# 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!

# INFORMATION

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):

$$\mathcal{H}(X) = -\sum_{x \in X} p(x) \log p(x)$$

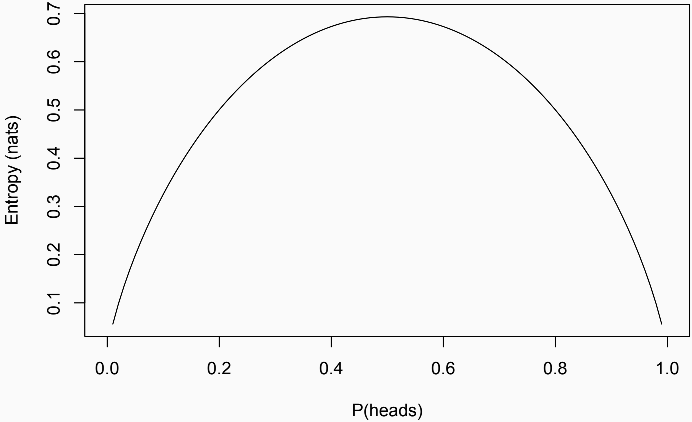
Let X be the outcome of a coin flip:

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

then:

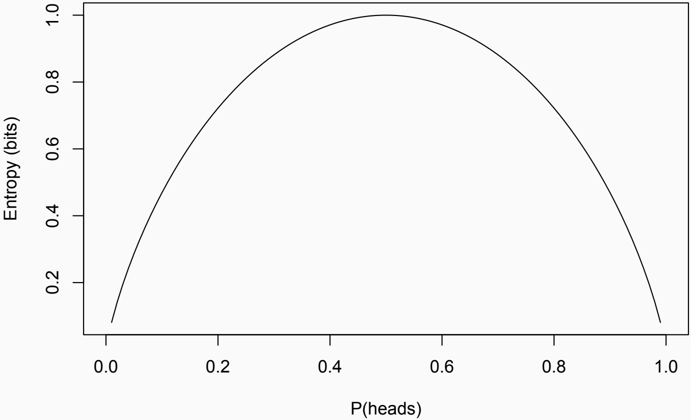
$$\mathcal{H}(X) = -p \log p - (1-p) \log(1-p)$$

```
coinEntropy <- function(p) -p * log(p) - (1-p) * log(1-p)
curve(coinEntropy, 0, 1)
```



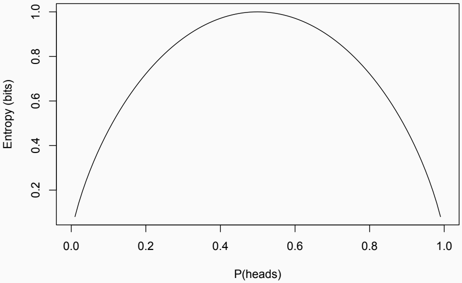
# ENTROPY

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)

$$\mathcal{I}(X;Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X,Y)$$

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

$$= \sum_{x \in X} \sum_{y \in Y} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right)$$

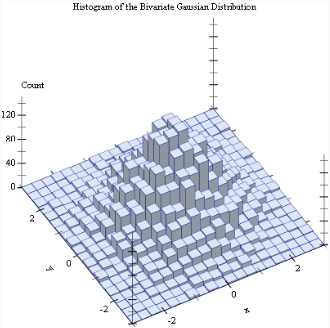
$\mathcal{I}(X;Y)$ is a measure of association between two random variables which captures linear and nonlinear relations

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

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

(Krafft, 2013)

