Article from

Predictive Analytics and Futurism

June 2017
Issue 15

Maximal Information Coefficient: An Introduction to Information Theory

By Bryon Robidoux

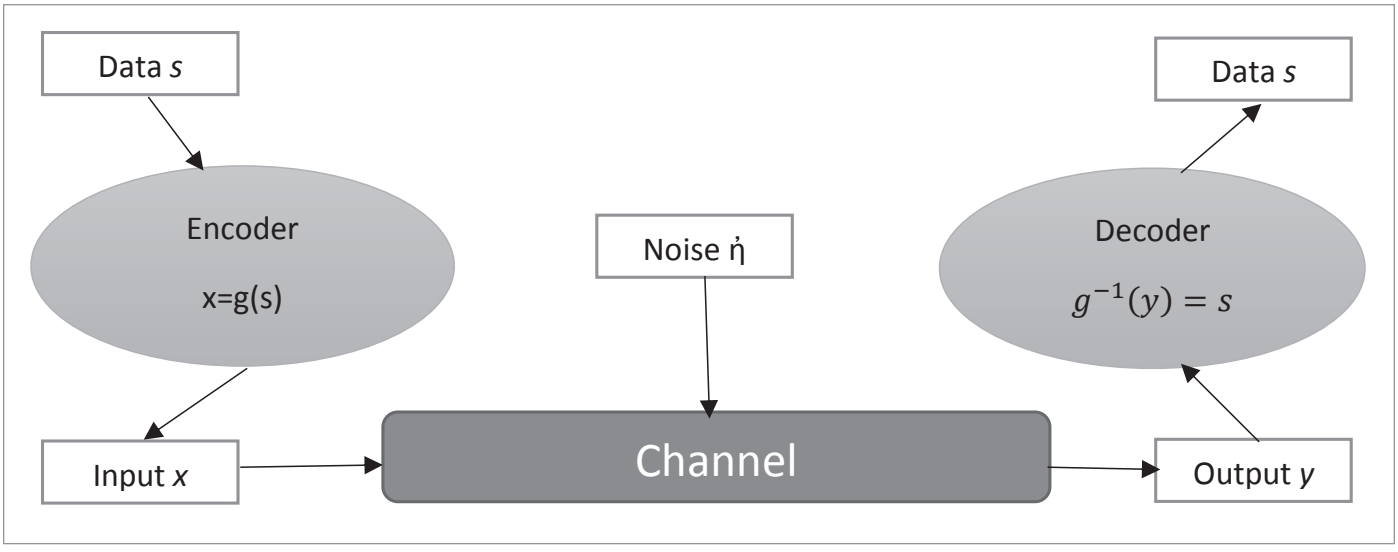
The maximal information coefficient (MIC) has been described as a 21st-century correlation that has its roots in information theory.[1] Information theory was developed by Claude Shannon back in 1948 when he published the paper "A Mathematical Theory of Communication" while working for Bell Labs. Scientists were trying to understand the limits of communication through a communication channel and how to send a signal and minimize the errors in the received message.[3] Even though this seems far removed from any problem in actuarial science, it turns out that it can be very useful for actuaries, such as:

1. Choosing between competing models for a stochastic phenomenon under investigation;
2. Adjusting mortality tables in a statistically valid manner to obtain exactly certain known or assumed individual characteristics, while simultaneously developing a table that is as close as possible to a given standard mortality table;
3. Smoothing observed insurance data to obtain smoothed estimates that are as close as possible to the observed data; and
4. Incorporating monotonicity constraints into a life table graduation.[4]

This article will describe the basic mechanics behind information theory, such as bits, entropy and mutual information, give some intuitive interpretation of its results and relate Pearson's correlation to MIC.

The basic unit of information theory is the bit, which stands for binary digit. This is unfortunate because a binary digit and a bit are different. A binary digit is the value of a binary variable, which can have only two values: zero and one. A bit is the amount of information required to choose between two equally probable alternatives. If there are m equally probable alternatives that can be arrived at by successively making n binary choices, then $n = \log_2 m$ bits of information are required. If the log is changed from base 2 to base e or base 10, then the units are nats or bans, respectively.[3] Information theory's original intent was to determine how to efficiently communicate information from point A to B with the least amount of error. Figure 1 shows the basic structure of communication.

