Kolmogorov Complexity and Model Selection

Nikolay Vereshchagin[*]

Moscow State University, Leninskie gory 1,
Moscow 119991, Russia, `ver@mccme.ru`
WWW home page: `http://lpcs.math.msu.su/~ver`

1 Stochastic strings

The goal of statistics is to provide explanations (models) of observed data. We are given some data and have to infer a plausible probabilistic hypothesis explaining it. Consider, for example, the following scenario. We are given a "black box". We have turned the box on (only once) and it has produced a sequence x of million bits. Given x, we have to infer a hypothesis about the black box.

Classical mathematical statistics does not study this question. It considers only the case when we are given results of many independent tests of the box. However, in the real life, there are experiments that cannot be repeated. In some such cases the common sense does provide a reasonable explanation of x. Here are three examples: (1) The black box has printed million zeros. In this case we probably would say that the box is able to produce only zeros. (2) The box has produced a sequence without any regularities. In this case we would say that the box produces million independent random bits. (3) The first half of the sequence consists of zeros and the second half has no regularities. In this case we would say that the box produces 500000 zeros and then 500000 independent random bits.

Let us try to understand the mechanism of such common sense reasoning. First, we can observe that in each of the three cases we have inferred a finite set A including x. In the first case, A consists of x only. In the second case, A consists of all sequences of length million. In the third case, the set includes all sequences whose first half consists of only zeros. Second, in all the three cases the set A can be described in few number of bits. That is A has low Kolmogorov complexity.[1] Third, all regularities present in x are shared by all other elements of A. That is, x is a "typical element of A".

It seems that the common sense reasoning works as follows: given a string x of n bits we find a finite set A of strings of length n containing x such that

(1) A has low Kolmogorov complexity (we are interested in simple explanations) and

---

[1] Roughly speaking, the Kolmogorov complexity of a string x is the length of a shortest program printing x. The Kolmogorov complexity of a finite set A is defined as the Kolmogorov complexity of the list of all elements of A in some fixed order, say, in the lexicographical one. For a rigorous definition we refer to [2, 4]

(2) x is a typical member of A, that is x has no regularities which allow to single out x from A.

How to define rigorously what means that x is a typical element of A? To this end we use the notion of the randomness deficiency [5]:

$$d(x|A) = \log_2 |A| - C(x|A).$$

Here $C(x|A)$ stands for the conditional Kolmogorov complexity of x given the list of A^2. The randomness deficiency has the following properties: $d(x|A)$ is non-negative for all $x \in A$ (up to a $O(\log n)$ error term) and for every finite A for almost all $x \in A$, $d(x|A)$ is negligible. Thus "random" elements of A have low randomness deficiency in A. We call x a "typical member of A" if $d(x|A)$ is small.

Strings that have explanations A with properties (1) and (2) are called *stochastic.* The first question, raised by Kolmogorov in 1983, was whether all strings are stochastic. Formally, a string x is called α, β-*stochastic* (where α, β are natural parameters) if there is a set $A \ni x$ of Kolmogorov complexity at most α such that $d(x|A) \leq \beta$. It turns out that some strings have no explanation: they are not α, β-stochastic for α and β proportional to n.

Theorem 1. *There is a constant c such that for all large enough n the following holds for all α, β. If $\alpha + \beta < n - c \log n$, then there is a string x of length n that is not α, β-stochastic. On the other hand, if $\alpha + \beta > n + c \log n$ then all strings of length n are α, β-stochastic (which is obvious).*

This theorem was first proved, in a weaker form, in [5] (with condition $2\alpha + \beta < n - c \log n$ in place of $\alpha + \beta < n - c \log n$). In the present form the theorem appeared in [7].

Note that we consider only uniform distributions on finite sets as possible probabilistic hypotheses. It is not hard to show that general distributions can be reduced to uniform ones [7].

2 Hypotheses selection

The second question is the following: assume that x is α, β-stochastic for some small α, β. How do we find a set $A \ni x$ with small $C(A)$ and $d(x|A)$? Obviously we look for sets $A \ni x$ of low complexity. To see that a set A has low complexity we somehow find a short description of A. But how can we verify that $d(x|A)$ is small? We can only verify that $d(x|A)$ is large by describing x conditional to A in much fewer than $\log |A|$ bits. That is, we can refute (2) and not prove it.

