Entropy and KL divergence

Arman Rysmakhanov, Cameron White

Introduction

In machine learning, we often need to compare how well our model's predicted probability distribution matches the true distribution of the data. Intuitive metrics, such as the Euclidean distance, are not very applicable. For example, Euclidean distance treats all differences equally, but with probabilities, the difference between 0.01 and 0.001 might be much more significant than the difference between 0.51 and 0.52. KL divergence addresses this issue by quantifying how much information we would need to fully describe the true distribution using our model's distribution. But what is information?

In everyday life, we carry a certain intuition for "information" as an informal measure of something we can know to a certain degree. In probability and statistics, we explore this intuition to develop rigorous mathematical structures surrounding the idea of information in terms of certainty and distributions. This understanding is further developed in Information Theory, which provides a backing context for this article on KL Divergence and machine learning.

Information theory is a field of probability and statistics that quantifies information and allows for its measurement and approximation. For an event x with probability $P(x)$, the mathematical definition of information for a specific event with probability P is given by the self-information function:

$$I(x) = -\log_2(P(x)).$$

One of the fundamental measures in Information Theory is entropy, which describes the average uncertainty in the outcome of a random process. Entropy is described by the equation

$$H(x) = -\sum_{x \in X} P(x) \log_2(P(x))$$

for the distribution associated with random variable $X \ni x$. Intuitively, entropy measures the amount of uncertainty associated with the value of a discrete random variable $X \ni x$ when only its distribution is known. Alternatively, the entropy associated with a certain outcome can be viewed as the "surprise" associated with that outcome.

Let us consider an example of entropy calculations with fair versus weighted dice. For a regular, fair six-sided die, we have that $p_f(x) = \frac{1}{6}$ for all $x \in [1, 6]$;

call this distribution P_f. Thus, our entropy calculation would be $H(P_f) = -6 \cdot \frac{1}{6} \log_2(\frac{1}{6}) \approx 2.58$ bits.

Now consider a weighted die following distribution P_{w1} where $p_{w1}(1) = \frac{1}{2}$ and $p_{w1}(i) = \frac{1}{10}$ for $i \in \{2, 3, 4, 5, 6\}$. Now we can see that the entropy will be $H(P_{w1}) = -(\frac{1}{2} \log_2(\frac{1}{2}) + 5 \cdot \frac{1}{10} \log_2(\frac{1}{10})) \approx 2.16$ bits.

Now we can analyze the entropy of one side of a die with an increasing number n of faces. In this case, the probability for each face is

$$p_u(x) = \frac{1}{n} \quad \text{for all} \quad x \in [1, n],$$

and the distribution is uniform. We calculate the entropy of this distribution to be:

$$H(P_u) = -\sum_{i=1}^{n} \frac{1}{n} \log_2 \left(\frac{1}{n} \right) = \log_2(n).$$

As the number of sides n increases, the entropy increases logarithmically, which aligns with our intuition. With an increasing number of sides, the number of total outcomes increases but the proportional increase slows, following the behavior of $f(x) = 1/x$.

We can graph the entropy with a fair die of increasing number of sides to visualize this pattern:

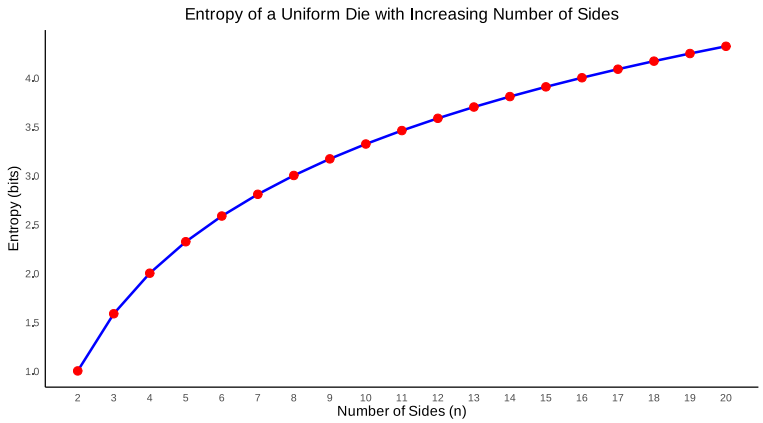
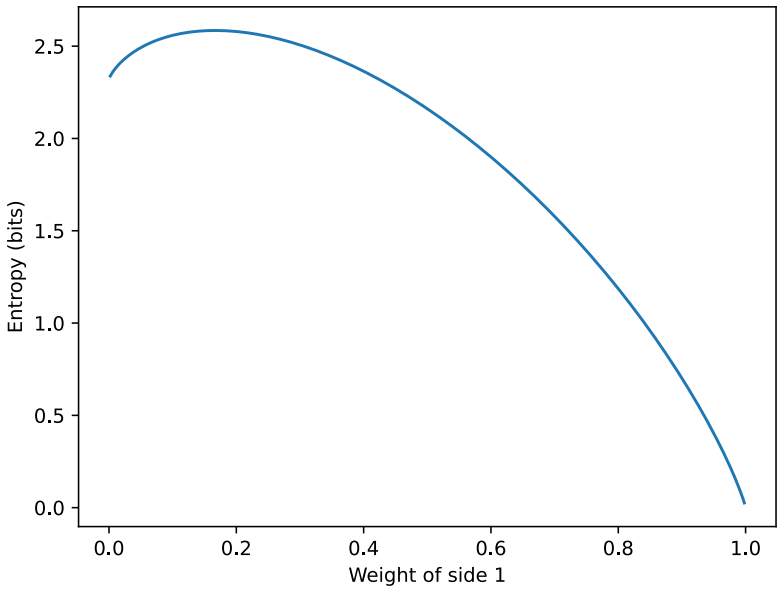


