Notes on Information Theory and Statistics

Rom Parnichkun

October 28, 2022

1 Background

1.1 Definition

Consider the *probability space*:

$$(x, y, \mu_i), \quad i = 1, 2. \tag{1}$$

x are the elements, y are the events (which is made up of elements), for which two hypotheses μ_1 and μ_2 are defined.

As an example, x may be the occurrence or non-occurrence of a signal pulse, and y may be a collection of possible sequences of a certain length of pulse and no pulse.

We assume that μ_1 and μ_2 are *absolutely continuous*, meaning that there exists no event E such that $\mu_1(E) = 0$ and $\mu_2(E) \neq 0$ and vice versa.

Let λ be a probability measure such that:

$$\mu_i(E) = \int_E f_i(x) d\lambda(x), \quad i = 1, 2, \tag{2}$$

where $0 < f_i(x) < \infty$.

$$d\mu_i(x) = f_i(x) d\lambda(x) \to f_i(x) = \frac{d\mu_i}{d\lambda}. \tag{3}$$

Given two hypotheses H_1 and H_2, the probability of a hypothesis given data x is as follows:

$$P(H_i \mid x) = \frac{P(H_i) f_i(x)}{P(H_1) f_1(x) + P(H_2) f_2(x)}, \quad i = 1, 2, \tag{4}$$

from which we obtain

$$f_i(x) = \frac{P(H_i \mid x)\big(P(H_1) f_1(x) + P(H_2) f_2(x)\big)}{P(H_i)}, \tag{5}$$

therefore,

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1 \mid x)}{P(H_2 \mid x)} - \log \frac{P(H_1)}{P(H_2)}. \tag{6}$$

It could be noted that in the equation above, the difference between the logarithm of the odds in favor of H_1 after the observation of $X = x$ and before the observation may be considered as the information resulting from the observation $X = x$.

The *mean information for discrimination in favor of* H_1 can be formulated as follows:

$$\begin{aligned} I(1:2) &= \int \log \frac{f_1(x)}{f_2(x)} d\mu_1(x) && (7) \\ &= \int f_1(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x) && (8) \\ &= \int \log \frac{P(H_1 \mid x)}{P(H_2 \mid x)} d\mu_1(x) - \log \frac{P(H_1)}{P(H_2)} && (9) \end{aligned}$$

1.2 Divergence

Following the previous section, we may define

$$I(2:1) = \int f_2(x) \log \frac{f_2(x)}{f_1(x)} d\lambda(x) \tag{10}$$

as the *mean information per observation from* μ_2 *for discrimination in favor of* H_2 *against* H_1, and

$$-I(2:1) = \int f_2(x) \log \frac{f_1(x)}{f_2(x)} d\lambda(x) \tag{11}$$

as the *mean information per observation from* μ_2 *for discrimination in favor of* H_1 *against* H_2.

We now define the *divergence* $J(1, 2)$ by

$$\begin{aligned} J(1,2) &= I(1:2) + I(2:1) && (12) \\ &= \int \big(f_1(x) - f_2(x)\big) \log \frac{f_1(x)}{f_2(x)} d\lambda(x) && (13) \\ &= \int \log \frac{P(H_1 \mid x)}{P(H_2 \mid x)} d\mu_1(x) - \int \log \frac{P(H_1 \mid x)}{P(H_2 \mid x)} d\mu_2(x), && (14) \end{aligned}$$

which can be said as the *total discrimination of one hypothesis over another.*

1.3 Entropy

Suppose H_2 is a hypothesis which must be true, meaning $P(H_2) = 1$, and $H_1 \in H_2$. We can compute the information for discrimination in favor of H_1 as:

$$\log \frac{f_1(x)}{f_2(x)} = \log \frac{P(H_1 \mid x)}{1} - \log \frac{P(H_1)}{1}. \tag{15}$$

If $P(H_1 \mid x) = 1$, information gain from the observation is $-\log P(H_1)$.

To carry this notion further, suppose a set of mutually exclusive and exhaustive hypotheses $H_1, H_2, \ldots, H_n$ exists and that we can infer which hypothesis is true from the observation. The mean information in an observation about the hypotheses is the mean value of $-\log P(H_i)$, $i = 1, \ldots, n$.

This expression is also called entropy, and it can be defined as

$$-\sum_{i=1}^{n} P(H_i) \log P(H_i). \tag{16}$$

2 Properties of Information

2.1 Additivity

$I(1:2)$ is additive for independent random events; that is, for X and Y independent under H_i, $i = 1, 2$,

$$I(1:2; X, Y) = I(1:2; X) + I(1:2; Y). \tag{17}$$

2.2 Convexity

$I(1:2)$ is almost positive definite; that is, $I(1:2) \geq 0$, with equality if and only if $f_1(x) = f_2(x)$.