It seems that instead of verifying that $d(x|A)$ is small we do what we are able: we try to refute that. If no such refutation is found for a long time, then it becomes plausible that $d(x|A)$ is indeed small. On the other hand, assume that we have found a "constructive refutation", that is, an easily described property

---

^{2} Roughly speaking, the conditional Kolmogorov complexity of a string x given a string y is the length of a shortest program that prints x given y as an input

P of elements of A such that x has the property P but the majority of elements of A do not. In this case we can switch to a new hypothesis $A' = \{x \in A \mid P(x)\}$. We then have $C(A') \approx C(A)$ (as P has a simple description) and $|A'| \ll |A|$ (as the majority of elements of A do not satisfy P). Therefore

$$d(x|A') = \log|A'| - C(x|A')$$

is much less than

$$\log|A| - C(x|A') \leq \log|A| - C(x|A) + C(P) = d(x|A) + C(P) \approx d(x|A)$$

(the first inequality holds up to an $O(\log n)$ error term). Here $C(P)$ stands for the Kolmogorov complexity of P, which is assumed to be negligible. Thus A' is much better than A as an explanation of x.

Actually, if by any means, in our search for explanations of x, we have found a hypothesis $A' \ni x$ with lower (or equal) complexity than the current explanation A and such that $\log|A'|$ is significantly smaller than $\log|A|$, we usually switch to such A'. This strategy is essentially based on the Maximal Likelihood (ML) principle from classical statistics. Recall that ML estimator chooses a distribution μ that maximises $\mu(x)$ (among all contemplated probability distributions). In the case of uniform probability distributions, the probability that x is obtained by picking a random element of A is equal to $1/|A|$. Thus maximising $\mu(x)$ corresponds to minimising $|A|$.

So assume that we just look for an explanation that minimises $|A|$ among all simple explanations. Do we finally obtain a hypothesis with small randomness deficiency? More specifically, let $\mathrm{ML}_x(\alpha)$ stand for a set A that minimises $|A|$ among all $A \ni x$ of Kolmogorov complexity at most α. Is it true that

$$d(x|\mathrm{ML}_x(\alpha)) \leq \beta + O(\log n)$$

for all α, β-stochastic x? Below we will show that this is indeed the case.

Let us see what explanations would we infer using the ML strategy in the examples (1)–(3) from the beginning of the paper. In the first example we would certainly choose the explanation $A = \{x\}$. In the second and third examples it depends on the complexity level α. If $\alpha = 100000$, say, then the ML strategy could choose the set A consisting of all sequences having the same prefix of length α as x has (in the second example) and the set A consisting of all sequences having the same prefix of length $\alpha + 500000$ as x has (in the third example). If α is very small, say $\alpha = 0$, then there will be no explanations at all of complexity at most α. For some small α the ML strategy might find the explanations obtained by common sense reasoning. However we do not know the right value α in advance.

We see that sometimes we prefer an explanation A' to an explanation A' even if $\log|A'| \gg \log|A|$ (the explanation A' is more general than A). This happens only when $C(A') \ll C(A)$. How do we compare hypotheses of different complexity? It seems that we use the Minimum Description Length principle (MDL). We prefer that hypothesis A for which $C(A) + \log|A|$ is smaller. And

among two hypotheses with the same value of $C(A)+\log|A|$ we prefer the simpler one. The explanation of the term MDL is the following: the pair (the shortest description A^* of A, the index i of x in the list of all elements of A) is a two-part description of x. The total length of this description is $C(A) + \log|A|$. The minimal possible value for $C(A) + \log|A|$ is obviously $C(x)$ (the Kolmogorov complexity of x). Those A with $C(A) + \log|A| \approx C(x)$ are called *sufficient statistics of x*.

In the above examples (1)–(3), the common sense explanations are sufficient statistics of minimal complexity. Such sufficient statistics are called *minimal*.

If A is a sufficient statistics of x then $d(x|A)$ is negligible, as

$$d(x|A) = \log|A| - C(x|A) \le \log|A| + C(A) - C(x) + O(\log n). \qquad (1)$$

Note that sufficient statistics always exist, which is witnessed by $A = \{x\}$. Thus MDL based search always returns in the limit a hypothesis with negligible randomness deficiency.

However we are interested in *simple* explanations and not only in those having negligible randomness deficiency. If there is a simple sufficient statistic, then the MDL based search will find such statistic in the limit. But is there always such statistic provided that x is α, β-stochastic?

2.1   The case of small α

If α is small then, obviously, the question answers in positive. Indeed, let A be a set witnessing α, β-stochasticity of x? Then

$$\log|A| + C(A) \le \log|A| + \alpha \le C(x|A) + \beta + \alpha \le C(x) + \beta + \alpha \qquad (2)$$

(the last inequality holds up to an $O(1)$ error term). Thus A itself is a sufficient statistic (we assume that β is small, too). Besides, A witnesses that $\mathrm{ML}_x(\alpha) \le \log|A|$, which together with (1) and (2) implies that

$$d(x|\mathrm{ML}_x(\alpha)) \le \beta + \alpha \qquad (3)$$

(with logarithmic precision). Thus $d(x|\mathrm{ML}_x(\alpha))$ is always small provided x is α, β-stochastic for small α, β.