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



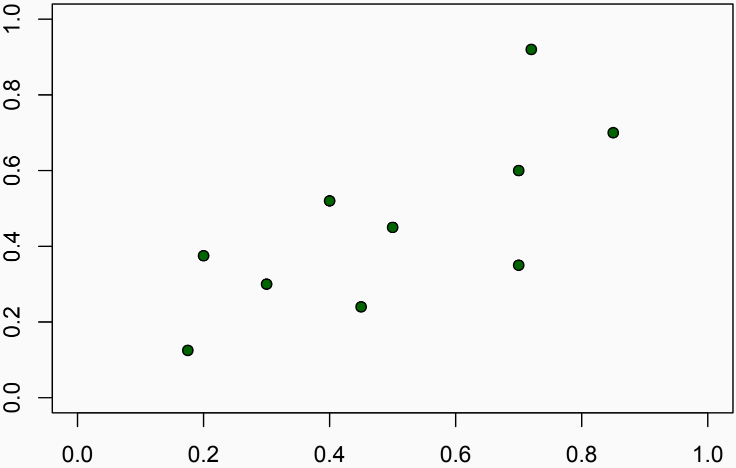

Other option: `kde` and then numerical integration.

This is an active field of research in ML (e.g., Gao et al., 2017).

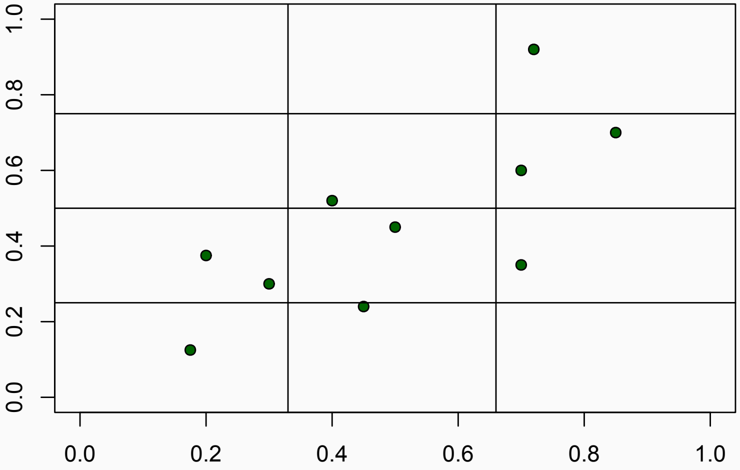
# MAXIMAL INFORMATION COEFFICIENT

# 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)


0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

# EXAMPLE



# EXAMPLE

$$\mathcal{H}(X) = -0.3 \log 0.3 - 0.3 \log 0.3 - 0.4 \log 0.4 = 1.09$$
$$\mathcal{H}(Y) = -0.2 \log 0.2 - 0.4 \log 0.4 - 0.3 \log 0.3 - 0.1 \log 0.1 = 1.28$$
$$\mathcal{H}(X, Y) = -0.6 \log 0.1 - 0.4 \log 0.2 = 2.03$$

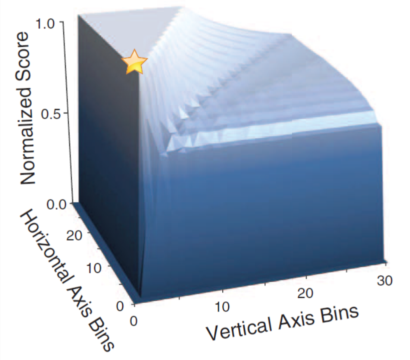
$$\mathcal{I}(X; Y) = \mathcal{H}(X) + \mathcal{H}(Y) - \mathcal{H}(X, Y) = 0.34$$

