International Journal of Business Analytics & Intelligence
10 (1) 2022, 01-03
http://publishingindia.com/ijbai/

Analytically Yours: Information and Dependence

Arnab Kumar Laha*

Introduction

The field of information theory has found interesting applications in many different areas. While information theory's primary focus was on data compression and transmission, its theories have now been applied to many other fields, including probability theory, statistical inference, complexity theory, ergodic theory and machine learning. The two key ideas of entropy and mutual information, which are functions of the underlying probability distribution generating the data, have proved to be immensely useful. Entropy is now routinely used as a criterion for deciding the optimal splitting of nodes while constructing a decision tree. In this article, we briefly introduce these important ideas with a specific focus on identifying dependence between two random variables. It is well known that Pearson's linear correlation coefficient (r) accounts only for linear relationships and that the value of this correlation coefficient can be misleading in the presence of non-linear dependence. In fact, it is easy to construct examples, where $r = 0$ even though Y is a function of X. In contrast, mutual information is capable of taking into account all types of dependence.

The *entropy* of a random variable X with a probability mass function (pmf) $p(x)$ is defined as

$$H(X) = -\sum_{x} p(x) \log_2 p(x) = E(-\log_2 p(x))$$

Here, we have used logarithms to base 2 and in this case, the unit of entropy is called *bits*. If instead the natural logarithm is used, the unit of entropy is then called *nats*. The entropy is a measure of the average uncertainty in the random variable.

Consider the toss of a coin, let the outcome "Head" be coded as 1 and "Tail" be coded as 0. Let the probability of the coin turning up "Head" be p. Then the pmf of X is given in Table 1.

Table 1

Outcome	*0(Tail)*	*1(Head)*
Probability	1-p	p

The entropy of the random variable, X, is then $H(X) = -(p \log_2 p + (1-p)\log_2(1-p))$. Using the convention $0.\infty = 0$, we obtain the graph given in Fig. 1, where the x-axis represents the values of p and the y-axis represents the values of H(X).

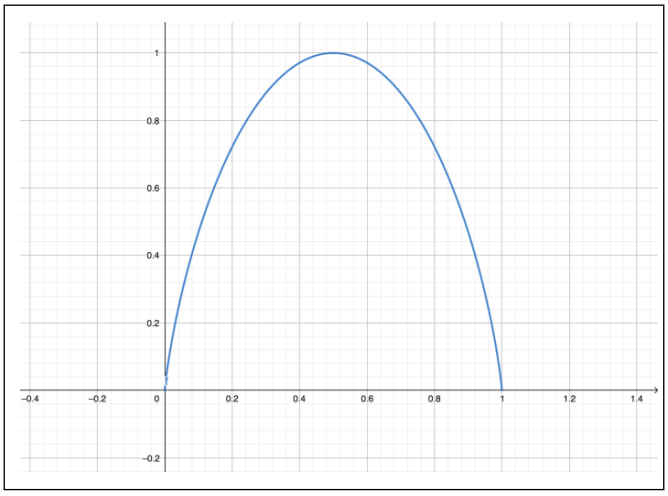

Fig. 1

From this graph, we see that the entropy is maximum when p = 0.5, i.e. when the coin is a fair coin with an equal chance of turning up Head or Tail. Intuitively, this is expected, since in this situation, the uncertainty is maximum. As the value of p deviates from 0.5 uncertainty decreases as one of the outcomes, Head or Tail, becomes

* Indian Institute of Management Ahmedabad, Gujarat, India. Email: arnab@iima.ac.in

more probable than the other. Of course, when p=0 or p=1, there is no uncertainty and the value of the entropy is then 0.

Entropy is the uncertainty of a single random variable. We can define *conditional entropy H(X/Y)*, which is the entropy of a random variable (X) conditional on the knowledge of another random variable (Y). The reduction in uncertainty of the random variable (X) due to another random variable (Y) is called the *mutual information*.

The mutual information $I(X; Y)$ is a measure of the dependence between the two random variables X and Y. For two random variables, X and Y with joint pmf $p_{X,Y}$ and marginal pmfs p_X and p_Y, the mutual information $I(X; Y)$ is defined as

$$I(X;Y) = H(X) - H(X \mid Y) = \sum_{x,y} p_{X,Y}(x,y) \log_2 \frac{p_{X,Y}(x,y)}{p_X(x)p_Y(y)}$$

We use the conventions $0 \log \frac{0}{0} = 0, \log \frac{0}{q} = 0$ and $p \log \frac{p}{0} = \infty$ if $p > 0$.

The mutual information $I(X; Y)$ is symmetric in X and Y and is always nonnegative. It is equal to zero if and only if X and Y are independent random variables. Also, note that $I(X; Y) = H(X)$.

The *joint entropy* of the random variables X and Y is defined as

$$H(X,Y) = -\sum_{x,y} p_{X,Y}(x,y) \ \log_2 p_{X,Y}(x,y)$$ The following can be shown using the simple algebra:

$$I(X; Y) = H(X) + H(Y) - H(X, Y).$$

Since $I(X; Y) \geq 0$, we have two interesting results follow immediately:

- *Conditioning reduces Entropy*: $H(X/Y) \leq H(X)$ and
- *Independence bound on Entropy*: $H(X/Y) \leq H(X) + H(Y)$.