2.2   The case of arbitrary α

Assume now that x was drawn at random from a set A that has large complexity. Say, x was obtained by adding noise to a clean musical record y. In other words, x was drawn at random from the set A consisting of all x' that can by obtained from y by adding noise of certain type. Then with high probability $d(x|A)$ is small. That is, we may assume that x is α, β-stochastic for small β and $\alpha = C(A) \approx C(y)$. Does MDL or ML based search work well for such x? The inequalities (2) and (3) do not guarantee that any more, if $C(y)$ is large. Nevertheless, the following theorem shows that both MDL search and ML search work well.

Theorem 2 ([7]). *If x is α, β-stochastic and $\alpha \leq C(x)$, then there is a set $A \ni x$ with $C(A) \leq \alpha + O(\log n)$ and $\log |A| \leq C(x) - \alpha + \beta$ and hence $C(A) + \log |A| \leq C(x) + \beta + O(\log n)$ (the set A is a sufficient statistic).*

Note that the explanation A from the theorem witnesses

$$d(x|\mathrm{ML}_x(\alpha + O(\log n))) \leq \beta + O(\log n). \tag{4}$$

3 Structure sets of a string

The next question is whether the inequality (4) is indeed an improvement over the inequality (3). That is, are there α, β-stochastic strings (for small β and large α) that are not α', β'-stochastic for much smaller α' and, may be, slightly larger β'. More generally, what shape can have the " structure set"

$$S_x = \{\langle \alpha, \beta \rangle \mid x \text{ is } \alpha, \beta\text{-stochastic}\}?$$

The next theorem shows that S_x can have almost any shape. For instance, for all large enough n there is $n/2, O(\log n)$-stochastic string that is not $n/3, n/3$-stochastic.

Theorem 3 ([7]). *For every string x of length n and Kolmogorov complexity k the set S_x is upward closed and contains some pairs that are $O(\log n)$-close[3] to the pairs $\langle k, 0 \rangle$ and $\langle 0, n - k \rangle$. On the other hand, for all n and $k \leq n$, if an upward closed set $S \subset \mathbb{N} \times \mathbb{N}$ contains the pairs $\langle k, 0 \rangle$, $\langle 0, n - k \rangle$, then there is x of length n and complexity $k + O(\log n + C(\tilde{S}))$ such that S_x is $O(\log n + C(\tilde{S}))$-close to S. Here $\tilde{S}$ stands for the set of minimal points in S.*

By Theorem 2 the set S_x is $O(\log n)$-close to another structure set

$$L_x = \{\langle \alpha, \gamma \rangle \mid \text{there is } A \ni x \text{ with } C(A) \leq \alpha, \ C(A) + \log |A| - C(x) \leq \gamma\}.$$

Thus Theorem 3 describes also all possible shapes of the set L_x. Theorem 3 also provides a description of possible shapes of the following set:

$$P_x = \{\langle \alpha, \delta \rangle \mid \text{there is } A \ni x \text{ with } C(A) \leq \alpha, \ \log |A| \leq \delta\}.$$

This set is called the Kolmogorov's structure set, as it was defined by Kolmogorov in [3]. Indeed, by Theorem 2, the set P_x is $O(\log n)$-close to the set

$$\{\langle \alpha, C(x) - \alpha + \beta \rangle \mid \alpha \leq C(x), \ \langle \alpha, \beta \rangle \in S_x\} \cup \{\langle \alpha, 0 \rangle \mid \alpha \geq C(x)\}.$$

---

[3] We say that u is ε-close to v if the Euclidean distance between u and v is at most ε. Sets U, V are ε-close if for every $u \in U$ there is $v \in V$ at the distance at most ε from u and vice versa.

References

1. P. Gács, J. Tromp, P.M.B. Vitányi, Algorithmic statistics, *IEEE Trans. Inform. Th.*, 47:6(2001), 2443–2463.
2. A.N. Kolmogorov, Three approaches to the quantitative definition of information, *Problems Inform. Transmission* 1:1 (1965) 1–7.
3. A.N. Kolmogorov, Talk at the Information Theory Symposium in Tallinn, Estonia, 1974.
4. M. Li and P.M.B. Vitányi, *An Introduction to Kolmogorov Complexity and its Applications* , Springer-Verlag, New York, 2nd Edition, 1997.
5. A.Kh. Shen, The concept of (α, β)-stochasticity in the Kolmogorov sense, and its properties, *Soviet Math. Dokl.*, 28:1(1983), 295–299.
6. A.Kh. Shen, Discussion on Kolmogorov complexity and statistical analysis, *The Computer Journal*, 42:4(1999), 340–342.
7. N.K. Vereshchagin and P.M.B. Vitányi, Kolmogorov's structure functions and model selection, *IEEE Trans. Information Theory*, 50:12 (2004) 3265-3290.