Information Theory Behind the Kullback-Leibler Divergence

Zhijie Chen

January 23, 2026

Why do we use the Kullback-Leibler (KL) divergence as a measure of distance between probability distributions? Why do we use the cross-entropy loss function in supervised machine learning?

Contents

1 Introduction

In supervised learning, we typically assume that the training set $\mathcal{D} = \{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$ is sampled from an underlying true distribution p, and our goal is to construct a model q_θ, parameterized by θ, that approximates this distribution. To evaluate our model, we need a measure of distance, or dissimilarity, between two probability distributions. A standard choice is the Kullback-Leibler (KL) divergence

$$D_{\mathrm{KL}}(p \parallel q) = \mathbb{E}_{x\sim p}\left[\log \frac{p(x)}{q(x)}\right] = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$$

(or $\int p(x) \log \frac{p(x)}{q(x)} \mathrm{d}x$ for continuous random variables). In practice, we often use the cross-entropy loss[1]

$$J(\theta) = -\frac{1}{N} \sum_{i=1}^{N} \log q_\theta(x^{(i)}, y^{(i)}).$$

But why do they work? The definitions above are not quite self-explanatory — it is not immediately clear from their forms why the KL divergence measures the distance between two distributions, or why minimizing the cross-entropy loss leads to better model performance. To answer this, we begin with the notion of surprisal.

2 Surprisal

Consider a random variable X with a discrete probability distribution $p(x)$. If we observe an outcome x, how surprised are we? Intuitively, the higher the probability of an event, the less surprised we are, and correspondingly the less information we gain from its occurrence. Conversely, events with low probability are more surprising and provide more information. This motivates the definition of information content, or self-information, or surprisal, of an event.

Definition 1. The information content, self-information, or surprisal of an event x is defined as

$$\mathrm{I}(x) = -\log p(x).$$

Here the choice of base b is somewhat arbitrary. Different choices of b correspond to different units of information: bit for $b = 2$, nat (for natural) for $b = e$, and Hart (for hartley) for $b = 10$.

In this article, we shall primarily be treating the concept above as surprisal, instead of the less intuitive notion of information content. The latter notion is also important and will be treated in 5.

The surprisal has some intuitive and desirable properties.

- $\mathrm{I}(x) = 0$ for x with $p(x) = 1$. We are completely unsurprised at events with probability 1. Such events carry no information.
- $\mathrm{I}(x)$ is a monotonically decreasing function of $p(x)$.
- For independent events x and y, $\mathrm{I}(x, y) = \mathrm{I}(x)\mathrm{I}(y)$. The information content of two independent events is the sum of their individual self-information; the surprisal of two independent events is the sum of their individual surprisals.

[1] The cross-entropy loss has the same form as the maximum likelihood estimation objective.

3 Entropy and Cross-Entropy

3.1 Entropy

Definition 2. The (Shannon) entropy of a random variable X, denoted $\mathrm{H}(p)$, is defined as the average surprisal we experience when drawing samples from it.

$$\mathrm{H}(p) = \mathbb{E}_{x\sim p}[\mathrm{I}(x)] = -\sum_{x\in\mathcal{X}} p(x)\log p(x)$$

(or $-\int p\log p$ for continuous distributions, called the differential entropy[2]).

Entropy measures the uncertainty inherent in a distribution. Indeed, the greater the entropy, the more surprisal we will experience (in terms of expectation) when sampling from it. Next we discuss what distributions extremize the entropy.

Discrete case, minimum Observe that the entropy is always nonnegative. Hence when the probability mass is concentrated at a single point, the entropy attains its minimum, 0.

Discrete case, maximum We use Lagrange multipliers. Without loss of generality, suppose that $\mathcal{X} = \{1, 2, \ldots, n\}$. Let $y = (y_1, \ldots, y_n) = (p(1), \ldots, p(n))$. The Lagrangian

$$\mathcal{L}(y, \lambda) = -\sum_{i=1}^{n} y_i \log y_i + \lambda\left(\sum_{i=1}^{n} y_i - 1\right).$$

The only stationary point is $y_1 = \cdots = y_n$, and the Hessian of the Lagrangian here is negative definite. Hence a uniform distribution maximizes the entropy (maximum $\log n$).

Continuous case, minimum When $X \sim U(a, b)$, $\mathrm{H}(p) = \log(b - a)$. Hence $\mathrm{H}(p) \to -\infty$ when $b - a \to 0$. There is no minimum.

Continuous case, maximum Suppose that the PDF $f(x) = 0$ when $x \notin [a, b]$. Similar to the proof of the discrete case, variational calculus (the Euler-Lagrange equation) leads to that a uniform distribution maximizes the (differential) entropy. Here we give a cleaner proof using the nonnegativity of the KL divergence (see 4.2).

Let u be the PDF of $U(a, b)$.

$$0 \le D_{\mathrm{KL}}(f \parallel u) = \int_a^b f\log f - \int_a^b f\log u = -\mathrm{H}(f) + \log(b - a).$$

Hence $\mathrm{H}(f) \le \log(b - a)$. Equality holds if and only if $f = u$ a.e.[3]

3.2 Cross-Entropy

In the context of machine learning, we often approximate an unknown, underlying true distribution p with a model q. Suppose we again want to measure the expected surprisal when sampling from the distribution. Samples are still drawn from the black box p, but this time we can only measure surprisal with our known, approximating distribution q.

