Complex Systems Methods — 2. Conditional mutual information, entropy rate and algorithmic complexity

Eckehard Olbrich

MPI MiS Leipzig

Potsdam WS 2007/08

Overview

Summary of Entropy and Information

- Random variables X, Y with values $x \in \mathcal{X}, y \in \mathcal{Y}$
- Entropy $H(X) = -\sum_{x \in \mathcal{X}} p(x) \log p(x) = E_{p(x)}[1/\log(p(x))]$
- Conditional entropy
  $H(X|Y) = H(X, Y) - H(Y) = E_{p(x,y)}[1/\log(p(y|x))]$
- Mutual information

$$MI(X:Y) = H(X) - H(X|Y) = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} p(x,y) \log \frac{p(x,y)}{p(x)p(y)}$$

- Relative entropy $D(p||q) = \sum_{x \in \mathcal{X}} p(x) \log \frac{p(x)}{q(x)}$
- Mutual information as relative entropy

$$MI(X:Y) = D(p(x,y)||p(x)p(y))$$

Convex functions and Jensens Inequality

Definition A function f(x) is said to be *convex* over an interval (a, b) if for every $x_1, x_2 \in (a, b)$ and $0 \leq \lambda \leq 1$,

$$f(\lambda x_1 + (1-\lambda)x_2) \leq \lambda f(x_1) + (1-\lambda) f(x_2)$$

A function is said to be *strictly convex* if equality holds only if $\lambda = 0$ or $\lambda = 1$.
A function f is *concave* if $-f$ is convex.
Examples x^2,$|x|$,e^x,$x \log(x)$ for $x \geq 0$ are *convex* functions, $\log x$ or $\sqrt{x}$ are *concave* functions.
Theorem *Jensens inequality* If f is a convex function and X is a random variable,

$$E_P[f(X)] \geq f(E_P[X]) \ .$$

If f is strictly convex equality implies $X = E_P[X]$ with probability 1 (i.e. X is a constant).

Information inequality

Now we are able to prove the non-negativity of the relative entropy and the mutual information:

Theorem $D(p||q) \geq 0$ with equality iff $p(x) = q(x)\ \forall x$.

Corollary: Non-negativity of the mutual information

$$MI(X:Y) = D(p(x,y)||p(x)p(y)) \geq 0$$

$MI(X:Y) = 0$ implies statistical independence, i.e. $p(x,y) = p(x)p(y)$.

Corollary: The uniform distribution over the range of X $u(x) = 1/|\mathcal{X}|$ is the maximum entropy distribution over this range.

$$D(p(x)||u(x)) = \sum_{x\in\mathcal{X}} p(x) \log \frac{p(x)}{u(x)} = log|\mathcal{X}| - H(X) \geq 0 \Rightarrow H(X) \leq log|\mathcal{X}|$$

Conditional mutual information

Lets have three random variables X, Y, Z we can ask, what we learn about X by observing Z knowing already Y. Answer:

$$MI(X:Z|Y) = H(X|Y) - H(X|Y,Z) = \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y|z)}{p(x|z)p(y|z)}$$

Properties:

1. Symmetry $MI(X:Z|Y) = MI(Z:X|Y)$
2. Non-negativity

$$MI(X:Z|Y) = \sum_{z} p(z) D(p(x,y|z)||p(x|z)p(y|z)) \geq 0$$

3. X is conditional independent on Z given Y, i.e. $p(x|y,z) = p(x|y)$, denoted by $X \perp Z|Y$, if and only if $MI(X:Z|Y) = 0$.

