Information Theory Behind the Kullback-Leibler Divergence

Zhijie Chen

January 22, 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 4.3.

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 $\mathsf{H}(p)$, is defined as the average surprisal we experience when drawing samples from it.

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

(or $-\int p(x)\log p(x)\mathrm{d}x$ for continuous distributions, called the differential entropy).

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, maximum

Continuous case, minimum When $X \sim U[a, b]$, $\mathsf{H}(p) = \log(b - a)$. Hence $\mathsf{H}(p) \to -\infty$ when $b - a \to 0$. There is no minimum. But we can show that

3.2 Cross-Entropy

4 Kullback-Leibler Divergence

4.1 Motivation and Definition

4.2 Properties

4.3 An Information-Theoretic Perspective