Definition 3. The cross-entropy, denoted $\mathrm{H}(p, q)$, is the average surprisal experienced when samples are drawn from p, but surprisal is measured via q.

$$\mathrm{H}(p, q) = \mathbb{E}_{x\sim p}[-\log q(x)] = -\sum_{x\in\mathcal{X}} p(x)\log q(x).$$

[2] In what follows, the definitions for continuous random variables apply mutatis mutandis.

[3] a.e. stands for almost everywhere.

4 Kullback-Leibler Divergence

4.1 Motivation and Definition

How much extra surprisal do we get by using our incorrect model q instead of the truth p?

Definition 4. The Kullback-Leibler (KL) divergence from q to p, denoted $D_{\mathrm{KL}}(p \parallel q)$, is the difference between the cross-entropy and the entropy.

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

4.2 Properties

Theorem 5. *$D_{KL}(p \parallel q) \geq 0$. Equality holds if and only if $p = q$ a.e.*

Proof. We prove the discrete case using Jensen's inequality; the continuous case holds mutatis mutandis. Notice that $p(x) \geq 0$ and $\sum_{x \in \mathcal{X}} p(x) = 1$. Because log is concave, by Jensen's inequality

$$D_{\mathrm{KL}}(p \parallel q) = -\sum_{x \in \mathcal{X}} p(x) \log \frac{q(x)}{p(x)} \geq -\log \sum_{x \in \mathcal{X}} p(x) \frac{q(x)}{p(x)} = 0.$$

The condition for equality follows from that of Jensen's inequality. □

Corollary 6. $\mathrm{H}(p, q) \geq \mathrm{H}(p) = \mathrm{H}(p, p)$.

Measuring surprisal using the true distribution yields least average surprisal. Any incorrect surprisal-measuring scheme incurs extra surprisal.

The KL divergence is asymmetric: $D_{\mathrm{KL}}(p \parallel q) \neq D_{\mathrm{KL}}(q \parallel p)$ generally.

5 An Information-Theoretic Perspective

The KL divergence quantifies the "inefficiency" in using distribution q to represent the truth p. Suppose we want to design a (binary) code (like Morse code or Huffman code) to transmit outcomes of a discrete random variable $X \sim p(x)$. To be efficient, we should assign shorter codewords to more probable outcomes and longer codewords to less probable ones.

Shannon's source coding theorem implies that the theoretically optimal length of a codeword representing an outcome x is $-\log p(x)$. In this subsection we use 2 as the base of the logarithm because we are constructing a binary code.

The optimal scenario Suppose we know the true distribution p and hence use it to construct the code. The expected length of a codeword is then

$$\mathbb{E}_{x \sim p}[-\log p(x)] = \mathrm{H}(p).$$

The suboptimal scenario Now suppose we only have an approximating distribution q, and construct our code based on this false belief. The length of the codeword assigned to an outcome x would then be $-\log q(x)$. The expected length of a codeword is therefore

$$\mathbb{E}_{x \sim p}[-\log q(x)] = \mathrm{H}(p, q).$$

We see that the entropy is the average number of bits required per message when using the most efficient code, and the cross-entropy is the average number of bits required per message when using an inefficient, suboptimal code based on the wrong distribution q. Therefore, the KL divergence $D_{\mathrm{KL}}(p \parallel q) = \mathrm{H}(p, q) - \mathrm{H}(p) \geq 0$ is the average number of *extra* bits required per message due to using the wrong distribution to optimize the code. This justifies the statement that the KL divergence quantifies the inefficiency in using distribution q to represent p.

6 The Machine Learning Perspective

Now the answers to the questions raised in 1 are clear.

(1) Why do we use the KL divergence as a measure of distance between distributions?

The answer is clear from 4 and 5.

(2) Why do we use the cross-entropy as the loss function in machine learning (classification problems in particular)?

To measure the dissimilarity between the data's underlying distribution p and our model q_θ, we use the KL divergence. Minimizing $D_{\mathrm{KL}}(p \parallel q_\theta) = \mathrm{H}(p, q_\theta) - \mathrm{H}(p)$ is equivalent to minimizing $\mathrm{H}(p, q_\theta)$ because p does not depend on θ.

One might further ask that the cross-entropy $\mathrm{H}(p, q_\theta) = -\sum_{x \in \mathcal{X}} p(x) \log q_\theta(x)$ looks different from the actual loss function we see in ML textbooks, and that how can we possibly calculate it without knowing the true distribution $p(x)$? The answer is we use the empirical distribution $\hat{\mathcal{D}}$, i.e., the uniform distribution over the training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^{N}$, to approximate p. The cross-entropy (loss)

$$J(\theta) = \mathrm{H}(\hat{\mathcal{D}}, q_\theta) = -\sum_{i=1}^{N} \frac{1}{N} \log q_\theta(x^{(i)}, y^{(i)}).$$

This is the form familiar to us for long.