Figure 1: Logarithmic increases in entropy for an n-sided die

If we have a highly biased die, say whose rolls follow distribution P_{w2} with $p_{w2}(1) = 0.8$, $p_{w2}(i) = 0.04$ for $i \in \{2, 3, 4, 5, 6\}$, we find the entropy in this case to be $H(P_{w2}) = -(0.8 \log_2(0.8) + 5 \cdot 0.04 \log_2(0.04)) \approx 1.15$ bits.

Let's see how entropy changes with the weight of one side, while other sides' weights are equal:



The peak is achieved when the weight is $\frac{1}{6}$. In other words, you are most surprised by the result when every outcome is equally possible.

These examples provide intuition for the relationship between certainty and entropy. As the dice roll becomes more biased, our uncertainty about the outcome decreases. This decrease in uncertainty corresponds to a decrease in entropy. In other words, a biased die roll provides more information about the likely outcome, reducing the overall information content or entropy of the system.

We saw that different probability distributions result in different entropy values—but how exactly can we measure the difference between them? In general, the concept of calculating entropy for a given distribution lays the groundwork for quantifying differences in information between distributions. If we have a certain defined set of events, and two different distributions to describe this set of events under a random variable, we want to find some way to measure distance between the two possible distributions.

Kullback-Liebler divergence provides one way of measuring this distance. Specifically, KL divergence compares the similarity of a given "new" probability distribution $Q(x)$ against an assumed "true" probability distribution $P(x)$. In this manner, KL divergence presents an asymmetrical comparison method in terms of information difference (or "divergence") for two distributions. In a statistical sense, given some data that follows distribution $P(x)$, if we apply the distribution $Q(x)$ to the data, the KL divergence of $P(x)$ given $Q(x)$ will measure the the number of average additional bits per datum necessary to compress the

data into the distribution described by $Q(x)$. The formula for KL divergence is as follows:

$$D_{KL}(P(x)||Q(x)) = \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}.$$

To make sense of this formulation, we can understand KL divergence from an entropy perspective. KL divergence is also known as "relative entropy" and describes the amount of extra entropy introduced by assuming the fit of distribution $Q(x)$ to a system which is described by the true distribution $P(x)$. In terms of entropy, we write KL divergence as

$$D_{KL}(P(x)||Q(x)) = H(P(x), Q(x)) - H(P(x)),$$

where $H(P(x), Q(x))$ is the cross-entropy, given by:

$$H(P, Q) = -\sum_{x \in X} P(x) \log Q(x).$$

The cross-entropy describes the average number of bits needed to identify a given event where the set of outcomes are fitted to approximating/estimating probability distribution $Q(x)$, rather than the true distribution $P(x)$. Using entropy, we can write KL divergence as

$$D_{KL}(P(x)||Q(x)) = H(P, Q) - H(P).$$

Now we see that

$$\begin{aligned} D_{KL}(P(x)||Q(x)) &= H(P, Q) - H(P) \\ &= -\sum_{x \in X} P(x) \log Q(x) + \sum_{x \in X} P(x) \log P(x) \\ &= \sum_{x \in X} P(x) \log \frac{P(x)}{Q(x)}. \end{aligned}$$

It is important to note that KL divergence is not a true distance metric as it fails to satisfy the triangle inequality.

Properties

Asymmetry

Note that KL divergence is not symmetric, that is, $D_{KL}(P(x)||Q(x)) \neq D_{KL}(Q(x)||P(x))$. In other words, the terms for distributions P and Q are not interchangeable. Intuitively, this makes sense: the expected surprise from using Q to predict "true" distribution P is not necessarily the same as the the expected surprise from using P to predict "true" distribution Q. Switching the true and assumed distributions P and Q re-frames the divergence measure in terms of the base distribution, making the measurement asymmetrical due to the differing weights assigned to the various events in both distributions. For example, if $P(x)$ is heavily concentrated on a single event while $Q(x)$ is more spread out, we will have that $D_{KL}(P(x)||Q(x)) > D_{KL}(Q(x)||P(x))$.