Mutual information turns out to be a special case of a more general quantity called *relative entropy (also known as. Kullback–Leibler divergence)* $D(p||q)$, which is a measure of the "distance" between two probability mass functions p and q. It is defined as

$$D(p \,||\, q) = \sum_x p(x) \ \log \frac{p(x)}{q(x)}$$

$D(p||q)$ is always nonnegative and is zero if and only if p = q. It is easy to observe that $I(X; Y) = \log_2 e \ D(p_{X,Y} \,||\, p_X p_Y)$.

As an example, let us compute the mutual information of two random variables X and Y having the joint pmf given in Table 2. Here,

$$I(X;Y) = 0.15 \log_2 \frac{0.15}{0.55 \times 0.4} + 0.25 \log_2 \frac{0.25}{0.45 \times 0.4}$$

$$+ 0.4 \log_2 \frac{0.4}{0.55 \times 0.6} + 0.2 \log_2 \frac{0.2}{0.45 \times 0.6} = 0.06$$

Table 2

$Y \downarrow / X \rightarrow$	0	1	p_Y
0	0.15	0.25	0.4
1	0.4	0.2	0.6
p_X	0.55	0.45	

Up to this point, we have discussed entropy and mutual information in the context of discrete distributions. Can these ideas be extended to continuous distributions? The answer is Yes, if these continuous distributions have probability density function (pdf), which most commonly occurring continuous probability distributions do. For the continuous case, the entropy is defined as

$$H(X) = -\int_{-\infty}^{\infty} f(x) \ \log f(x) \ dx$$

where f is the pdf of X. For two random variables, X and Y with joint pdf $f_{X,Y}$ and marginal pdf's f_X and f_Y respectively the mutual information is defined as

$$I(X;Y) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x,y) \ \log \frac{f_{X,Y}(x,y)}{f_X(x) f_Y(y)} \ dx dy.$$

As before it is easy to observe that $I(X; Y) = 0$. if and only if X and Y are independent random variables.

As an example, let us consider the $(X,Y) \sim$ BVS $(0,0,1,1,p)$. Straight forward computations yield $I(X;Y) = -\frac{1}{2} \log(1 - \rho^2)$. Thus, when $p = 0$ (i.e. when X and Y are independent) we have $I(X; Y) = 0$ as expected and when $p^2 \rightarrow 1$ then $I(X; Y) \rightarrow \infty$. Thus, a large value of $I(X; Y)$ is indicative of strong dependence between the random variables X and Y. Figure 2 gives the plot of $I(X; Y)$ (y-axis) and p (x-axis).

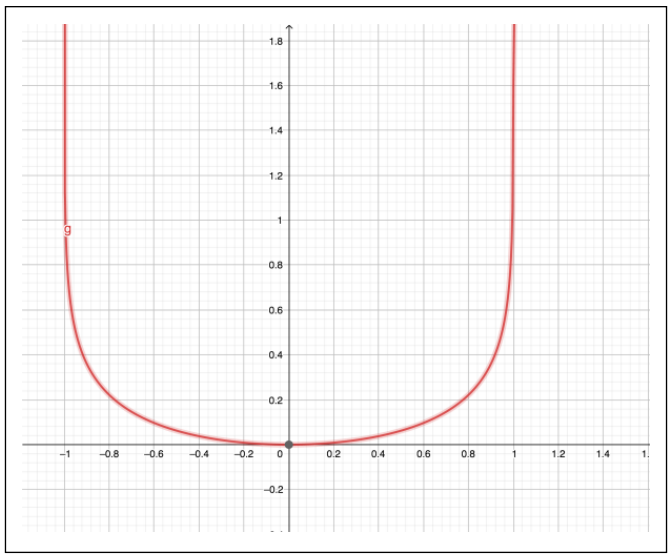

Fig. 2

Linfoot (1957) defined a correlation measure between two random variables X and Y, which is based on mutual information and called it *informational coefficient of correlation* (r_j). It is defined as

$$r_I = \sqrt{1 - e^{-2I(X;Y)}}$$

It is easy to observe that $0 \le r_1 \le 1$ with $r_1 = 0$ if and only if X and Y are independent random variables. Returning to our example, where $(X, Y) \sim \text{BVS}(0,0,1,1,p)$ we find that $r_1 = |p|$. Being based on mutual information, $r_I r_I$ is able to take into account both linear and non-linear dependence between the random variables X and Y.

If you find the topic of information theory interesting and would like to know more, you may find the book a good place to start (Cover & Thomas, 2006).

References

Linfoot, E. H. (1957). An informational measure of correlation. *Information and Control, 1,* 85-89.

Cover, T. M., & Thomas, J. A. (2006). *Elements of information theory* (2nd ed.). Wiley - Interscience.