Then, normalise so that $\mathcal{I}_n(X; Y) \in [0, 1]$

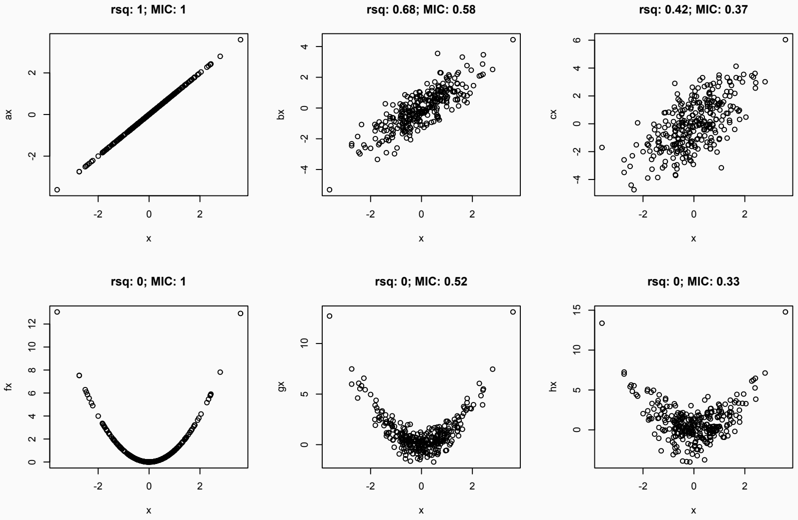
$$\mathcal{I}_n(X; Y) = \frac{\mathcal{I}(X; Y)}{\log \min(n_x, n_y)} = \frac{0.34}{\log 3} = 0.31$$

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.



# EQUITABILITY



# FUNCTIONAL FORMS

| Relationship Type | MIC | Pearson | Spearman | Mutual Information *(KDE)* | Mutual Information *(Kraskov)* | CorGC (Principal Curve-Based) | Maximal Correlation |
|---|---|---|---|---|---|---|---|
| Random | 0.18 | -0.02 | -0.02 | 0.01 | 0.03 | 0.19 | 0.01 |
| Linear | 1.00 | 1.00 | 1.00 | 5.03 | 3.89 | 1.00 | 1.00 |
| Cubic | 1.00 | 0.61 | 0.69 | 3.09 | 3.12 | 0.98 | 1.00 |
| Exponential | 1.00 | 0.70 | 1.00 | 2.09 | 3.62 | 0.94 | 1.00 |
| Sinusoidal *(Fourier frequency)* | 1.00 | -0.09 | -0.09 | 0.01 | -0.11 | 0.36 | 0.64 |
| Categorical | 1.00 | 0.53 | 0.49 | 2.22 | 1.65 | 1.00 | 1.00 |
| Periodic/Linear | 1.00 | 0.33 | 0.31 | 0.69 | 0.45 | 0.49 | 0.91 |
| Parabolic | 1.00 | -0.01 | -0.01 | 3.33 | 3.15 | 1.00 | 1.00 |
| Sinusoidal *(non-Fourier frequency)* | 1.00 | 0.00 | 0.00 | 0.01 | 0.20 | 0.40 | 0.80 |
| Sinusoidal *(varying frequency)* | 1.00 | -0.11 | -0.11 | 0.02 | 0.06 | 0.38 | 0.76 |

# QUESTIONS?

# LET'S PLAY!

```r
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: `plot(x, f(x))`.

Gao, W., Kannan, S., Oh, S., and Viswanath, P. (2017). Estimating Mutual Information for Discrete-Continuous Mixtures. pages 1–25.

Krafft, P. (2013). Correlation and mutual information – building intelligent probabilistic systems.

Rényi, A. (1961). On measures of entropy and information. Fourth Berkeley Symposium on Mathematical Statistics and Probability, 1(c):547–561.

Reshef, D., Reshef, Y., Finucane, H., Grossman, S., Mcvean, G., Turnbaugh, P., Lander, E., Mitzenmacher, M., and Sabeti, P. (2011). Detecting Novel Associations in Large Data Sets. Science, 334(6062):1518–1524.

Shannon, C. E. (1948). A mathematical theory of communication. The Bell System Technical Journal, 27(July 1928):379–423.

---

read more:

`http://science.sciencemag.org/content/334/6062/1502.full`

# MY TOP FUNCTION

```r
f <- function(x) abs(9 %% x)
mine(x, f(x))$MIC
# [1] 0.4969735
```