Chain rules

- Entropy

$$H(X,Y) = H(X) + H(Y|X)$$

- Mutual information

$$MI(X:Y,Z) = MI(X:Y) + MI(X:Z|Y)$$

- Relative Entropy:

$$D(p(x,y)||q(x,y)) = D(p(x)||q(x)) + D(p(y|x)||q(y|x))$$

Stationary stochastic processes

- A **stochastic process** is indexed sequence of random variables. The process is characterized by joint probabilities $Pr\{(X_1, X_2, ..., X_n) = (x_1, x_2, \ldots, x_n)\} = p(x_1, \ldots, x_n), (x_1, \ldots, x_n) \in \mathcal{X}^n$.
- A stochastic process is said to be **stationary** if the joint distribution of any subset of random variables is invariant with respect to shifts in the time index; that is

$$
\begin{aligned}
& Pr\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} \\
& \quad = Pr\{X_{1+l} = x_1, X_{2+l} = x_2, \ldots, X_{n+l} = x_n\}
\end{aligned}
$$

for every n and every shift l and for all $x_1, x_2, \ldots, x_n \in \mathcal{X}$.

Entropy rate

Entropy rate as **entropy per symbol:**

$$h_\infty = \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)$$

Entropy rate as **conditional entropy given the past:**

$$h'_\infty = \lim_{n\to\infty} H(X_n | X_{n-1}, \ldots, X_1)$$

Theorem: For a stationary stochastic process the limits exists and are equal.

Can be proven using
Theorem: (*Cesáro mean*) If $a_n \to a$ and $b_n = \frac{1}{n}\sum_{i=1}^{n} a_i$, then $b_n \to a$.

Markov chains

Definition A discrete stoachstic process $X_1, X_2, \dots$ is said to be a Markov chain or a Markov process if for $n = 1, 2, \dots$

$$Pr(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_1 = x_1) = Pr(X_{n+1} = x_{n+1} | X_n = x_n)$$

or in a shortened notation:

$$p(x_{n+1} | x_n, \dots, x_1) = p(x_{n+1} | x_n) \ .$$

for all $x_1, x_2, \dots, x_n \in \mathcal{X}$.

Markov chains

- The past is conditional independent of the future, given the present:

$$MI(X_{n+1} : X_{n-1}, \dots, X_1 | X_n) = 0 \ .$$

- Joint probability distribution factorizes:

$$p(x_1, x_2, \dots, x_n) = p(x_1)p(x_2|x_1)p(x_3|x_2)\dots p(x_n|x_{n-1})$$

- Transition matrix $P_{ij} = Pr(X_{n+1} = j | X_n = i)$.
- Dynamics

$$p(x_{n+1}) = \sum_{x_n} p(x_n) P_{x_n, x_{n+1}} \ .$$

- Stationary distribution $p(x_{n+1}) = p(x_n) = \mu(x)$
- Entropy rate

$$\begin{aligned} h_\infty &= H(X_2|X_1) \\ &= -\sum_{ij} \mu_i P_{ij} \log P_{ij} \end{aligned}$$

The entropy rate of natural language

- Consider English as a stationary ergodic process
- Alphabet with 26 letters and the space symbol
- Letters occur non-uniform (E with 13%, Q and Z with 0.1%). Most frequent correlations between T and H or Q and U.
- Entropy rates: Zeroth order $\log 27 = 4.76$ bits per letter. Second order Markov approximation 4.03 bits per letter and fourth order Markov approximation 2.8 bits per letter.
- Entropy rate from guessing the next letter by humans: 1.3 bits per letter (Shannon 1950).
- Gambling estimate with 12 subjects and a sample of 75 letters from the text used by Shannon: 1.34 bits per letter (Cover and King 1978)

Algorithmic complexity

Definition The *Kolmogorov complexity* $K_{\mathcal{U}}(x)$ of a binary string x with respect to a universal computer $\mathcal{U}$ is defined as

$$K_{\mathcal{U}}(x) = \min_{p:\mathcal{U}(p)=x} l(p)$$

with $l(p)$ the length of the string p and running the program p on the universal computer $\mathcal{U}$ produces the output x and halts.