Figure 1
Basic Structure of Communication



1. A source *s* generates a message, which is an ordered sequence of *k* symbols $s = (s_1,\ldots,s_k)$.
2. The source can be coded from an alphabet A_s which can have α letters, so $A_s = (s_1,\ldots,s_\alpha)$.
3. A message *s* is encoded as an input *x* by some function *g* into code words $x = (x_1,\ldots,x_n)$.
4. These code words can have their own alphabet with m letters, hence $A_x = (x_1,\ldots,x_m)$.
5. The input *x* is transmitted through the channel where noise ή is added to the output $Y = X + \acute{\eta}$.
6. The output y code words are decoded back into the original message.

Both the input *X* and output *Y* code words are defined as random variables, so there is a probability associated with each one of the encoded and decoded code words. The probability *p* of all the possible letters in an alphabet need to sum to unity. The output may not be the same as the input because the noise could have added errors into the transmission and changed the resulting alphabet character. The encoder is responsible for compressing and adding error-detecting redundancy. The decoder is responsible for decompressing the message and using the redundancy to remove errors from the message. The error rate in the transmission is the number of incorrect inputs associated with the output per the number of possible inputs.[3] Now that the original problem has been explained, it is time to formally define information and entropy.

Suppose that a biased coined is flipped and the probability of a heads is 95 percent. When the coin comes up heads, there is little information provided or surprise in this result. But if the result is tails, this is a lot more surprising and informative. The Shannon information is the amount of uncertainty or surprise in a random variable. It is defined as the $\log_2 1/p(z)$ bits, where *z* is any random variable, so the uncertainty of a variable should decrease as the probability of an event increases. The entropy $H(Z)$ is the expected value of the Shannon information $H(Z) = -\sum_i p(z_i)\log_2 p(z_i)$. A variable with $H(Z)$ bits entropy will have enough Shannon information to choose between $2^{H(z)}$ equally probable outcomes.[3] The calculation for the entropy is different for discrete versus continuous random variables. To see the problem, the entropy differential $H(Z^\Delta)$ needs to be defined: $H(Z^\Delta) = \sum_i p(z_i)\Delta z \log_2 \frac{1}{p(z_i)\Delta z}$. It is obvious that as $\Delta z \to 0$ then $H(Z^\Delta) \to \infty$. This can be interpreted as saying that as the precision of a variable increases, so dos the bits of information provided by the variable.[3] This means that integrals cannot be used to calculate the continuous entropy. To do the calculation, the random variables must be discretized by dividing the ranges into variable bins and counting how many values fall in the histogram grid.[1] Entropy has some very nice properties regardless if discrete or continuous:

- Continuity—the amount of information associated with an event increases or decreases continuously;
- Symmetry—the amount of information associated with a sequence of events does not depend on the order in which they occurred;
- Maximal Value—the amount of information associated with a set of events cannot be increased if the events are equally likely;
- Additive—the information associated with a set of events is obtained by adding the events together;
- Positive—it will always be greater than equal zero.[3]

The conditional entropy $H(Y|X)$ is the average amount of uncertainity in *Y* given that *X* has occurred, or, to phrase it another way, it is the amount of uncertainty in *Y* that cannot be contributed to *X*.[3] If the focus is put back on signal processing then the output *Y* is nothing more than the input *X* + random noise. The $H(Y|X) = H(X + \acute{\eta} \mid X) = H(\acute{\eta})$ so the average uncertainty in the output given the input is equivalent to the average uncertainty in the noise.[3]

The **relative entropy** between two distributions can be calculated by the Kullback-Liebler (KL) divergence. The relative entropy is a measure of the dissimilarity between probability distributions *p* and *q*: $KL(p||q) = \sum_k^K p_k \log_2 \frac{p_k}{q_k} = \sum_k^K p_k \log_2 p_k - \sum_k^K p_k \log_2 q_k$ where $\bullet_k^K p_k \log_2 q_k$ is called the **cross entropy**. The cross entropy is the average number of bits needed to encode data coming from a source with distribution *p* when model *q* is used. The KL divergence is the average number of extra bits needed to encode the data, due to the fact that the distribution *q* was used to encode the data versus *p*: $KL(p||q) \geq 0$ unless $p = q$.[1] Note that in general the relative entropy is not symmetric under interchange of the distributions *p* and *q*: in general $KL(p||q) \uparrow KL(q||p)$, so *KL*, although it is sometimes called the "KL distance," is not strictly a distance. The relative entropy is important in pattern recognition and neural networks, as well as in information theory.[2]

