S3: Kolmogorov complexity – Conditional Kolmogorov complexity and Cox Theorem

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

Scribe:3

Student: Pranjal Umesh Kalekar

CS 7840: Foundations and Applications of Information Theory (fa24)

https://northeastern-datalab.github.io/cs7840/fa24/

Lecturers: Wolfgang Gatterbauer, Javeed Aslam

Version Jan 1, 2001 <replace with the date of submission>

Kolmogorov Complexity

For a string x, its Kolmogorov Complexity, K(x), is defined as:

K(x)=Length of the shortest program (in bits) that outputs x

Examples like the once we saw in the class:

For a string x="1111111111” → example program print("1" * 10)

The program's length is much shorter than x, so K(x) is small.

For x="1100101010"x="1100101010” → it’s random so print("1100101010")

K(x) is nearly equal to the string's length.

- Kolmogorov Complexity can theoretically help identify the simplest model that fits the data (*aligning with Occam’s Razor*).
- But as we saw that calculating Kolmogorov complexity K(X) is not possible as there is no algorithm to calculate K(X) for any arbitrary X
- It is approximated using methods like compression algorithms (e.g., gzip), which give a practical lower bound for K(x)K(x).

Conditional Kolmogorov Complexity

The conditional Kolmogorov complexity of x relative to y is defined similarly as the length of a shortest binary program to compute x, if y is furnished as an auxiliary input to the computation

The conditional Kolmogorov complexity is commonly taken to be a finite binary string. – which makes it more practical than the Kolmogorov Complexity . In the talk by Ilya Sutsekever: "An Observation on Generalization" He talks about how conditional Kolmogorov compressor is the solution to the unsupervised learning.

The way Conditional Kolmogorov Compressor works is, It finds the shortest algorithm (or program) that outputs x, using y as an input. Essentially, it compresses x by extracting patterns in x that can be explained or predicted based on y.

Example and derivation of Conditional Kolmogorov complexity

Conditional Kolmogorov Complexity extends the concept to measure the complexity of a string x given some auxiliary information/condition y.

K(x|y)=Length of the shortest program that outputs x given y

Example: Let x="1111111111"x="1111111111" and y="111”

x can be described as "repeat y three times and add 11." This makes K(x|y) very small.

One way to imagine this is, thinking if K(x|y) is smaller y has almost all the information needed to get x not a lot of additional information is needed.

Now if we look at the complexity in this way then we can see the following applications very clearly:

- Compression – this where Ilya’s talk comes in! and when he says compress all the data
- Similarity – this was a very surprising observation for me, but it is so very obvious if we have a information about a pre-existing cluster
- Feature Importance: it can help assess the importance of features by evaluating how much additional information a feature provides about the target variable.

Cox theorem

Cox's theorem is a fundamental result in probability theory and Bayesian reasoning. It states that any system of reasoning about uncertainty that satisfies certain reasonable rules is equivalent to probability theory. If you want to measure degrees of belief (how confident you are about a proposition being true) consistently and logically, you will end up using rules that align with the rules of probability.

The Key Idea of Cox's Theorem:

Cox's theorem formalizes the idea that degrees of belief should follow certain logical principles to remain consistent:

- Belief should be updated logically when new evidence is presented.
- Combining beliefs should follow consistent rules.

The theorem shows that any system adhering to these principles must:

1. Represent degrees of belief as real numbers between 0 and 1.
2. Combine beliefs according to the rules of **probability theory**:
    - Addition Rule: If you believe in multiple possibilities, their combined belief must follow the sum rule.
    - Multiplication Rule: Beliefs about independent events should combine multiplicatively.

Using Cox’s Theorem, Bayesian inference prefers simpler models by assigning them higher priors, consistent with the data and **Occam’s Razor**.

Both frameworks reject overly complex models that overfit the data; but Cox’s Theorem focuses on reasoning and updating beliefs under uncertainty and Kolmogorov Complexity focuses on the intrinsic complexity of data or models, independent of reasoning processes.

References:

- https://arxiv.org/pdf/1206.0983, Conditional Kolmogorov Complexity and Universal Probability, Paul M.B. Vitányi
- https://northeastern-datalab.github.io/cs7840/fa24/, Fnd and Apl of information theory
- https://link.springer.com/chapter/10.1007/978-3-642-15387-7_49, Clustering Based on Kolmogorov Information, Fouchal Said, Ahat Murat, Lavallée Ivan, Bui Marc & Benamor Sofiane
- https://en.wikipedia.org/wiki/Kolmogorov_complexity#:~:text=Kolmogorov%20complexity%20is%20the%20length,have%20been%20proposed%20and%20reviewed.
- http://jimbeck.caltech.edu/summerlectures/references/ProbabilityFrequencyReasonableExpectation.pdf, R.T. Cox, one of the best papers I have read on probability.
- https://www.sciencedirect.com/science/article/pii/S0888613X03000513?ref=cra_js_challenge&fr=RR-1, Constructing a logic of plausible inference: a guide to Cox’s theorem, Kevin S Van Horn