Theorem (*Universality of Kolmogorov complexity*) If $\mathcal{U}$ is a universal computer, for any other computer $\mathcal{A}$ there exists a constant $c_{\mathcal{A}}$ such that

$$K_{\mathcal{U}}(x) \leq K_{\mathcal{A}}(x) + c_{\mathcal{A}}$$

for all strings $x \in \{0,1\}^*$, and the constant $c_{\mathcal{A}}$ does not depend on x.

Upper and lower bounds

- *Conditional Kolmogorov complexity* knowing $l(x)$

$$K_{\mathcal{U}}(x|l(x)) = \min_{p:\mathcal{U}(p,l(x))=x} l(p)$$

- Upper bounds

$$\begin{aligned} K(x|l(x)) &\leq l(x) + c \\ K(x) &\leq K(x|l(x)) + \log^* l(x) + c \end{aligned}$$

with $\log^* n = \log n + \log\log n + \log\log\log n + \ldots$ as long as the terms are positive.

- Lower bound: The number of strings x with complexity $K(x) < k$ satisfies

$$|K(x) < k| < 2^k$$

because there are only $2^k - 1$ strings and therefore possible programs with length $k - 1$.

- A sequence $x_1, x_2, \ldots, x_n$ is said to be algorithmically random if
$$K(x_1, x_2, \ldots, x_n | n) \geq n \, .$$
- There exists for each n at least one sequence x^n such that
$$K(x^n | n) \geq n$$
- A string is called incompressible if
$$\lim_{n \to \infty} \frac{K(x_1, x_2, \ldots, x_n | n)}{n} = 1 \, .$$
- *Strong law of large numbers for incompressible binary sequences*
$$\frac{1}{n} \sum_{i=1}^{n} x_i \to \frac{1}{2} \, ,$$
i.e. the proportion of 0's and 1's in any incompressible string are almost equal.

Algorithmic complexity and entropy

Let the stochastic process $\{X_i\}$ be drawn i.i.d. according to the probability distribution $p(x)$, $x \in \mathcal{X}$, where $\mathcal{X}$ is a finite alphabet. There exists a constant c such that

$$H(X) \leq \frac{1}{n} \sum_{x^n} p(x^n) K(x^n|n) \leq H(X) + \frac{(|\mathcal{X}| - 1)\log n}{n} + \frac{c}{n}$$

for all n. x^n is denoting $x_1, \ldots, x_n$. Consequently

$$E[\frac{1}{n} K(X^n|n)] \to H(X)$$

More general (*Brudno's Theorem*): The entropy rate of an ergodic dynamical system is equal to the rate of the Kolmogorov complexity of almost all of its trajectories encoded by its generating partition.

Kolmogorov sufficient statistic

The *Kolmogorov structure function* $K_k(x^n|n)$ of a binary string $x \in \{0,1\}^n$ is defined as

$$K_k(x^n|n) = \min_{\substack{p\,:\,l(p) \leq k \\ \mathcal{U}(p,n) = S \\ x^n \in S \subseteq \{0,1\}^n}} \log |S|$$

The set S is the smallest set that can be described with no more than k bits and which includes x^n. $\mathcal{U}(p,n) = S$ means, that running p with data n on the computer $\mathcal{U}$ will print out the indicator function of the set S. For a given small constant c, let k^* be the least k such that

$$K_k(x^n|n) + k \leq K(x^n|n) + c\ .$$

The corresponding program p^{**} that prints out the indicator function on the corresponding set S^{**} is a Kolmogorov minimal sufficient statistic for x^n.

Further remarks

- Independently developed by Solomonoff (1964), Kolmogorov (1965), Chaitin (1966)
- The Kolmogorov complexity is uncomputable (related to the Halting problem, Gödels incompleteness theorem, ...).
- Further reading: An introdution to Kolmogorov Complexity and Its Applications by Ming Li and Paul Vitányi, Springer 1997
  Stochastic Complexity in Statistical Inquiry by Jorma Rissanen, World Scientific, 1989 — Minimum description length