The Mathematical Engineering of Deep Learning

Book Draft

Benoit Liquet, Sarat Moka and Yoni Nazarathy

December 6, 2022

Contents

B. Cross Entropy and Other Expectations with Logarithms - DRAFT

This appendix expands on basic properties of cross entropy, the KL divergence, and basics of information theory. It is not meant to be an extensive review of these areas but rather provides key definitions, properties, and results needed for the content of the book.

B.1. Divergences and Entropies

We first define the relative entropy (KL divergence), cross entropy, and entropy in the context of probability distributions. We then define the Jensen–Shannon divergence. The information theoretic meaning of these quantities and a few related measures is described in Section B.2 below. At this point, we simply make formal use of these definitions.

Assume two probability distributions, $p(\cdot)$, and $q(\cdot)$, over elements in some discrete set $\mathcal{X}$[1]. That is, for each $x \in \mathcal{X}$ let $p(x)$ or $q(x)$ denote the respective probabilities, and assume that for $x \notin \mathcal{X}$, $p(x)$ and $q(x)$ are both zero. The definitions we present here also extend to continuous, multi-variate, and more general distributions, yet for simplicity we restrict the notation to summations over elements of discrete distributions.

A key measure for the proximity between the distributions $p(\cdot)$ and $q(\cdot)$ is the *Kullback–Leibler divergence*, also shortened as *KL divergence* or called the *relative entropy*. It is denoted $D_{\mathrm{KL}}(p \parallel q)$ and is the expected value of $\log p(X)/q(X)$ where X is a random variable following the probability law $p(\cdot)$. Namely,

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}.$$

Note that the expression for $D_{\mathrm{KL}}(p \parallel q)$ can be decomposed into the difference of $H(p)$ from $H(p, q)$ via,

$$D_{\mathrm{KL}}(p \parallel q) = \underbrace{\sum_{x \in \mathcal{X}} p(x) \log \frac{1}{q(x)}}_{H(p,q)} - \underbrace{\sum_{x \in \mathcal{X}} p(x) \log \frac{1}{p(x)}}_{H(p)}.$$

Here

$$H(p, q) = -\sum_{x \in \mathcal{X}} p(x) \log q(x) \tag{B.1}$$

is called the *cross entropy* of p and q and

$$H(p) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) \tag{B.2}$$

[1] The set $\mathcal{X}$ can be taken as the union of the supports of the distributions.

is called the entropy of p. Hence in words, the KL divergence or relative entropy of p and q is the cross entropy of p and q with the entropy of p subtracted. Note that in case where there are only two values in $\mathcal{X}$, say 0 and 1, where we denote $p(1) = p_1$, and $q(1) = q_1$, we have

$$H(p) = -\big(p_1 \log p_1 + (1 - p_1) \log(1 - p_1)\big) \tag{B.3}$$

$$H(p, q) = -\big(p_1 \log q_1 + (1 - p_1) \log(1 - q_1)\big). \tag{B.4}$$

Some observations are in order. First observe that $D_{\mathrm{KL}}(p \parallel q) \geq 0$. Further note that in general $D_{\mathrm{KL}}(p \parallel q) \neq D_{\mathrm{KL}}(q \parallel p)$ and similarly $H(p, q) \neq H(q, p)$. Hence as a "distance measure" the KL divergence is not a true metric since it isn't symmetric over its arguments. Nevertheless, when $p = q$ the KL divergence is 0 and similarly the cross entropy equals the entropy. In addition, it can be shown that $D_{\mathrm{KL}}(p \parallel q) = 0$ only when $p = q$. Hence the KL divergence may play a role similar to a distance metric in certain applications. In fact, in some applications one may consider a sequence $q^{(1)}, q^{(2)}, \ldots$ which has decreasing $D_{\mathrm{KL}}(p \parallel q^{(t)})$ approaching 0 as $t \to \infty$. For such a sequence the probability distributions $q^{(t)}$ approach the target distribution p since the KL divergence convergences to 0.

A related measure to the KL divergence which is symmetric in arguments is the *Jensen-Shannon divergence* denoted $\mathrm{JSD}(p \parallel q)$. It is defined by considering a mixture distribution,

$$m(x) = \frac{1}{2}\big(p(x) + q(x)\big),$$

and then averaging the KL divergence between each of the distributions and $m(\cdot)$, namely,

$$\mathrm{JSD}(p \parallel q) = \frac{D_{\mathrm{KL}}(p \parallel m) + D_{\mathrm{KL}}(q \parallel m)}{2}. \tag{B.5}$$

The square square root of $\mathrm{JSD}(p \parallel q)$, sometimes called the *Jensen-Shannon distance* is a metric in the mathematical sense.

B.2. Interpretation via Information Theory