Kolmogorov Complexity, Data Compression, and Inference

Thomas M. Cover
Stanford University

Abstract

If a sequence of random variables has Shannon entropy H, it is well known that there exists an efficient description of this sequence which requires only H bits. But the entropy H of a sequence also has to do with inference. Low entropy sequences allow good guesses of their next terms. This is best illustrated by allowing a gambler to gamble at fair odds on such a sequence. The amount of money that one can make is essentially the complement of the entropy with respect to the length of the sequence.

Now suppose that the sequence is not random. Although the entropy of such a sequence is not defined, there is a notion of its intrinsic descriptive complexity. This idea, put forth by Kolmogorov, Chaitin, and Solomonoff, says that the intrinsic complexity of a sequence is the length of its shortest description. Here too there is a tradeoff between complexity and inference. Low complexity sequences allow a high degree of inference. Again there is a gambling tradeoff.

Finally, it will be shown that if a sequence is random and has entropy H, then with high probability its Kolmogorov complexity will also be H.

Special attention will be given to the so-called Kolmogorov H function, a function that has not yet made its appearance in the literature. We argue that it plays the role of a minimal sufficient statistic. Thus, we can assert that there is a sufficient statistic for the Mona Lisa. This idea will capture the fundamental structure of geometrical patterns, probability distributions and the laws of nature.

1. Kolmogorov Complexity.

Let N denote the natural numbers $\{0,1,2,...\}$. Let $x \in \{0,1\}^\infty$ denote an infinite binary sequence $x = (x_1,x_2,...)$ and let $x(n) = (x_1,x_2, \ldots, x_n)$ denote the first n terms. Let $\{0,1\}^*$ denote all binary sequences of finite length. Let A be a partial recursive function A: $\{0,1\}^* \times N \rightarrow \{0,1\}^*$. We restrict A to have a prefix free domain, i.e., no program p accepted by A is the prefix of another. Let $l(x)$ denote the length of the sequence x. Then

$$K_A(x(n)\,|\,n) \;=\; \min_{A(p,n)=x(n)} l(p) \tag{1}$$

is defined to be the complexity of $x(n)$ with respect to the algorithm A, given the length n of the sequence $x(n)$. Similarly, let

$$K_A(x) \;=\; \min_{A(p,0)=x} l(p) \tag{2}$$

If A is a universal partial recursive function, then K_A , or simply K, is called the Kolmogorov complexity [1,2,3,4,5,9,10,11]. We know that

$$i)\quad K(x(n)\,|\,n) \le K_B(x(n)\,|\,n) + c_B \quad \text{for all} \quad n \in N\,,\ \forall\ x \tag{3}$$

$$ii)\quad |\{x \in \{0,1\}^* : K(x) < k\}| \le 2^k\,,\ \forall\ k \in N\,. \tag{4}$$

Now we define a complexity measure for functions $f : D \rightarrow \{0,1\}$, where the domain D is some finite set. Let A be a universal partial recursive function

Definition (Function complexity)

$$K_A(f\,|\,D) \;=\; \min_{\substack{A(p,x)=f(x)\\ \forall\ x \in D}} l(p) \tag{5}$$

Thus the complexity of f given the domain D is the minimum length program p such that a Turing machine A, or equivalently a mechanical algorithm A, can compute $f(x)$ in finite time, for all $x \in D$.

2. Some More Properties of the Kolmogorov Complexity K.

First some examples. Let all sequences $x \in \{0,1\}^n$. Let n be known to the computer. Let 0^n denote a sequence of n 0's. Examples:

1. $K(0^n \mid n) = c$ (some constant independent of n). (6)

2. $K(\pi_1\pi_2\cdots\pi_n \mid n) = c$, where $\pi_1\cdots\pi_n$ are the first n bits of π . (7)

3. $K(1st\ n$ bits of Shakespeare $\mid n) \approx n/4$. (8)

4. $K(\alpha_1\alpha_2\cdots\alpha_n \mid n) = ?$,where α_i is the ith bit in the binary expansion of the fine structure constant $\alpha = e^2 / hc$. (9)

5. $K(x_1,x_2 \cdots x_n \mid n) \leq nh(\frac{1}{n} \sum_{i=1}^{n} x_i) + \log n + c$.

If $X_i \sim$ Bernoulli (p) , (10)

then $Pr\{\mid \frac{1}{n} K(X_1X_2 \cdots X_n \mid n) - h(p) \mid > \epsilon\} \to 0$ (11)

We investigate some additional properties of K. Again, we assume n known to the computer.

Proposition 1:

$$K(x(n) \mid n) \leq n + c , \quad \text{for all} \quad x \in \{ 0,1 \}^n . \tag{12}$$

Proof: The program "Print $x_1x_2 \cdots x_n$ " achieves the bound.

