IA703 – AIAI
(Algorithmic information and AI)

Jean-Louis Dessalles

Evaluation – January 2022

Examples of answers

Q1. Starting from a large integer n, we can form a huge number by computing $N = n! \wedge n! \ldots \wedge n!$ ($n!$ exponentiations). Can you provide an estimate of the complexity of $\log(N)$?

$\log(N)$ is obtained from n through a small program. So its complexity is about log(n).

Q2. What is Zipf's law? How is it related to complexity?

Zipf's law says that in human generated texts, the rank of words according to frequency is in inverse proportion to their frequency: $r = k/f$.
For most words, the log of inverse frequency $\log(1/f)$ is equal to Kolmogorov complexity K.
On the other hand, the rank offers a proxy to K: $K = \log(r)$. Zipf's law is explained by equalizing these to estimates of K.

Q3. On a search engine, the word 'peptide' is claimed to be found 2510 million times, while 'enzyme' is found 4870 million times. The two words are found about 150 million times together. Can you provide an estimate of the conditional Kolmogorov complexity K('peptide' | 'enzyme')?

We may write K('peptide' & 'enzyme') $\leq K$('enzyme') + K('peptide' | 'enzyme').
So K('peptide' & 'enzyme') – K('enzyme') provides an estimate of K('peptide' | 'enzyme').
If we approximate K by a log-frequency, we get : K('peptide' | 'enzyme') ≈ log(4870/150) ≈ log(32) = 5.

Q4. Astronomers collected 128 observational data about an exoplanet in a double star system. A standard hypothesis E_1 about the mass of the planet and its distance to the stars accounts for m of the 128 data points. The $(128-m)$ exceptions remain a mystery, unless the two stars follow a particular (never observed) elongated elliptic trajectory around each other (hypothesis E_2). If the two stars happen to obey E_2, it accounts for all data. However, the precision of the additional parameter required to specify E_2 is found to add 48 bits to the complexity of E_1 alone.
For which value of m does E_1+E_2 offer a better theory than E_1 alone?

The cost of the exceptions to E_1 is $(128 - m) \times \log_{2+}(128) = 7 \times (128 - m)$ bits.
This cost has to be larger than E_2's 48 bits for E_2 to be useful.
The number m of data explained by E_1 alone should be smaller than: $m \leq 128 - 48/7) \approx 121$ for E_1+E_2 to be preferred.

Q5. In your region, one cat out of 10 000 is absolutely white. Let's write C_d('white') the complexity of the concept 'white' out of context. Since the word 'white' is among the 200 most frequent words, C_d('white') ≈ 8 bits. How unexpected (in bits) would an all-white cat be, as compared with a standard cat?

The complexity of selecting one all-white cat among 10 000 cats requires a causal complexity of $C_w(wc) \approx \log_2(10000) = 14$ bits. Description complexity involves the notion of 'white' and, perhaps the notion of 'cat'. So unexpectedness reads:

$$U(wc) \leq 14 - C_d(\text{'cat'}) - C_d(\text{'white'})$$

where C_d('cat') should be omitted if 'cat' already belongs to the context.