David Bridston Osteyee
Irving John Good

Information, Weight of Evidence, the Singularity Between Probability Measures and Signal Detection

Lecture Notes in Mathematics
376

Springer

References for Section 1.2 are, for example, Good (1950) and (1956).

1.3 Mutual Information Between Events

The mutual information between A and B is defined as

$$I(A:B) = I(A) - I(A|B) = I(A) + I(B) - I(AB)$$
$$= I(B) - I(B|A) = I(B:A)$$

from (1.2.1) provided that $P(A) > 0$ and $P(B) > 0$. Thus $I(A:B)$ is symmetric in A and B.

Intuitively, this is a measure of the decrease (if positive) or increase (if negative) in the uncertainty about the occurrence of A (or B) caused by the occurrence of B (or A). This can also be thought of as a measure of the decrease (if positive) or the increase (if negative) in the gain in information when A (or B) occurs, caused by the occurrence of B (or A).

Therefore, $I(A:B)$ can be thought of as a measure of the positive or negative information concerning the occurrence of A (or B) provided by the occurrence of B (or A). $I(A:B)$ can also be expressed as

$$I(A:B) = \log \frac{P(A|B)}{P(A)} = \log \frac{P(AB)}{P(A)P(B)} = \log \frac{P(B|A)}{P(B)} . \qquad (1.3.1)$$

It follows that $I(A:B)$ is positive if and only if $P(A|B) > P(A)$ or equivalently $P(B|A) > P(B)$.

1.3.1 Properties of Mutual Information

Some of the properties of $I(A:B)$ are as follows.

(i) If A and B are independent then $I(A:B) = 0$ since the occurrence of B gives us no information about the occurrence of A.

(ii) If $B \subseteq A$ then $I(A:B) = I(A)$ since the occurrence of B provides all the information about the occurrence of A.

(iii) If $P(A) > 0$, $P(B) > 0$, and $P(AB) = 0$ then $I(A:B) = -\infty$ which would be the case if A and B were mutually exclusive.

For a further discussion of the ideas in Section 1.3 see Good (1961a).

1.4 Weight of Evidence

The expression "weight of evidence" has been used independently by Good (1950), Peirce (1878), and Minsky and Selfridge (1961). Most of the results of Section 1.4 are from Good (1950) and Kullback (1959). Weight of evidence will now be defined.

1.4.1 Definitions and Expressions

Let H_1 and H_2 be two competing hypotheses related to some evidence B. H_1 and H_2 may be thought of as "events" both in the a priori probability spaces (Ω,H,P) and (Ω,H,P_B), where P_B is the conditional probability given B. Then the weight of evidence in favor of H_1 as opposed to H_2, provided by B, may be defined as

$$W(H_1/H_2:B) = \log \frac{O(H_1/H_2|B)}{O(H_1/H_2)} , \qquad (1.4.1.1)$$

where $O(H_1/H_2|B)$ is the odds in favor of H_1 as opposed to H_2 given B, and $O(H_1/H_2)$ is the odds in favor of H_1 as opposed to H_2. Weight of evidence may be positive or negative.

Another expression for $W(H_1/H_2:B)$ is, from (1.4.1.1) and (1.3.1),

$$W(H_1/H_2:B) = \log \frac{P(H_1|B)}{P(H_2|B)} - \log \frac{P(H_1)}{P(H_2)}$$
$$= \log \frac{P(H_1|B)}{P(H_1)} - \log \frac{P(H_2|B)}{P(H_2)}$$
$$= I(H_1:B) - I(H_2:B). \qquad (1.4.1.2)$$

Therefore, the weight of evidence may be interpreted as the difference in information about H_1 compared to H_2 provided by B.

1.4.2 Another Expression for the Weight of Evidence

If W is interpreted as the difference in information about B provided by

H_1 compared to H_2 then it can be expressed from equation (1.4.2.1) as follows, where B is an event in the probability spaces $(\Omega,\mathcal{B},P_{H_1})$ and $(\Omega,\mathcal{B},P_{H_2})$.

$$W(H_1/H_2{:}B) = \log \frac{P(B|H_1)}{P(B)} - \log \frac{P(B|H_2)}{P(B)} = \log \frac{P(B|H_1)}{P(B|H_2)} \qquad (1.4.2.2)$$

which is the log of the likelihood ratio when the hypotheses are simple. Otherwise the non-Bayesian does not necessarily regard the probabilities $P(B|H_1)$ and $P(B|H_2)$ as meaningful. It is historically interesting that the expression "weight of evidence", in its technical sense, anticipated the term "likelihood" by over forty years.

W may also be expressed as

$$W(H_1/H_2{:}B) = I(B{:}H_1) - I(B{:}H_2) = I(B|H_2) - I(B|H_1) \qquad (1.4.2.2)$$

because

$$I(B{:}H_i) = I(B) - I(B|H_i) \quad (i = 1, 2)$$

from Section 1.3.

1.4.3 A Special Case

When $H_2 = H_1^C$ (the complement of H_1), then $W(H_1/H_2{:}B)$ may be abbreviated by $W(H_1{:}B)$ and the expression in (1.4.1.1) becomes

$$W(H_1{:}B) = \log \frac{O(H_1|B)}{O(H_1)} = \log \frac{P(B|H_1)}{P(B|H_1^C)},$$

where $O(H_1|B)$ is the odds in favor of H_1 given B and $O(H_1)$ is the odds in favor of H_1. Good (1950, p. 63, and 1969, p. 25) calls $\frac{O(H_1|B)}{O(H_1)}$ the "Bayes-Jeffreys-Turing factor" in favor of H_1 provided by B.

II. ENTROPY

The entropy I(X) of a random variable or random vector is discussed in this chapter. For the discrete case, it is shown that the entropy may be finite, even when the random variable takes on a denumerable number of values, as shown in a quantum mechanics model in Section 2.1.2. For the continuous case, however, the entropy is always infinite or undefined. Conditional entropy is also discussed.

2.1 Entropy of Discrete Random Variables

Let X be a discrete random variable. Then the expected gain in information from a single observation (or the expected uncertainty before an observation is taken) is

$$I(X) = E[I(x)] = -\sum_{i=1}^{\infty} p_i \log p_i ,$$

[Shannon (1948, p. 50)], where E stands for the expected value operation and $p_i = P[X = x_i]$ $(i = 1, 2, \ldots)$. I(X) is called the entropy of X since it is analogous to the average amount of disorder in a system of possible states in thermodynamics. [See, for instance, Slater (1939, p. 33).]

2.1.1 Properties of Entropy

Most of these properties are from Shannon (1948, pp. 49-52) except for (i), (vii) and (x).

(i) If the variable takes a denumerable number of values, the entropy may be infinite. [Balakrishnan (1968, p. 193).]

(ii) If the variable takes a finite number of values, then the entropy is a maximum when all the p_i's are equal.

(iiia) I(X) is a continuous function of the p_i's.

(iiib) If all the p_i's are equal to $\frac{1}{n}$ then

$$I(X) = -\sum_{i=1}^{n} \frac{1}{n} \log \frac{1}{n} = \log n$$