Complexity of finite strings

2016年10月13日 14:44

We turn to the theory of Randomness

We plan to talk about three approaches to characterize our intuition about randomness

The first approach can be viewed as an information theoretical approach.

To measure how much information a (finite) string carries, we refer to how hard to compress it

Example The strings 1000000000000000000000001

or 0101010101010101010101

carrying relatively small amount of information, since it can be easily described / compressed

Intuitively, a randomly generated string should not be easily described, i.e. containing relatively large amount of information.

To characterize the complexity (hardness to be compressed) of a finite string:

Def Let $f: 2^{<\omega} \to 2^{<\omega}$ be a partial computable function, and M is a corresponding Turing machine. We define the (plain) Kolmogorov complexity of string σ with respect to f / M to be

$$C_f(\sigma) = C_M(\sigma) = \min\{|\tau| \mid f(\tau) = \sigma\}$$

where $\min \emptyset = \infty$, i.e. if $\sigma \notin \mathrm{ran} f$ then $C_f(\sigma) = \infty$

Here we think of f / M as a "description system" or "decompression program"

Def Let $f/M: 2^{<\omega} \to 2^{<\omega}$ be partial computable. σ be a finite string we say σ is random (in the sense of plain Kolmogorov complexity) relative to f / M if $|\sigma| \leq C_f(\sigma) = C_M(\sigma)$

To get rid of the dependence on the choice of f / M, we use a universal description system.

Def A partial computable function / machine $U: 2^{<\omega} \to 2^{<\omega}$ is universal if for each partial computable function f, there is a constant string ρ_f s.t.

for all string $\sigma \in 2^{<\omega}$,

$$U(\rho_f \sigma) = f(\sigma)$$

We say ρ_f is a coding string, $|\rho_f|$ is a coding constant of f in U

A universal machine is optimal in the following sense: there is a constant e_M
For each machine $M: 2^{<\omega} \to 2^{<\omega}$, there is a constant e_M s.t.

$$\forall \tau \forall \sigma \left[M(\sigma) = \tau \to \exists \theta \, (U(\theta) = \tau \wedge |\theta| \leq |\sigma| + e_M) \right]$$

or equivalently, there is a constant e_M s.t. for all τ,

$$C_U(\tau) \leq C_M(\tau) + e_M$$

(same meaning)

In this case, we also write $C_U(\tau) \leq C_M(\tau) + O(1)$

Fact There exists a universal machine / partial computable function.

Let U be the following program:

$$U(\underbrace{0\cdots01}_{e}\sigma) = \Phi_e(\sigma)$$

So the coding string for Φ_e is $\underbrace{0\cdots01}_{e}$, and the coding constant is $e+1$

From now on, we fix a universal Machine U

Def For $\sigma \in 2^{<\omega}$, define the (plain) Kolmogorov complexity of σ to be

$$C(\sigma) = C_U(\sigma)$$

Clearly C is defined on all strings, i.e. $\mathrm{dom}\, C = 2^{<\omega}$
[Consider the program $M(\sigma) = \sigma$]

Proposition

(1) $C(\sigma) \leq |\sigma| + O(1)$
(2) $C(\sigma\sigma) \leq C(\sigma) + O(1)$
(3) If $h: 2^{<\omega} \to 2^{<\omega}$ is partial computable, then $C(h(\sigma)) \leq C(\sigma) + O(1)$ for $h(\sigma)$ defined

Proof (3) Let M be the program: $M(\tau) = h(U(\tau))$

So $M(\theta)=n(\sigma) \Leftrightarrow U(\theta)=\sigma$, therefore $C_M(n(\sigma)) = C(\sigma)$

Then $C(n(\sigma)) \le C_M(n(\sigma)) + O(1) \le C(\sigma) + O(1)$

$C(\sigma)$

(2) is a special case of (3)

For natural number $n \in N$, when we write $C(n)$ or $C_M(n)$, we think of n as its binary representation

Note $C(n) \le \log n + O(1)$

We also use $C(\sigma,\tau)$ to denote $C(\langle\sigma,\tau\rangle)$

the particular choice of pairing function $p: 2^{<\omega} \xrightarrow[\text{onto}]{1-1} 2^{<\omega} \times 2^{<\omega}$ is independent modulo constant.

For each $\sigma \in 2^{<\omega}$, assume $C(\sigma)=n$, then there is a least (leftmost) string τ s.t. $U(\tau)=\sigma$ and $|\tau|=n$, we define σ^*_C to be this τ.

The information contained in σ together with $C(\sigma)$ is no more than σ^*_C

Proposition $C(\sigma, C(\sigma)) = C(\sigma^*_C) + O(1)$

Proof Consider partial computable function

$$h(\tau) = \langle U(\tau), |\tau| \rangle$$

Then $h(\sigma^*_C) = \langle \sigma, |\sigma^*_C| \rangle = \langle \sigma, C(\sigma) \rangle$ for all $\sigma \in 2^{<\omega}$

and $C(h(\tau)) \le C(\tau) + O(1)$

so $C(\sigma, C(\sigma)) \le C(\sigma^*_C) + O(1)$ for all $\sigma \in 2^{<\omega}$ □

Def For $\sigma \in 2^{<\omega}$, we say σ is C-random if

$$C(\sigma) \ge |\sigma|$$

More carefully, Let d be a constant, we say σ is C-random for d if

$$C(\sigma) \ge |\sigma| - d$$

Proposition 1) For each n there is a σ s.t. $|\sigma|=n$ and $C(\sigma) \ge n$

2) Fix d, for any n, at least $2^n - 2^{n-d}$ many strings of length n are C-random for d.

Proof 1) $|\{\sigma \mid |\sigma|=n\}| = 2^n$

but $|\{\tau \mid |\tau| < n\}| = 2^0 + \cdots + 2^{n-1} = 2^n - 1$

2) $|\{\tau \mid |\tau| < n-d\}| = 2^{n-d} - 1$

so $|\{s \mid |s|=n \text{ and } C(s) \geq n-d\}| \geq 2^n - (2^{n-d} - 1)$

$> 2^n - (2^{n-d} - 1) - 1 = 2^n(1 - 2^{-d})$ □

Example Fix n. the number of strings of length n, which are "half random"

$$|\{s \mid |s|=n \ C(s) \geq \tfrac{n}{2}\}| = 2^n(1 - 2^{-\frac{n}{2}})$$

with large n, almost all $(1 - 2^{-\frac{n}{2}})$ the strings are half random