Algorithmic Interpretability

Michael Yao, Allison Chae, Mark Yatskar PhD

Learning Objectives

1. **Define** what it means for an algorithm to be interpretable and **highlight** key ways that the definition is subjective and user-dependent.

2. **Describe** the accuracy-interpretability tradeoff and why it is observed in many real-world algorithms.

3. **Reflect** on the role of interpretability in algorithms used in clinical practice.

> Food for Thought: Interpretability or Accuracy?
>
> Suppose I ask you the following question: **is 99 a prime number?** Which of the following generated answers is more helpful to you?
>
> 1. **No**, 99 is not a prime number.
> 2. **Yes**, 99 is a prime number. To figure out if a number is prime we can list out all of the numbers that are greater than 1 and less than or equal to the square root of input number. In our case, the numbers that satisfy these criteria are 2 through 9 inclusive. We then see if the original number is divisible by any of the numbers in this list - if not, then the number is prime. Otherwise, the number is composite. 99 is not divisble by any of 2 through 9 inclusive. Therefore, we can conclude from that 99 is prime.
>
> Would your answer to this question change if the original question posed instead was **what disease does this sick**

> **patient in front of me have?**
> Which is more important to you: getting the right answer (accuracy) or understanding how to approach similar problems in the future (interpretability)?

What is Interpretability?

Unlike our past few modules defining topics like fairness and anonymity exactly, it is challenging to give a rigorous, objective definition of interpretability. One commonly cited definition is **interpretability is the degree to which a human can understand *why* an algorithm made its prediction**.[1] If an algorithm is interpretable, then it is easier for someone to understand why certain predictions were made. Note that the definition of interpretability is *entirely independent* from the the accuracy of the algorithm - we only seek to explain why an algorithm made its own prediction, which may or may not be necessarily correct.

However, even this definition of interpretability remains underspecified.[2] For example, here are some other factors of algorithmic design that are closely related to - if not paramount for - interpretability:

1. **Trust**: What is an objective notion of trust? Is it confidence that a model will perform well? Do you care about how *often* an algorithm is correct, or for *which inputs* the algorithm is correct for?
2. **Causality**: Can we use algorithms to learn potential hypotheses about the world around us? Is this important in labelling a model as interpretable?
3. **Transferrability**: Do algorithms generalize to new patient populations? Can we predict when an algorithm

---

[1] Miller T. Explanation in artificial intelligence: Insights from the social sciences. Art Intel 267: 1-38. (2019). doi: 10.1016/j.artint.2018.07.007
[2] Lipton ZC. The mythos of model interpretability. Proc ICML Workshop on Human Interpretability in Machine Learning. (2016). 10.48550/arXiv.1606.03490

might generalize and when it won't? How might deployment of models alter the user's environment to simultaneously *invalidate* the model?

4. **Informativeness**: As previously explored in the **Food for Thought** discussion above, is it more important for you to have an explanation to approach future problems, or to get the correct answer? Does the model serve the role of an oracle, a colleague, or a mentor?

5. **Fair and Ethical Decision Making**: How can we be sure that predictions do not discriminate on the basis of race, age, gender, and other patient attributes?

💡 A recent study described a machine learning algorithm showing that patients are less likely to die from pneumonia if they also had asthma. Why might this be the case? Would you want to deploy this system in the ED?

In 2015, Caruana et al. showed that concurrent asthma was a predictor of lower risk of death in patients with pneumonia. This is because physicians that knew about the patient's asthma history always always treated these asthmatic patients more agressively. As a result, asthmatic patients paradoxically had *lower* death rates from pneumonia.

Imagine that we naively implement this model for patient triage without understanding the reasoning behind its predictions. In this case, model deployment would *prevent* physicians from treating asthmatic patients more agressively because the model would label these patients as having a low risk of death. In other words, deploying the model would actually end up invalidating the model and *hurt* patients!

Caruana R, Lou Y, Gehrke J, Koch P, Sturm M, Elhadad N. Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. Proc Conf Knowl Disc Data Mining: 1721-30. (2015). doi: 10.1145/2783258.2788613

These factors that contribute to how we label if an algorithm is interpretable often vary due to a number of factors:

1. **Difficulty of the Task**: Algorithms trained to perform complex, domain-specific tasks are inherently less interpretable by the average user due to the nature of the task itself.
2. **Expertise of the User**: A domain expert may require less explanation in order to call an algorithm interpretable compared to someone with less experience in the field.
3. **Expertise with the Algorithm**: Just like with any other software, technicians with years of experience using an algorithm are more likely able to explain its predictions more due to having more experience using the technology.

