MUTUAL INFORMATION AS A MEASURE OF DEPENDENCE

Erik-Jan van Kesteren
October 5, 2017

Methods & Statistics Data Science lab
Information

Mutual Information

Maximal Information Coefficient

Questions?

Let’s play!
Entropy is a measure of uncertainty about the value of a random variable.

Formalised by Shannon (1948) at Bell Labs.

Its unit is commonly shannon, bits, or nats.

In general (discrete case):

\[ H(X) = - \sum_{x \in X} p(x) \log p(x) \]
Let $X$ be the outcome of a coin flip:

$$X \sim \text{bernoulli}(p)$$

then:

$$\mathcal{H}(X) = -p \log p - (1 - p) \log(1 - p)$$
coinEntropy <- function(p) -p * log(p) - (1-p) * log(1-p)
curve(coinEntropy, 0, 1)
When we use 2 as the base of the log, the unit will be in **shannon** or **bits**.
Uncertainty = Information

“the amount of information we gain when we observe the result of an experiment is equal to the amount of uncertainty about the outcome before we carry out the experiment” (Rényi, 1961)
We can also do this for multivariate probability mass functions:

$$\mathcal{H}(X, Y) = \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$
MUTUAL INFORMATION
Mutual Information is the information that a variable $X$ carries about a variable $Y$ (or vice versa)

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

$$= - \sum_{x \in X} p(x) \log p(x) - \sum_{y \in Y} p(y) \log p(y) + \sum_{x \in X} \sum_{y \in Y} p(x, y) \log p(x, y)$$

$$= \sum_{x \in X} \sum_{y \in Y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right)$$
Mutual information $I(X; Y)$ is a measure of association between two random variables which captures linear and nonlinear relations.

If $X \sim \mathcal{N}(\mu_1, \sigma_1)$ and $Y \sim \mathcal{N}(\mu_2, \sigma_2)$, then

$$I(X; Y) \geq -\frac{1}{2} \log(1 - \rho^2)$$

(Krafft, 2013)
Common estimation method: *discretize* and then calculate $\mathcal{I}(X, Y)$.

Other option: *kde* and then numerical integration.

This is an active field of research in ML (e.g., Gao et al., 2017).
MAXIMAL INFORMATION COEFFICIENT
We need a measure of dependence that is *equitable*: its value should depend only on the amount of noise and not on the functional form of the relation between $X$ and $Y$. (Reshef et al., 2011, paraphrased)
\[ H(X) = -0.3 \log 0.3 - 0.3 \log 0.3 - 0.4 \log 0.4 = 1.09 \]
\[ H(Y) = -0.2 \log 0.2 - 0.4 \log 0.4 - 0.3 \log 0.3 - 0.1 \log 0.1 = 1.28 \]
\[ H(X, Y) = -0.6 \log 0.1 - 0.4 \log 0.2 = 2.03 \]

\[ I(X; Y) = H(X) + H(Y) - H(X, Y) = 0.34 \]

Then, normalise so that \( I_n(X; Y) \in [0, 1] \)

\[ I_n(X; Y) = \frac{I(X; Y)}{\log \min(n_x, n_y)} = \frac{0.34}{\log 3} = 0.31 \]
How to calculate the Maximal Information Criterion (MIC)

1. For all grids of size $n_x \times n_y$ up to $n_x \cdot n_y \leq N^{0.6}$ calculate maximum normalised MI for different bin sizes.
2. Pick the maximum value of these normalised MIs.
<table>
<thead>
<tr>
<th>Relationship Type</th>
<th>MIC</th>
<th>Pearson</th>
<th>Spearman</th>
<th>Mutual Information (KDE)</th>
<th>CorGC (Principal Curve-Based)</th>
<th>Maximal Correlation</th>
</tr>
</thead>
<tbody>
<tr>
<td>Random</td>
<td>0.18</td>
<td>-0.02</td>
<td>-0.02</td>
<td>0.01</td>
<td>0.03</td>
<td>0.19</td>
</tr>
<tr>
<td>Linear</td>
<td>1.00</td>
<td>1.00</td>
<td>1.00</td>
<td>5.03</td>
<td>3.89</td>
<td>1.00</td>
</tr>
<tr>
<td>Cubic</td>
<td>1.00</td>
<td>0.61</td>
<td>0.69</td>
<td>3.09</td>
<td>3.12</td>
<td>0.98</td>
</tr>
<tr>
<td>Exponential</td>
<td>1.00</td>
<td>0.70</td>
<td>1.00</td>
<td>2.09</td>
<td>3.62</td>
<td>0.94</td>
</tr>
<tr>
<td>Sinusoidal (Fourier frequency)</td>
<td>1.00</td>
<td>-0.09</td>
<td>-0.09</td>
<td>0.01</td>
<td>-0.11</td>
<td>0.36</td>
</tr>
<tr>
<td>Categorical</td>
<td>1.00</td>
<td>0.53</td>
<td>0.49</td>
<td>2.22</td>
<td>1.65</td>
<td>1.00</td>
</tr>
<tr>
<td>Periodic/Linear</td>
<td>1.00</td>
<td>0.33</td>
<td>0.31</td>
<td>0.69</td>
<td>0.45</td>
<td>0.49</td>
</tr>
<tr>
<td>Parabolic</td>
<td>1.00</td>
<td>-0.01</td>
<td>-0.01</td>
<td>3.33</td>
<td>3.15</td>
<td>1.00</td>
</tr>
<tr>
<td>Sinusoidal (non-Fourier frequency)</td>
<td>1.00</td>
<td>0.00</td>
<td>0.00</td>
<td>0.01</td>
<td>0.20</td>
<td>0.40</td>
</tr>
<tr>
<td>Sinusoidal (varying frequency)</td>
<td>1.00</td>
<td>-0.11</td>
<td>-0.11</td>
<td>0.02</td>
<td>0.06</td>
<td>0.38</td>
</tr>
</tbody>
</table>
QUESTIONS?
LET’S PLAY!
install.packages("minerva")
library("minerva")
set.seed(142857)
x <- rnorm(300)

# Define functional form
f <- function(x) log(abs(x))

# Get the MIC
mine(x, f(x))$MIC
THE RULES

1. Don’t add errors! The goal is to cheat the system!
2. You can only use x once in f(x).
3. f(x) can only perform 2 operations.
4. Any number in f(x) needs to be a 9.
5. Top tip: \texttt{plot(x, f(x))}.


read more:
http://science.sciencemag.org/content/334/6062/1502.full
f <- function(x) abs(9 %% x)
mine(x, f(x))$MIC
# [1] 0.4969735

f(x) = abs(9 %% x)