Proposition 2:

$$K(x) \leq K(x \mid l(x)) + 2 \log l(x) + c , \quad \text{for all} \quad x \in \{ 0,1 \}^* . \tag{13}$$

Theorem: (Complexity version of law of large numbers) (Fine [6])

$$K(x_1x_2 \cdots x_n \,|\, n) \geq n(1-\epsilon) => \left| \frac{1}{n} \sum_{i=1}^{n} x_i - \frac{1}{2} \right| \leq \epsilon' , \qquad (14)$$

where $\epsilon' \rightarrow 0$ as $\epsilon \rightarrow 0$. Thus high complexity sequences satisfy the most frequently used test for randomness.

3. Gambling on Patterns.

We shall now develop some properties of the function complexity defined in Equation (2). This section follows the development in Cover [7].

Given a domain D of patterns $D = \{x_1, x_2, \ldots, x_n\}$ and an unknown classification function $f : D \rightarrow \{0,1\}$ assigning the patterns to two classes, we ask for an intelligent way to learn f as the correctly classified elements in D are presented one by one. We ask this question in a gambling context in which a gambler, starting with one unit, sequentially bets a portion of his current capital on the classification of the new pattern. We find the optimal gambling system when f is known a priori to belong to some family F. We also exhibit a universal optimal learning scheme achieving $\exp_2(n - K(f \,|\, D) - log(n+1))$ units for each f, where $K(f \,|\, D)$ is the length of the shortest binary computer program that calculates f on its domain D. In particular it can be shown that a gambler can double his money aproximately $n(1 - H(d/n))$ times, where $H(p) = -p \log p - (1\text{-}p)log(1\text{-}p)$, if f turns out to be a linear threshold function on n patterns in d-space.

Let F denote a set of (classification) functions $f : D \rightarrow \{0,1\}$. For example, F might be the set of all linear threshold functions. Let $|F|$ denote the number of elements in F.

The interpretation will be that D is the set of patterns, and $f(x)$ is the classification of the pattern x in D.

Consider the following gambling situation. The elements of D are presented in any order. A gambler starts with one dollar. The first pattern $x_1 \in D$ is exhibited. The gambler then announces amounts b_1 and b_0 that he bets on the true class being $f(x_1) = 1$ and $f(x_1) = 0$, respectively. Without loss of generality we can set $b_1 + b_0 = 1$. The true value $f(x_1)$ is then announced, and the gambler loses the incorrect bet and is paid fair odds (2 for 1) on the correct bet. Thus his new capital is

$$S_1 = \begin{cases} 2b_1, & f(x_1) = 1 \\ 2b_0, & f(x_1) = 0 . \end{cases}$$

Now a new pattern element $x_2 \in D$ is exhibited. Again, the gambler announces proportions b_1 and b_0 of his current capital that he bets on $f(x_2) = 1$ and $f(x_2) = 0$ respectively. Without loss of generality, let $b_0 + b_1 = 1$. Thus the bet sizes are $b_1 S_1$ and $b_0 S_1$. Then $f(x)$ is announced and the gambler's new capital is

$$S_2 = \begin{cases} 2b_1 S_1, & f(x_2) = 1 \\ 2b_0 S_2, & f(x_2) = 0 . \end{cases}$$

Continuing in this fashion, we define

$$b_1^{(k)}[\, x_k \,|\, (x_1, f(x_1)), \ldots, (x_{k-1}, f(x_{k-1}))] , \quad x_k \in D ,$$

and

$$b_0^{(k)} = 1 - b_1^{(k)}, b_0^{(k)} \geq 0 , \quad b_1^{(k)} \geq 0 ,$$

as a gambling scheme that depends only on the previously observed properly classified (training) set.

The accrued capital after all patterns $x_1, x_2, \ldots, x_n$, $n = |D|$, have been observed is

$$S_k = \begin{cases} 2b_1^{(k)}S_{k-1}\,, & f(x_k)=1 \\ 2b_0^{(k)}S_{k-1}\,, & f(x_k)=0\,, \end{cases}$$

for $k = 1,2,\ldots,n$ and $S_0 = 1$. Let

$$b = (\,(b_0^{(1)},b_1^{(1)}\,),(\,(b_0^{(2)},b_1^{(2)}\,),\,.\,.\,.\,,(\,(b_0^{(n)},b_1^{(n)}\,),$$

denote a sequence of gambling functions.