Discussion Questions

💡 What other factors might influence the subjectivity of the interpretability of an algorithm?

Which of the following algorithms are interpretable?

Mean Arterial Pressure (MAP)

On one end of the spectrum, clinical algorithms like computing the mean arterial pressure (MAP) are pretty clearly interpretable. We can exactly right down the formula to compute this quantity as

$$\text{MAP} = \frac{1}{3}(\text{Systolic Blood Pressure (mmHg)}) + \frac{2}{3}(\text{Diastolic Blood Pressure (mmHg)})$$

We might even be able to reason *why* this formula works - heuristically, the arterial pressure might spend about 2/3rds of the time in diastole and 1/2rd in systole, and so the time-weighted average of these two quantities is the MAP.

MELD Score

The MELD Score is a clinical algorithm used quantifying the degree of end-stage liver disease in potential transplant candidates. Similar to MAP, the MELD score also has an exact formula:

$$\text{MELD} = 9.57 \times \log(\text{Cr}) + 3.78 \times \log(\text{Bilirubin}) + 11.20 \times \log(\text{INR}) + 6.43$$

Is this equation still interpretable? From the equation above, we still clearly have transparency into how a MELD score is calculated, but the equation itself is a little more complicated and may not be easily understand by everyone. After looking at the above equation, we're still left with a number of remaining questions: How were the decimal coefficients derived? Why is there a logarithmic relationship between the MELD score and patient lab values?

A Machine Learning Algorithm

Suppose we now have a ML algorithm that predicts a patient's risk of breast cancer given their genomic data. Such algorithms are often referred to as ***black-box algorithms*** because the algorithm's user cannot see the inner workings of the algorithm. However, is such an algorithm truly "black-box"? Similar to the MELD Score, I can exactly write down the specific formula for the algorithm, with all of its inputs, internal functions, decimal coefficients, etc. It would be an incredibly long and complex equation, but *any* ML algorithm can be written down exactly just like the MELD score and MAP calculations above.

> Does being able to write down a mathematical formula for an algorithm (even if it's extraordinarily complex as with machine learning models) mean that the algorithm is interpretable? Why or why not?
>
> Unfortunately, we don't have a "right" answer to share with you here. Comparing the above three scenarios, some experts may describe them as interpretable and humans just lack the ability to understand them. In other cases, some might claim that this is not sufficient. Is it the algorithm's "job" to explain itself to humans when we lack the ability to understand how they work?

A Probabilistic Algorithm

Finally, consider the following algorithm that utilizes a fair two-sided coin: if I flip the coin and it lands heads, then I admit a patient from the ED. Otherwise, I discharge them and send the patient home. Is this algorithm (or any probability-based algorithm) "interpretable"?

An Accurate Algorithm Trained on An Unknown Dataset

After reading a recently published paper on a new machine learning algorithm to diagnose a rare disease, you try testing the algorithm on your own patients' data and find that it has almost perfect accuracy! However, the paper does not include any details about how the model was trained - including any information on the patient demographics in the published study.

Interpretability versus Accuracy

A key insight that we hope you take away from considering the discussion question posed above is that there is often times a trade-off between the *complexity* and *interpretability* of an algorithm. If an algorithm is more complex, such as machine learning models and the MELD score, then they may be less interpretable. At the same time, algorithms that are more complex can often times represent more complex relationships between inputs and outputs, leading to better predictive accuracy. In other words, we have the following:

> The Accuracy-Interpretability Tradeoff
>
> The more accurate an algorithm model is, the less likely it is to be interpretable.[3]

What are some strategies for building interpretable algorithms?

1. **Feature Attribution** is one way that researchers try to build interpretable algorithms is by being able to attribute the individual contributions from input variables to the final model prediction. A good example of this is linear regression. Suppose that we have two variables - *A* and

---

[3]There's a great blog post discussing the accuracy-interpretability tradeoff in more detail here: Ndungula S. Model accuracy and interpretability. Medium. (2022). Link to article

B - that contribute to a final output y according to the following model:

$$y = m_A A + m_B B + y_0$$

If $|m_A| \gg |m_B|$, then we know that the the importance of the variable A to the final score y is greater than that of variable B.

