Information Theory and Signal Processing

Muchang Bahng

Spring 2024

Contents

This final section does not relate directly to the workflow of training a machine learning model. Rather, it provides some very nice tools for us when analyzing the performance of these models.

First, we want to quantitatively measure the “surprise" of an event E happening in a probability space by assigning it a value $I(E)$. We want it to satisfy the following:

1. $I(E) \geq 0$. The surprisal of any event is nonnegative.
2. $I(E) = 0$ iff $\mathbb{P}(E) = 1$. No surprisal is gained from events with probability 1.
3. If E_1 and E_2 are independent events, then $I(E_1 \cap E_2) = I(E_1) + I(E_2)$. The information from two independent events should be the sum of their informations.
4. I should be continuous, i.e. slight changes in probability correspond to slight changes in surprisal.

Definition 0.1 (Surprisal)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, the **surprisal**, or **self-information**, of an event $E \in \mathcal{F}$ is

$$\sigma_{\mathbb{P}}(E) :== -\log \mathbb{P}(E)$$

and the **expected surprisal** of E is

$$h_{\mathbb{P}}(E) = \mathbb{P}(E)\sigma_{\mathbb{P}}(E)$$

Now we can define entropy as the expected surprisal of a random variable, which seems now more motivated and intuitive.

Definition 0.2 (Entropy)

Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a $\mathbb{P}$-almost partition is a set family $\mathcal{G} \subset \mathcal{F}$ such that $\mu(\cup_{G \in \mathcal{G}} G) = 1$ and $\mathbb{P}(A \cap B) = 0$ for all distinct $A, B \in \mathcal{G}$ (this is a relaxation of the usual conditions for a partition). The **entropy** of the subfamily $\mathcal{G}$ is

$$H_{\mathbb{P}}(\mathcal{G}) := \sum_{G \in \mathcal{G}} h_{\mathbb{P}}(G)$$

The **entropy** of the σ-algebra $\mathcal{F}$ is defined

$$H_{\mathbb{P}}(\mathcal{F}) = \sup_{\mathcal{G} \subset \mathcal{F}} H_{\mathbb{P}}(\mathcal{G})$$

Example 0.1 ()

For a discrete random variable, since we are working with its power set, the entropy reduces to

$$H[X] := \mathbb{E}[-\ln p(X)] = -\sum_x \mathbb{P}(X = x) \ln \mathbb{P}(X = x)$$

Intuitively, this represents the element of surprise of a certain data point, and distributions that have relatively sharp peaks will have lower entropy (since we expect most of the samples to come from the peaks) while uniform distributions have higher entropy. The entropy also demonstrates the average length (if base is 2) number of bits required to transmit the state of a random variable.

Definition 0.3 (Joint, Conditional Entropy)

We can define the joint entropy and conditional entropy between two discrete random variables X, Y

as

$$H(X,Y) = \mathbb{E}_{X \times Y}[-\log \mathbb{P}(X = x, Y = y)]$$
$$H(X \mid Y) = \mathbb{E}[-\log \mathbb{P}(X = x \mid Y = y)]$$

0.0.1 Kullback Leibler Divergence

The **relative entropy**, or **Kullback-Leibler divergence**, of distributions $p(x)$ and $q(x)$ is defined

$$\begin{aligned} \mathrm{KL}(p||q) &:= -\int p(\mathbf{x}) \, \ln q(\mathbf{x}) \, d\mathbf{x} - \left(-\int p(\mathbf{x}) \, \ln p(\mathbf{x}) \, d\mathbf{x} \right) \\ &= -\int p(\mathbf{x}) \, \ln \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} \right) d\mathbf{x} \end{aligned}$$

We can show that this quantity is always greater than or equal 0 by Jensen's inequality using the fact that $-\ln(x)$ is concave

$$\int p(\mathbf{x}) \, -\ln \left(\frac{q(\mathbf{x})}{p(\mathbf{x})} \right) d\mathbf{x} \geq -\ln \int p(\mathbf{x}) \, \frac{q(\mathbf{x})}{p(\mathbf{x})} \, d\mathbf{x} = -\ln \int q(\mathbf{x}) \, d\mathbf{x} = -\ln(1) = 0$$

and it is precisely 0 if $p = q$, so it behaves similarly to a metric. However, it isn't exactly since it is not symmetric.

Let's demonstrate how entropy and the KL divergence applies to maximum likelihood estimation. Suppose that iid samples $\mathcal{D} = \{(x^{(n)}, y^{(n)}\}$ are given in a regression problem. Let $P^* = (X, Y)$ be the true data generating function. Then, we want to compute an approximation of P^* with P_θ, where P_θ is some parameterized distribution. The negative log likelihood of the y's being generated is

$$\ell(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log P_\theta(y_i \mid x_i)$$

which asymptotically converges to

$$\mathbb{E}_{P^*}[-\log P_\theta(y_i \mid x_i)] = \mathrm{KL}(P^* || P) + H[P^*]$$

and since the entropy is constant, this is equivalent to minimizing the KL divergence between P and P^*.

We assume that the $y^{(n)}$'s come from a conditional distribution P_{θ, x_i}, where the parameters of the distribution is θ and x_i

0.0.2 Mutual Information

Definition 0.4 (Differential Entropy)

For a continuous random vector, the **differential entropy** is defined

$$H[\mathbf{X}] = -\int p(\mathbf{x}) \ln p(\mathbf{x}) \, d\mathbf{x}$$

Definition 0.5 (Mutual Information)

The **mutual information** between random variables X, Y is the decrease in entropy when we condition X by Y.

$$I(X;Y) = H(X) - H(X \mid Y) = H(Y) - H(Y \mid X)$$

This can be conditioned on another random variable Z.

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid Y, Z) = H(Y \mid Z) - H(Y \mid X, Z)$$