High-probability bias

Since the formula for KL divergence consists of a summation over expectation terms of the form $P(x) \log \frac{P(x)}{Q(x)}$, we can see that the logarithmic difference terms are not weighted equally for all probabilities in the true distribution $P(x)$.

Specifically, for large $P(x)$, we will have large $P(x) \log \frac{P(x)}{Q(x)}$, which means terms of this form will contribute more to the sum.

This property is a feature of KL divergence in which, unlike in some other measures of probability distance, terms of higher probabilities in the true distribution have a greater impact on the information gain from the assumed distribution.

We can formalize this property by taking the partial derivative of expectation terms with respect to $P(x)$:

$$\frac{\partial}{\partial P(x)} \left[P(x) \log \frac{P(x)}{Q(x)} \right] = \log \frac{P(x)}{Q(x)} + 1.$$

We see that when $P(x) > Q(x)$, the derivative is positive, meaning that increasing the probability of an event under P increases its contribution to the KL divergence. Therefore, higher values of $P(x)$ correspond to a great effect of the ratio between $P(x)$ and $Q(x)$ on the overall KL divergence measure.

Non-negativity

Having explored KL divergence's bias towards P, you might intuitively suspect that it is always non-negative. Indeed, when $P(x)$ is large relative to $Q(x)$, the term $P(x) \log \frac{P(x)}{Q(x)}$ is a large positive value. While there are negative contributions when $Q(x) > P(x)$, the bias suggests that they might be outweighed by the positive terms.

This intuition turns out to be correct. We can prove D_{KL} to be always non-negative using Jensen's inequality.

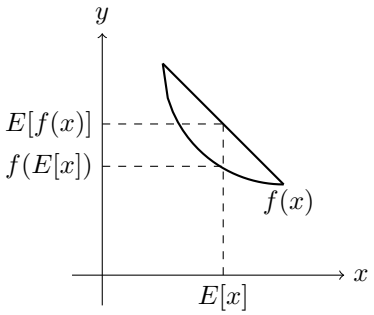


Figure 2: Visual illustration of Jansen's inequality for a convex function f.

Remark (Jensen's inequality)**.** For a convex function $f : \mathbb{R} \to \mathbb{R}$ and a probability space $(\Omega, \mathcal{F}, P)$, Jensen's inequality states that for any integrable random variable X:

$$\mathbb{E}[f(X)] \geq f(\mathbb{E}[X])$$

In other words, the function of an average is less than or equal to the average of the function.

Visually, it means that, for any two distinct point on the function, the line between them will always lie above the function. This powerful property will help us establish non-negativity.

$$D_{KL}(P(x)||Q(x)) = \sum_x P(x) \log \frac{P(x)}{Q(x)} = -\sum_x P(x) \log \frac{Q(x)}{P(x)}$$

Since $-\log$ is convex, and D_{KL} is an expectation, i.e. 'average', of the logarithmic distance between P and Q, we can apply Jensen's inequality:

$$-\sum_x P(x) \log \frac{Q(x)}{P(x)} \geq -\log \sum_x P(x) \frac{Q(x)}{P(x)} = -\log \sum_x Q(x) = -\log(1) = 0$$

Non-negativity of D_{KL} has important implications. First, it allows us to interpret KL divergence as a measure of dissimilarity of P and Q, even though it is asymmetric. In machine learning, non-negativity is desirable since it provides a natural objective function for optimization: when training models to match target distributions, we know that a KL divergence of zero indicates perfect matching, while positive values indicate room for improvement, allowing us to iteratively minimize D_{KL}.

$$D_{KL} = 0 \iff P = Q$$

There is one more concern that we need to address. Since D_{KL} operates in terms of information, one could think that, for a given P, there is Q such that the terms of the sum cancel each other out, thus resulting in a "distance" of 0 for different distributions. Fortunately, we can show that $D_{KL}(P(x)||Q(x)) = 0$ if and only if $P(x) = Q(x)$.

Assume we know $D_{KL}(P(x)||Q(x)) = \sum_x P(x) \log \frac{P(x)}{Q(x)} = 0$. Using Jensen's inequality and the strict convexity of $-\log(x)$,

$$0 = -\sum_x P(x) \log \frac{Q(x)}{P(x)} \geq -\log \sum_x P(x) \frac{Q(x)}{P(x)} = -\log \sum_x Q(x) = 0.$$

Equality holds if and only if $\frac{Q(x)}{P(x)}$ is constant for all x where $P(x) > 0$, therefore $Q(x) = kP(x)$, where k is the constant in question. Since $1 = \sum_x Q(x) = k \sum_x P(x) = k$, $P(x) = Q(x)$.