Theorem 1: For any $F \subseteq D^{\{0,1\}}$, there exists a gambling scheme b^* achieving $S_n(f) = S^* = 2^{n-log|F|}$ units, for all f in F and for all orders of presentation of the elements $x \in D$. Moreover, there exists no b that dominates b^* for all f; thus, b^* is minimax. This gambling scheme is given by the expression

$$b_1^{(k)*}(x) = \frac{|\{g \in F{:}g(x_1) = f(x_1),\ i = 1,2,\ldots,k-1,\ \text{and}\ \ g(x) = 1\}|}{|\{g \in F{:}\ g(x_1) = f(x_i),\ i = 1,2,\ldots,k-1\}|}$$

Remark: This gambling scheme simply asserts at time k, "Bet all of the current capital on the hypotheses $f(x_k) = 1$ and $f(x_k) = 0$ in proportion to the number of functions g in F that *agree* on the training set and assign the new pattern x_k to classes $g(x_k) = 1$ and $g(x_k) = 0$ respectively."

The proof will not be given here but can be found in [1].

Applications and Examples:

1. Let F be all 2^n functions $f : D \rightarrow \{0,1\}$, where $n = |D|$. Then $log|F| = n$, and $S^* = 1$. No money can be gained. The training set gives no information about future pattern classifications. This is the worst case.

2. Let D denote a set of n vectors in Euclidean d-space R^d. Let us also assume that $\{x_1,x_2,\,.\,.\,.\,,x_n\} = D$ is in *general position* in the sense that every d-element subset of D is linearly independent. Let F be the set of all

linear threshold functions on D; i.e., $f \in F$ implies there exists $w \in R^d$, $T \in R$, such that

$$f(x) = sgn(w^t x - T), \ \forall \ x \in D ,$$

where

$$sgn(t) = \begin{cases} 1, & t \geq 0 \\ 0, & t < 0 . \end{cases}$$

Then from Cover [8], we have

$$|F| = 2 \sum_{k=0}^{d} \binom{n-1}{k}, \ \forall \ d,n .$$

Using bounds derived from Stirling's approximation, it can be shown that

$$log\left(2 \sum_{k=0}^{d} \binom{n-1}{k} \right) \approx nH\left(\frac{d}{n}\right), \quad \text{for} \quad n \geq 2d ,$$

where $H(p) = -p \log p - (1-p)\log(1-p)$ is the Shannon entropy function. Thus we conclude, for $n \geq 2d$, that an amount $S_n = 2^{n(1-H(d/n))}$ can be won if in fact the n patterns are linearly separable in R^d. Note also that $H(d/n)$ is the Kolmogorov complexity of most of the linear threshold function $f \in F$. Finally, we observe that S_n is not much greater than 1 until $n \geq 2d$, at which point the behavior of S_n is exponential. This is yet more evidence that $n = 2d$ is a natural definition of the capacity of a linear threshold pattern recognition device with d variable weights.

3. Let F be the set of all functions $f : D \rightarrow \{0,1\}$ that can be represented by rth degree polynomial discriminant functions:

$$f(x) = sgn\left[\sum_{i_1,i_2,\dots,i_r} w_{i_1 i_2 \cdots i_r} x_{i_1} x_{i_2} \cdots x_{i_r} - T \right]$$

If the elements of D are in general position with respect to rth degree polynomials, we see [8] that there are precisely $2 \sum^{\bar{d}-1} \binom{n-1}{k}$ elements in F

where d' is the number of coefficients in an arbitrary rth degree polynomial in d variables. For example, for $r = 2,\ d = 2$, we have

$$f(x) = a_{11}x_1^2 + a_{22}x_2^2 + a_{12}x_1x_2 + a_1x_1 + a_2x_2 + a_0 ,$$

$$\text{and} \quad d' = 6 .$$

The point is that d' is the number of degrees of freedom of the manifold $\{x : f(x) = 0\}$. Again by the theorem, we have $S_n \geq 2^{n(1 - H(d'/n))}$, where now d' is the number of degrees of freedom of the family of separating surfaces F.

4. Suppose it is not known what degree polynomial is needed to classify D correctly. Since the degree r need take on only $(n + 1)$ values before the degree is sufficient to make an arbitrary assignment f, we merely invest an initial amount $1/(n + 1)$ in the betting system for each degree $r = 0,1,...,n$. Then the theorem becomes

$$S(f) > 2^{n(1 - H(d(f)/n)) - log(n + 1)}, \text{ for all } f : D \rightarrow \{0,1\}$$

where $d(f)$ is the number of degrees of freedom of an rth degree polynomial, and r is the minimal degree necessary to yield f.

