Eric Sullivan, Nina Scolieri
Amulya Yadav
IST 402
10/10/2019

Week 6 Write Up

Week 6 focuses on interpretability, which is the degree to which a human can understand the cause of a decision that a machine learning model makes. There are arguments both for and against why interpretability is necessary. However, it has been concluded that interpretability is needed when problems or objectives aren't completely specified or when there is an incompleteness in problem formalization. It also allows humans to figure out if the ML model is meeting a specific set of criteria with its predictions.

Interpretability is desirable for the following 5 reasons: trust, causality, transferability, informativeness, and fair and ethical decision making. Trust is desirable because people want more than just good performance on standard metric models. If a user is able to understand a model, they feel more comfortable using certain systems. People want to know that the systems they are using won't discriminate based on race, gender, sexual orientation, or other identifiable characteristics.

Causality is important because most ML models are based on correlations, not causations. People can draw from certain correlations and try to interpret causal relationships from that data. Without an interpretable model, these correlations would not be understandable.

Transferability is important because machine learning models must need to be able to perform well on tasks it already learned from and in addition, should be able to transfer its knowledge to slightly different tasks. An example of this would be a model that is trained to recognize pictures of a certain animal such as a panda. If noise is added to that model, it may inaccurately predict an image as a different animal or object. Without interpretability, it would be unclear as to why the model was making inaccurate predictions.

Informativeness is useful when people don't intend to use machine learning models for decision-making purposes. Accuracy is not a factor in this circumstance because the model is not being used to make a decision. You cannot have informativeness for your model if you cannot use interpretability to understand its output.

Fair and ethical decision making is necessary because people want the models they are utilizing to be non-discriminatory. An example of this would be the recidivism prediction. In this case, a machine learning model could be used to decide whether a person with a criminal record should be granted bail or be sent to prison. Because criminal offenses can vary, it is important that an ML model does not make decision-based on characteristics such as a person's race, gender, age, etc. Without an interpretable model, we would not be able to understand why or how unfair predictions were being made, if this was the case.

There are arguments both for and against interpretability. Interpretability is important for instances where you might have a causal relationship. It helps to provide clarity in areas where completeness is lacking. An example of this would be an ML model that predicts a medical patient's risk of getting pneumonia. This model predicts that people with asthma have a lower chance of contracting pneumonia, which is not true. With an interpretable model, it would be more evident that people with asthma have taken more preventative measures so that they don't develop pneumonia, which is why their risk is lower. The data would show that there is a positive correlation between the amount of time it would take to get medical care and there risk from dying from pneumonia. This also highlights why causal relationships are so important.

An argument against interpretability is that we as humans don't need to understand the inner workings of every model. For example, Netflix often makes recommendations for movies on numerous different factors based on data from its users. These recommendations have very low impact on users' lives. Someone may not fully understand how Netflix is able to recommend these movies to its viewers. But people are usually confident that the system will make accurate predictions most of the time.

This concludes the key takeaways from week 6 readings and lectures. Ranging from the Justice System to Netflix, interpretability in machine learning models is becoming more prevalent in the environment around us. The readings were an interesting and easy to follow, as they covered a debate between two prominent figures in the machine learning community. It's important to have discussions both for and against interpretability in models to acknowledge their ability to benefit us or potentially harm us. Overall, this week was very interesting and we hope future lessons build on the lessons learned.