Assume now that we know $P(x) = Q(x)$. Then,

$$D_{KL}(P(x)||Q(x)) = \sum_x P(x) \log \frac{P(x)}{Q(x)} = \sum_x P(x) \log(1) = 0$$

Therefore, $D_{KL}(P(x)||Q(x)) = 0$ if and only if $P(x) = Q(x)$.

Approximation

Monte Carlo Estimation

In machine learning, computing KL divergence directly is often undesirable. If a model has a lot parameters, we need to come up with efficient approximations. A good estimator would have both low bias and variance. Low bias means producing results that are close the actual distribution on average, which is important for accuracy. Low variance means producing results that are not too distant from the mean, which is important to decrease the number of samples needed to accurately estimate KL divergence.

One such approach is Monte Carlo estimation, as outlined by John Schulman. Instead of evaluating the expectation over all the possible values, we choose samples from the distribution and average over them,

Starting with the standard form of KL divergence:

$$D_{KL}(P(x)||Q(x)) = \mathbb{E}_{x\sim P(x)} \left[\log \frac{P(x)}{Q(x)}\right]$$

The naive Monte Carlo estimator simply draws samples from p(x) and averages:

$$\frac{1}{N} \sum_{i=1}^{N} \log \frac{P(x_i)}{Q(x_i)}$$

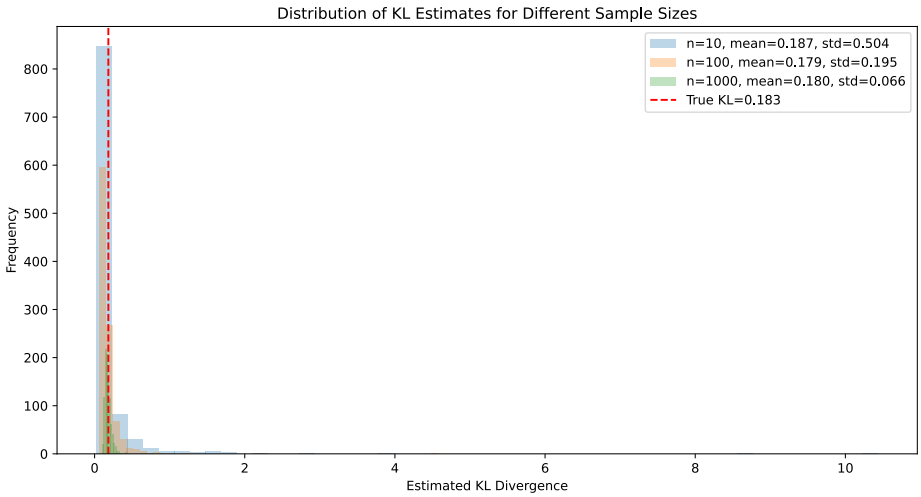
While this estimator is unbiased, it can have high variance. We can derive a better estimator using control variates. The idea is to add a term with zero expectation that is negatively correlated with our original estimator to reduce variance.

Let $r = \frac{Q(x)}{P(x)}$. Since q is a probability distribution, we know that $\mathbb{E}_{x\sim P(x)}[r] = \sum_x P(x)\frac{Q(x)}{P(x)} = 1$, and therefore $\mathbb{E}_{x\sim P(x)}[r - 1] = 0$. This means that for any constant k, $-\log(r) + k(r - 1)$ is an unbiased estimator of $D_{KL}(P(x)||Q(x))$.

We also know that $\log(x) \leq x - 1$ for all $x > 0$. Therefore, setting $k = 1$ guarantees our estimator is non-negative. This gives us our improved estimator:

$$\frac{1}{N} \sum_{i=1}^{N} \left[\frac{Q(x_i)}{P(x_i)} - 1 - \log \frac{Q(x_i)}{P(x_i)}\right] = \frac{1}{N} \sum_{i=1}^{N} [r_i - 1 - \log r_i]$$

This estimator has lower variance than the naive estimator. When $r > 1$, the first term $r - 1$ is positive while $-\log r$ is negative (and vice versa when $r < 1$), causing these terms to partially cancel out and reduce the overall variance. As a result, it is close to the true KL in practice:



Conclusion

In this exposition, we have provided an introduction to entropy and KL divergence. Beginning with entropy to quantify uncertainty, we developed KL divergence as a natural measure of the "difference" of probability distributions. Approximation methods, exemplified here by Monte Carlo estimation with variance reduction, bridges the gap between theoretical elegance and practical implementation in machine learning applications. We hope this article serves as an accessible entry point for those who want to learn the theoretical part of machine learning.