IA703 – AIAI
(Algorithmic information and Artificial Intelligence)

Jean-Louis Dessalles

Evaluation – January 2023
Examples of possible answers

Q1. A weekly lottery issues a 6-digit number between 000000 and 999999. We suppose that all previous draws are available from the Web. Today's draw is 767433. It was 767439 three weeks ago. More generally, what is a good estimate of the complexity of today's draw if it differs by d from the k^{th} previous draw?

We need $\log_{2+}(1+k) + 1 + \log_{2+}(1+d) + O(1)$ bits to retrieve a previous draw at distance k in memory and to tell the difference (the term +1 is the bit needed to indicate the sign (positive or negative) of the difference; the constant $O(1)$ represents the code needed to switch to a memory retrieval representation). This method for describing the draw should not be more complex than a direct description. So a rigorous answer is: $\min(1+\log_{2+}(k \times d) + O(1), \log_{2+}(10^6))$.

Q2. The Normalized Google Distance (NGD) uses a Web search engine to measure the resemblance between concepts by substituting $\log(N/\#\text{hits})$ for K in the expression of the normalized information distance (reminder: Information distance ID is $K(x, y) - \min(K(x), K(y))$). Using "Bing" as Web search engine, we found:

- kimono 18 600 000 hits
- sushi 54 300 000 hits
- kimono+sushi 424 000 hits

Which value do we get for NGD(kimono, sushi), if we suppose that Bing crawled $N = 8\times10^9$ pages?

NGD(kimono, sushi) = (log(54300000, 2) - log(424000, 2)) / (log(8000000000, 2) - log(18600000, 2)) = 0.800

Q3. The number $N = 1234567891$ is a prime number. What can you say about its complexity?

A first upper bound of the complexity of n is $\log_{2+}(N) = 31$ bits.
We could bound the complexity of N by the complexity of its rank in the list of prime numbers (the prime number theorem says it is about $N/\ln(N)$, which gives a complexity of 26 bits).
If we note that $M = 123456789$ has a low complexity, then N can be described with operators such as 'integers', 'increment', 'concatenation' and the constants 9 and 1. Since these operators are simple in the context of arithmetic, one can get a representation of n of less than 12 bits (variable depending on context). The number can also be described as the first 11 digits of the Champernowne constant.

Q4. A linear regression has been performed from a dataset $\{(x_i, y_i)\}$, $I \in [1, N]$. With this linear model L, $y_i = L(x_i) + \varepsilon_i$, where $\{\varepsilon_i\}$ are the errors. So the dataset can be represented by L and $\{(x_i, \varepsilon_i)\}$. All coordinates are measured with precision δ. Express the amount of compression that the linear regression achieves.

Thanks to L, we substitute ε_i for y_i and thus need a smaller amount of information, as $|\varepsilon_i| << |y_i|$ in general. The informational gain is:

$\Sigma \log_2(1+|y_i|/\delta) - C(L) - \Sigma \log_2(1+|\varepsilon_i|/\delta)$ (note that the signs of the ε_i and y_i are counted on both sides and cancel out).

Note that L can be defined by two numbers, so its complexity $C(L)$ is limited, often negligible if N is large.

Q5. Jean-Louis bought a brand new smartphone for p €. Just d=10 days later, he dropped the phone and the window broke. Understandably, he got very frustrated. He had never broken any of his four previous smartphones' window, though he kept them about u=3 years each.

(1) Provide an estimate of the probability b of breaking the phone's window on the very first day.

(2) How unexpected (in bits) is it to break the phone on the first day?

(3) How unexpected (in bits) is it to break the phone on day d?

(1) $b \leq 1/(4 \times u \times 365)$. We take equality to get an estimate.

(2) $C_w(d=0) = \log_2(1/b) = \log_2(1460) + \log_2(u)$. $C(d=0) = 0$. Hence $U(d=0) = Cw(d=0) - C(d=0) \approx 10.5 + \log_2(u) = 12$ bits.

(3) $C_w(d) = C_w(d=0)$. $C(d) = \log_2(1+d)$. Hence $U(d) \approx C_w(d) - C(d) \approx 10.5 + \log_2(u) - \log_2(d) = 9$ bits.