Theorem: These results are special cases of the following theorem:

Theorem: There exists a betting scheme b^* such that the total accumulated capital satisfies

$$S(f) \geq 2^{n - K(f\,|\,D) - log(n + 1)} .$$

Comment: If f is a linear threshold function, then

$$K(f\,|\,D) \leq \log 2\ (\sum_{k=0}^{d-1} \binom{n-1}{k} + c\ .$$

Simply write a program saying "f is the ith function in the lexicographically ordered list of linear threshold functions on D^n. Thus i requires $\log 2 \left(\sum_{k=0}^{d-1} \left({n-1 \atop k} \right) \right)$ bits and c is the length of the rest of the program specified above.

Similarly, the polynomial threshold functions can be seen to be special cases of this theorem.

4. **Kolmogorov's** H_k Function.

Consider the function $H_k : \{0,1\}^n \rightarrow N$, $H_k(x) = \min_{p:l(p) \leq k} \log |S|$, where the minimum is taken over all subsets $S \subseteq \{0,1\}^n$, such that $x \in S$, $U(p) = S$, $l(p) \leq k$. This definition was introduced by Kolmogorov in a talk at the Information Theory Symposium, Tallin, Estonia, in 1974. Thus $H_k(x)$ is the log of the size of the smallest set containing x over all sets specifiable by a program of k or fewer bits. Of special interest is the value

$$k^*(x) \; = \; \min\{k{:}H_k(x) + \; k = K(x)\} \; .$$

Note that $\log |S|$ is the maximal number of bits necessary to describe an arbitrary element $x \in S$. Thus a program for x could be written in two stages: "Use p to print the indicator function for S; the desired sequence is the ith sequence in a lexicographic ordering of the elements of this set." This program has length $l(p) + \log |S|$, and $k^*(x)$ is the length of the shortest program p for which this 2-stage description is as short as the best 1-stage description p^* . We observe that x must be maximally random with respect to S -- otherwise the 2-stage description could be improved, contradicting the minimality of $K(x)$. Thus $k^*(x)$ and its associated program p constitute a minimal sufficient description for x.

Example: Let $x \in \{0,1\}^n$, $\sum_{i=1}^{n} x_i = k$. Then $k^*(x) \approx log(n + 1)$, and the associated program is "S is the set of all $x \in \{0,1\}^n$ such that $\sum x_i = k$."

Arguments can be provided to establish that $k^*(x)$ and its associated set S^* describe all of the "structure" of x. The remaining details about x are conditionally maximally complex. Thus pp^{**}, the program for S^*, plays the role of a sufficient statistic.

References

[1]. A.N. Kolmogorov, "Three Approaches to the Concept of the Amount of Information," *Problemy Peredachi Informatsii,* 1, (1965), pp. 3-11.

[2] A.K. Zhvonkin and L.A. Levin, "The Complexity of Finite Objects and the Development of the Concepts of Information and Randomness by Means of the Theory of Algorithms," Russian Mathematical Surveys 25, (1970), pp. 83-124.

[3] C.P. Schnorr, "A Unified Approach to the Definition of Random Sequences," *Math. Systems Theory,* 5, No. 3, (1971), pp. 246-258.

[4] R.J. Solomonoff, "A Formal Theory of Inductive Inference, Part I," *Information and Control,* 7, (1964), pp. 1-22.

[5] R.J. Solomonoff, "A Formal Theory of Inductive Inference, Part II," *Information and Control,* 7, (1964), pp. 224-254.

[6] T. Fine, Theories of Probability, 1974.

[7] T. Cover, "Generalization on Patterns Using Kolmogorov Complexity," *Proc. 1st Internatinal Joint Conference on Pattern Recognition,* Washington, D.C. (1973).

[8] T. Cover, "Geometrical and Statistical Properties of Linear Threshold Functions with Applications in Pattern Recognition," *IEEE Trans. Elec. Comp.*, (1965).

[9] T. Cover and S.K. Leung-Yan-Cheong, "Some Equivalences between Shannon Entropy and Kolmogorov Complexity," *IEEE Trans. on Information Theory*, Vol. IT-24, No. 3, May 1978, pp. 331-338.

[10] T. Cover, "Universal Gambling Schemes and the complexity Measures of Kolmogorov and Chaitin," Technical Report No. 12, (1974) Dept. of Statistics, Stanford University.

[12] G. Chaitin, "A Theory of Program Size Formally Identical to Information Theory," *J. of the Assoc. for Computing Machinery* Vol. 22, No. 3, July 1975, pp. 329-341.

nological
es.

on with the

The Impact of Processing Techniques on Communications

edited by

J.K. Skwirzynski

Marconi Research Centre
GEC Research Laboratories
Great Baddow, Essex, UK

1985 **Martinus Nijhoff Publishers**
Dordrecht / Boston / Lancaster
Published in cooperation with NATO Scientific Affairs Division