If there was a goal to state how one variable depends on another, one measurement that would suffice is to calculate the Pearson's correlation ρ that we are all so familiar with: $cov_{xy} = \sum_i^N \frac{(x_i - \bar{x})(y_i - \bar{y})}{N-1}$ *and* $\rho = \frac{cov_{xy}}{\sigma_x \sigma_y}$. Pearson's correlation measures only the linear relationship between random

variables x and y within a range [–1, 1] where 0,–1, 1 implies no relationship, perfectly negative relationship and perfectly positive relationship, respectively. Even though 0 implies no relationship, it does not imply that the random variables are independent. This is a limiting measure of dependence because many relationships are nonlinear. Mutual information is a more general approach of calculating how random variables depend on each other. It has a range from [0, ∞). There are actually several different formulas with corresponding interpretations for mutual information:

1. $I(X,Y) = KL(p(x,y) \| p(x)p(y)) = \sum_x \sum_y p(x,y) \log_2 \frac{p(x,y)}{p(x)p(y)}$.
   This is the extra bits needed to encode the data given independent distributions were used versus the joint distribution of X and Y.
2. $I(X,Y) = H(X) + H(Y) - H(X,Y)$. This is the intersection between the average uncertainty of the input and output.
3. $I(X,Y) = H(X) - H(X \mid Y)$. This is the difference between the average uncertainty of the input and the average uncertainty of the input knowing the output.
4. $I(X,Y) = H(Y) - H(Y \mid X)$. This is the average uncertainty of the output less the average uncertainty of the output given the input.
5. $I(X,Y) = H(Y) - H(noise)$. This is the difference between the average uncertainty of the output and the noise.[3]

Given that the mutual information is derived from the entropy, it suffers from the same problem of being infinite for continuous variables. Unfortunately, the number of bins used, and the location of the bin boundaries, can have a significant effect on the results of MIC. The maximal information coefficient is an approach to try many different bin sizes and locations, and to compare the maximum mutual information received. It is defined as $MIC \triangleq max_{x,y:xy<B} m(x,y)$ such that $m(x,y) = \frac{max_{G \in G(x,y)} I(X(G), Y(G))}{log_2 \min(x,y)}$ where B is some sample-size dependent bound on the number of bins that can be used to reliably estimate the distribution and $G(x,y)$ is the set of two-dimensional grids of size $x \times y$ and $X(G)$, $Y(G)$ represents a discretization of the variables onto this grid. The MIC lies in a range [0,1], where 0 represents no relationship between variables and 1 represents a noise-free relationship of any form, not just linear. MIC will not give any indication of the type of the relationship, though. It is possible with the MIC to find interesting relationships between variables in a way that simpler measures, such as the correlation coefficient, cannot.[1] With MIC the goal is equitability: similar scores will be seen in relationships with similar noise levels regardless of the type of relationship. Because of this, it may be particularly useful with high dimensional settings to find a smaller set of the strongest correlations. Where distance correlation might be better at detecting the presence of (possibly weak) dependencies, the MIC is more geared toward the assessment of strength and detecting patterns that we would pick up via visual inspection.[5]

In conclusion, this article has taken you from the elementary beginnings of information theory. The concepts of bits and nats were explained, which led to the definition of entropy and its many flavors as well as the definition of the KL divergence and its interpretation. This led to the description mutual information for the discrete and continuous cases. Last, the familiar Pearson's correlation coefficient was compared to MIC. MIC is important to pattern recognition because it is a general approach to measure the dependency between two random variables, whereas Pearson's correlation measures only linearity between two random variables. ■

Bryon Robidoux, FSA, is director and actuary at AIG in Chesterfield, MO. He can be reached at *Bryon.Robidoux@aig.com.*

ENDNOTES

1 Murphy, Kevin P. 2012. *Machine Learning: A Probabilistic Perspective*. Cambridge, MA: MIT Press.

2 MacKay, David J. C. 2003. *Information Theory, Inference, and Learning Algorithms,* 2nd ed. Cambridge: Cambridge University Press.

3 Stone, James V. 2015. *Information Theory: A Tutorial Approach*. Sheffield, UK: Sebtel Press.

4 Brocket, Patrick L. 1991. "Information theoretic approach to actuarial science: A unification and extension of relevant theory and applications," *Transactions of the Society of Actuaries,* vol. 43.

5 Clark, Michael. 2013. "A comparison of correlation measures," Center for Social Research, University of Notre Dame, *https://m-clark.github.io/docs/Correlation-Comparison.pdf.*