Mathematical Engineering of Deep Learning

Book Draft

Benoit Liquet, Sarat Moka and Yoni Nazarathy

February 28, 2024

Contents

B Cross Entropy and Other Expectations with Logarithms - DRAFT

This appendix expands on basic properties of cross entropy, the KL-divergence, and related concepts, also in the context of the multivariate normal distribution. It is not meant to be an extensive review of these concepts 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 discrete probability distributions. We then present we then provide a definition of the KL-divergence for continuous random variables. Finally we define the Jensen–Shannon divergence.

The KL-divergence for Discrete Distributions

Assume two probability distributions $p(\cdot)$ and $q(\cdot)$ over elements in some discrete sets $\mathcal{X}_p$ and $\mathcal{X}_q$ respectively. That is, $p(x)$ or $q(x)$ denote the respective probabilities, which are strictly positive unless $x \notin \mathcal{X}_p$ for which $p(x) = 0$ (or similarly $x \notin \mathcal{X}_q$ for which $q(x) = 0$).

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

$$D_{\mathrm{KL}}(p \parallel q) = \sum_{x \in \mathcal{X}_p} p(x) \log \frac{p(x)}{q(x)}. \tag{B.1}$$

Further if $\mathcal{X}_p \not\subseteq \mathcal{X}_q$, that is if there is some element in $\mathcal{X}_p$ that is not in $\mathcal{X}_q$, then by definition $D_{\mathrm{KL}}(p \parallel q) = +\infty$. This definition as infinity is natural since we would otherwise divide by 0 for some $q(x)$.

Observe that the expression for $D_{\mathrm{KL}}(p \parallel q)$ from (B.1) 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,

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

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

$$H(p) = -\sum_{x\in\mathcal{X}} p(x)\log p(x) \tag{B.3}$$

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

$$H(p) = -\big(p_1\log p_1 + (1-p_1)\log(1-p_1)\big), \tag{B.4}$$
$$H(p,q) = -\big(p_1\log q_1 + (1-p_1)\log(1-q_1)\big). \tag{B.5}$$

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 is not 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, 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[1] the target distribution p since the KL-divergence convergences to 0.

The KL-divergence for Continuous Distributions

The KL-divergence in (B.1) naturally extends to arbitrary probability distributions that are not necessarily discrete. In our case let us consider continuous multi-dimensional distributions. In this case $p(\cdot)$ and $q(\cdot)$ are probability densities, and the sets $\mathcal{X}_p$, and $\mathcal{X}_q$ are their respective supports. Now very similarly to (B.1), as long as $\mathcal{X}_p \subseteq \mathcal{X}_q$ we define,

$$D_{\mathrm{KL}}(p \parallel q) = \int_{x\in\mathcal{X}_p} p(x)\log\frac{p(x)}{q(x)}\,dx. \tag{B.6}$$

The Jensen-Shannon Divergence

A related measure to the KL-divergence which is symmetric in arguments is the *Jensen-Shannon divergence* denoted $\mathrm{JSD}(p \parallel q)$. Either for the discrete or continuous case, is defined by considering a mixture distribution with support $\mathcal{X}_p \cup \mathcal{X}_q$,

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

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

$$\mathrm{JSD}(p \parallel q) = \frac{D_{\mathrm{KL}}(p \parallel m) + D_{\mathrm{KL}}(q \parallel m)}{2}. \tag{B.7}$$

[1] There are multiple ways to define convergence of such a sequence of probability distributions. The exact form is out of our scope.

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

B.2 Computations for Multivariate Normal Distributions

A univariate (single variable) *normal*, or *Gaussian*, distribution has a probability density function,

$$\mathcal{N}(x\,;\,\mu,\sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}, \qquad \text{for} \qquad x \in \mathbb{R},$$

and is parameterized by $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ which are the mean and variance of the distribution respectively. The *standard normal* case has $\mu = 0$ and $\sigma^2 = 1$.

An m dimensional multivariate normal distribution is characterized by a mean vector $\mu \in \mathbb{R}^m$ and a covariance matrix $\Sigma \in \mathbb{R}^{m\times m}$ which is assumed to be symmetric and positive definite. The probability density function (pdf) of a multivariate normal distribution is,

$$\mathcal{N}(x\,;\,\mu,\Sigma) = \frac{1}{(\det\Sigma)^{1/2}(2\pi)^{m/2}} e^{-\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu)}, \qquad \text{for} \qquad x \in \mathbb{R}^m,$$

where $\det\Sigma$ stands for the determinant of a matrix Σ. There are many useful formulas associated with this distribution with one particular case being the *log-density*,

$$\log\mathcal{N}(x\,;\,\mu,\Sigma) = -\frac{1}{2}(x-\mu)^\top\Sigma^{-1}(x-\mu) - \frac{m}{2}\log(2\pi) - \frac{1}{2}\log(\det\Sigma). \tag{B.8}$$

It is also useful to consider the KL-divergence between two multivariate normal distributions. For short, denote such a distribution as $\mathcal{N}_{\mu,\Sigma}$ when the mean vector is μ and the covariance matrix is Σ. Then if we consider two such distributions on $\mathbb{R}^m$ with corresponding mean vectors μ_1 and μ_2, and corresponding covariance matrices Σ_1 and Σ_2, then it is possible to show that,

$$D_{\mathrm{KL}}(\mathcal{N}_{\mu_1,\Sigma_1} \parallel \mathcal{N}_{\mu_2,\Sigma_2}) = \frac{1}{2}\Big((\mu_1-\mu_2)^\top\Sigma_2^{-1}(\mu_1-\mu_2) - m + \mathrm{tr}(\Sigma_2^{-1}\Sigma_1) + \log\frac{\det(\Sigma_2)}{\det(\Sigma_1)}\Big). \tag{B.9}$$

A particularly useful case is one where $\Sigma_2 = \sigma_2^2 I$ for some constant $\sigma_2^2 > 0$. In this case,

$$D_{\mathrm{KL}}(\mathcal{N}_{\mu_1,\Sigma_1} \parallel \mathcal{N}_{\mu_2,\sigma_2^2 I}) = \frac{1}{2\sigma_2^2}\|\mu_1-\mu_2\|^2 - \frac{m}{2} + \frac{\mathrm{tr}(\Sigma_1)}{2\sigma_2^2} + \frac{m\log\sigma_2^2}{2} - \frac{\log\det(\Sigma_1)}{2}. \tag{B.10}$$

Furthermore, if the second distribution is standard, i.e., $\mu_2 = 0$ and $\sigma_2^2 = 1$, then

$$D_{\mathrm{KL}}(\mathcal{N}_{\mu_1,\Sigma_1} \parallel \mathcal{N}_{0,I}) = \frac{1}{2}\|\mu_1\|^2 - \frac{m}{2} + \frac{\mathrm{tr}(\Sigma_1)}{2} - \frac{\log\det(\Sigma_1)}{2}. \tag{B.11}$$