(s)}{P(s)}$$

log function is
concave or
convex?

$$\begin{aligned}\sum_s P(s) \log \frac{Q(s)}{P(s)} &\leq \log \sum_s P(s) \frac{Q(s)}{P(s)} && \text{By Jensen Inequality} \\ &= \log \sum_s Q(s) = \log 1 = 0\end{aligned}$$

So $\mathbf{KL}[P||Q] \geq 0$. Equality iff $P = Q$

When $P = Q$, $KL[P||Q] = 0$