2. **Chain-of-Thought Reasoning** is a technique used specifically with large language models (LLMs) like ChatGPT. The goal of CoT prompting is to ask an algorithm to generate a step-by-step explanation or reasoning process before arriving at a final answer. Savage et al. (2024). include illustrative examples of CoT reasoning as it pertains to clinical medicine.

3. **Post-Hoc Explanations** refers to analyzing the outputs of an algorithm to see if we can better understand how algorithms work by analyzing a lot of their past outputs. Unfortunately, this body of work on interpretability is often implemented as "fishing for explanations" in practice. In general, we want to avoid generating hypotheses for how a model might work *after* the data has already been observed. It is important to carefully consider how post-hoc explanations are generated and how they may potentially lead to biased interpretations of algorithms.

4. **Constraining How Input Data is Represented** is another way to achieve model interpretability. For example, we might want an algorithm to only focus on particular clinical lab values to arrive at a prediction based on prior domain-specific knowledge of the pathophysiology of a disease.

What other methods can we use to make algorithms more interpretable?

💡 Is the CoT "explanation" generated by a large language model truly how the LLM reasons to get to a predicted answer? Turpin et al. Proc NeurIPS (2023) explain how this is not always the case.

Evidence-Based Medicine Discussion

Do algorithms need to be interpretable in order for clinicians to leverage them for patient care?

---

1. Overview Article

Imrie F, Davis R, van dr Schaar M. Multiple stakeholders drive diverse interpretability requirements for machine learning in healthcare. Nat Mach Intell 5: 824-9. (2023). doi: 10.1038/s42256-023-00698-2

tl;dr: Machine learning (ML) algorithms are becoming increasingly commonplace in healthcare settings. Key stakeholders in healthcare systems - such as algorithm developers, researchers, clinicians, and patients - often have different (and sometimes conflicting) definitions for interpretability of different algorithms used in clinical practice.

---

2. Yes, interpretability ensures that algorithms are aligned with clinical reasoning.

Antony M, Kakileti ST, Shah R, Sahoo S, Bhattacharyya C, Manjunath G. Challenges of AI driven diagnosis of chest X-rays transmitted through smart phones: A case study in COVID-19. Sci Rep 13: 18102. (2023). doi: 10.1038/s41598-023-44653-y. PMID: 37872204

tl;dr: Retrospective study using multiple publicly available chest X-ray (CXR) imaging datasets from approximately 40,000 patients. Researchers found that state-of-the-art machine learning models could accurately predict which patients had COVID-19 from CXR imaging studies, but were actually diagnosing patients by focusing on parts of the CXR scans completely outside of the lung fields, and even outside the patient's body in certain instances.

---

3. No, using black-box models allows us to discover new clinical insights and provide better care.

Ling J, Liao T, Wu Y, Wang Z, Jin H, Lu F, Fang M. Predictive value of red blood cell distribution width in septic

shock patients with thrombocytopenia: A retrospective study using machine learning. J Clin Lab Anal 35(12): e24053. (2021). doi: [10.1002/jcla.24053](#). PMID: 34674393 **tl;dr**: The red blood cell distribution width (RDW) is a lab value most commonly used in the workup of anemias. However, a retrospective study using the [a large patient dataset](#) showed using non-interpretable machine learning methods that red blood cell distribution width (RDW) was the second most important lab value in predicting 28-day mortality from sepsis. Non-interpretable algorithms therefore helped clinicians "discover" new clinical applications of the RDW.

Summary

Interpretability is a subjective property of an algorithm that characterizes the degree to which a human can understand why and how and algorithm made its prediction. There are many reasons why interpretability can vary, including the difficulty of the clinical task, the complexity of the algorithm, and expertise of the user among many others. While there is often a tradeoff observed between accuracy and interpretability in practice, many experts believe that interpretability is important for algorithms used in patient care.

Additional Readings

1. Model interpretability. Amazon Web Services Whitepapers. Accessed 20 May 2024. [Link to article](#)
2. Amann J, Blasimme A, Vayena E, Frey D, Madai VI. Explainability for artificial intelligence in healthcare: A multidisciplinary perspective. BMC Med Inform Decis Mak 20(1): 310. (2020). doi: [10.1186/s12911-020-01332-6](#). PMID: 33256715
3. Teng Q, Liu Z, Song Y, et al. A survey on the interpretability of deep learning in medical diagnosis. Multimed Syst 28(6): 2335-55. (2022). doi: [10.1007/s00530-022-00960-4](#). PMID: 35789785