Module 22: Interpretability and Causality in Models

Quick Reference Guide

Learning outcomes:

- Identify examples of interpretability issues.
- Describe what makes a model more or less interpretable.
- Apply integrated gradients to a neural network.
- Identify potential interpretability limitations of current AI models.
- Use a causal model to determine updates to a pricing model using third-party data.

In this module, there are two approaches to model interpretability, practical applications of Shapley values and integrated gradients for auditing black box models by delving into the complexities of evaluating medical AI devices. The examples highlight the importance of ensuring models are focusing on the right features and illustrate the use of integrated gradients to predict emotions in images. Further, it examines interpretability issues in AI applications and explores beyond interpretability with causal models. Deep learning involves building complex representations of unstructured data and how models can generate new content. Due to the complexity of the models, it can be difficult to understand how they arrive at their decisions, which raises concerns about interpretability and causality.

Black box models are interpretable. The example of predicting the risk of cervical cancer based on several predictive features, demonstrates how the Shapley values technique can be used to calculate the contribution of each input feature to the model's output, providing a more interpretable measure of feature importance. Understanding which features are most important to the model's decision-making process increases trust in the model and highlights ways to improve its performance.

Shapley values are effective in interpreting black box models, but they are computationally expensive. Integrated gradients, developed by researchers at Google, provide a more efficient approach to cutting holes into black boxes for interpreting black box models, particularly with complex inputs such as images. They are also visual and easy to interpret, measuring how sensitive the model is to changes in its inputs.

Further in the module it can be seen that there are limitations to AI medical devices. Several AI devices aim at spotting and alerting doctors to suspected blood clots in the lungs. Some analyze mammograms and ultrasound images for signs of breast cancer, while others examine brain scans for signs of hemorrhage. Cardiac AI devices can now flag a wide range of hidden heart problems. A new study led by researchers at Stanford, some of whom are themselves developing devices, suggests that the

evidence isn't as comprehensive as it should be and may miss some of the peculiar challenges posed by artificial intelligence. Many devices were tested solely on historical — and potentially outdated — patient data. Few were tested in actual clinical settings, in which doctors were comparing their own assessments with the AI-generated recommendations. And many devices were tested at only one or two sites, which can limit the racial and demographic diversity of patients and create unintended biases.

The limitations of interpretability in machine learning and the desire for causality is further broken down using a personal YouTube playlist as an example. The questions of how much of the success of recommended videos is due to their qualities versus their position in the list. To address this problem, place training models on three inputs — the video's content, the user's preference, and the position of the video — this demonstrates how the approach works.

In conclusion, in this module we not only understand how to build state of the art models, we also are beginning and have begun to appreciate some of the subtleties and what we need to do to address those subtleties when we take these deep learning models and put them into production in practice.