Kolmogorov Complexity

Speaker Topic 13: Complexity
Artificial Life
School of Informatics and Computing
Indiana University
Spring 2011

Joseph Biberstine

How complex are these messages

- Examples
  - **1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20**
  - **1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39**
  - **1 2 1 4 1 2 1 8 1 2 1 4 1 2 1 16 1 2 1 4 1 2 1 8 1 2 1 4 1 2**
- Do they seem increasingly complex?
  - How can we quantify this?
    - "Complexity" is a word with many meanings..
    - .. but one thought to pursue might be:
      They seem increasingly **difficult to describe**

Definition

- Kolmogorov complexity = descriptive complexity

- First, we need to pick a language strong enough to describe all messages of interest
  - Turing-completeness allows all computable messages
  - Let's use English as an accessible example language
  - In many cases, we will find descriptions shorter than simply restating the message
- The Kolmogorov complexity of a message is defined to be **the length of its shortest description in this language**
  - Description should be exact and unambiguous

Kolmogorov complexity of examples

- **1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20**
  - *integers from 1 to 20*
    - Kolmogorov complexity = Length = 21

- **1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31 33 35 37 39**
  - *odd integers from 1 to 39*
    - Length = 25

- **1 2 1 4 1 2 1 8 1 2 1 4 1 2 1 16 1 2 1 4 1 2 1 8 1 2 1 4 1 2**
  - *write "1 ", then repeatedly concatenate your work to itself and double the last number. take first 30 of result.*
    - Length = 113
  - *greatest power of 2 dividing integers from 1 to 30*
    - Length = 50

Puzzles

- How complex (in our sense) are these?
  - **0 1 4 0 8 4 5 0 7 0 4 2 2 5 3 5 2 1 1 2 6 7 6 0 5 6 3 3**
  - **1 2 5 4 8 9 0 6 7 3 3 0 2 9 7 5 0 9 5 4 2 1 8 6 3 7 3 7**

Solutions(?)

- **0 1 0 3 0 9 2 7 8 3 5 0 5 1 5 4 6 3 9 1 7 5 2 5 7 7 3 1 9 5 8 7 6 2 8 8 6 5 9 7 9 3 8 1 4 4 3 2 9 8 9 6 9 0 7 2 1 6 4 9 4 8 4 5**
  - *1st 64 digits of 1/97*
    - But maybe we could do better?
      - Could only be certain if we checked **every** smaller description

- **1 2 5 4 8 9 0 6 7 3 3 0 2 9 7 5 0 9 5 4 2 1 8 6 3 7 3 7 9 6 5 4 7 1 0 2 3 8 1 5 0 7 2 3 9 6 8 4 4 1 8 0 7 9 3 6 1 9 5 4 0 8 7 2**
  - ?
  - 64 random digits
  - Is there a shorter description than the sequence itself?
    - Who knows?

Languages: Natural

- Any Turing-complete language could serve as the basis
- But which one?
- English is convenient, but it's **not rigorous** for this purpose
  - *the number of people on Earth January 1st 2000 at midnight GMT*
    - Does this describe exactly one numeric message?
      - What about people jumping or flying ...
      - Being born or dying or ...

Languages: Formal

- **Formal languages** solve this problem, but they are less intuitive than natural language
  - How do we **find descriptions** shorter than a literal restatement?
    - And if we do, how can we be certain that any one we've found is the **shortest**?

In practice

- Though Kolmogorov complexity is tricky to calculate **exactly**, except perhaps for the shortest of strings, we can get reasonable **approximations**
  - Consider a compression algorithm such as one used to create zip archives
    - It will produce descriptions of any input, trying very thoroughly to be as succinct as possible
    - Such algorithms are as good an attempt as we are likely to find in full generality

What does it measure?

- Now that we have this concept defined well and on a solid theoretical footing using a T-C formal language.. what's it telling us, anyway?
- We call it "complexity", but can we say more specifically what it means? Following Feldman, it:
  - measures "**randomness**"
    - incompressibility/unstructuredness
  - does not measure "pattern or structure or correlation or organization"

In artificial life

- Could we use Kolmogorov complexity to guide evolution in a simulation?
  - Selection could be wholly or partially determined by the complexity of the genome
    - Minimize it: Maybe we'll get elegantly simple agents!
      - Instead: Boringness
    - Maximize it: Maybe we'll get complex agents!
      - Instead: Randomness
        - Instead of generating agents with behavior that is complex in another, more desirable, sense of the word, we would more likely find a structureless noise

So it's a wash?

- Perhaps not; consider these experiments involving a fitness function:
  - When it begins to plateau across the population, add a component favoring simpler genomes
    - Maybe we can drive inessential complexity out of the solution
  - Add a small component favoring complexity and reduce the global mutation rate
    - Instead of changing alleles blindly, maybe we will occasionally pinpoint those areas that will effect the most ambitious search
- What do you think would happen?

Interesting properties

- How great can a message's Kolmogorov complexity be?
  - Always bounded by whatever literal restatement we could lazily give it, potentially plus a little padding amounting to "the following is literal:"
    - This is important, if we ignored it and encoded the message **integers between 1 and 50** as *integers between 1 and 50* we will probably cause confusion

Confusing *descriptions* with literal **messages**

E-MAIL ADDRESSES IT WOULD BE REALLY ANNOYING TO GIVE OUT OVER THE PHONE.

BY MICHAEL WARD

MikeUnderscore2004@yahoo.com

MikeAtYahooDotCom@hotmail.com

Mike_WardAllOneWord@yahoo.com

AAAAAThatsSixAs@yahoo.com

One1TheFirstJustTheNumberTheSecondSpelledOut@hotmail.com

http://www.mcsweeneys.net/links/lists/27MichaelWard.html

Interesting properties

- For theoretical purposes, the base language used doesn't matter too much
  - For any two potential base languages, there is a constant upper bound on the difference between the complexities calculated using either
    - Constant never depends on a particular message, only on the languages
    - Still, the constant almost surely enormous

Interesting properties

- Joint Kolmogorov complexity satisfies an equality reminiscent of joint entropy
  - $H(X,Y) = H(X) + H(Y \mid X)$
  - $K(X,Y) = K(X) + K(Y \mid X) + O(\log(K(X,Y)))$
  - Joint means we are to describe both
  - Conditional means we can use the given message as input
    - Consider X = **01** and Y = **0101**
      - Our description for calculating $K(Y \mid X)$ could be *X twice*
      - Our description for calculating $K(X, Y)$ could be *01, then the first two characters twice*
  - Logarithmic factor essentially allows for space in the description to explain which part encodes message X and which encodes message Y

That's all

Andrey Kolmogorov

Born	25 April 1903 Tambov, Imperial Russia
Died	20 October 1987 (aged 84) Moscow, USSR
Nationality	Russian
Known for	Probability theory, topology, intuitionistic logic, turbulence, classical mechanics, mathematical analysis