CIS 8392
Topics in Big Data Analytics

#Explainable AI

Yu-Kai Lin

Agenda

In this session, we will look into an emerging topic in machine learning: **Explainability**

- Importance of Explainability
- Taxonomy of Explainability Methods
- Scope of Explainability
- Evaluation of Explainability
- Properties of Explanations

[**Acknowledgements**] The materials in the following slides are based on the source(s) below:

- Interpretable Machine Learning by Christoph Molnar
- AAAI 2020 Tutorial on Explainable AI

Explainability in AI and ML

- The vast majority of machine learning (ML) models are designed to make good predictions, and they typically perform really well in that regard
- As artificial intelligence (AI) becomes more advanced, humans are challenged to comprehend and retrace how the algorithm came to a result
- **Explainability** is the degree to which a human can understand the cause of a decision
    - The higher the explainability of a ML model, the easier it is for someone to comprehend why certain decisions or predictions have been made
    - A model is better explainable than another model if its decisions are easier for a human to comprehend than decisions from the other model

Explainability or Interpretability?

- In the context of AI/ML, we often use the terms "explainability" and "interpretability" interchangeably
    - You could argue that interpret and explain have different definitions in the dictionary. But in my view, the differences are really not meaningful here.
- Since AI is more than just ML, explainable AI (XAI) involves more than interpretable ML (IML). However, since we are mainly interested in ML (and not other areas of AI, such as planning, search, robotics, etc.), XAI and IML are equivalent in the scope of this lecture.

The dark secret at the heart of AI

> No one really knows how the most advanced algorithms do what they do. That could be a problem.

Will Knight

The dark secret at the heart of AI

Image source: https://dilbert.com/strip/2000-01-03

Glass-box vs. black-box ML models

- Glass-box models are the ones that intrinsically permit humans, at least the experts in the domain, to understand how a prediction was made
- Black-box models, on the other hand, are extremely hard to explain and can hardly be understood even by domain experts
    - How difficult? Check out the article "Visualizing CNN architectures side by side with mxnet" by Joseph Paul Cohen to see some examples
- Aside from certain simple models (regressions, decision trees, etc.), most AI/ML models are black-box models and are not designed to offer explanations

Accuracy vs. Explainability

- There is often a clear trade-off between accuracy vs. explainability
- To increase accuracy, ML models often rely on complex non-linear functions (e.g., deep learning) or combine multiple models (e.g., random forests)
- As a result, these make it very difficult to interpret what is happening inside an algorithm

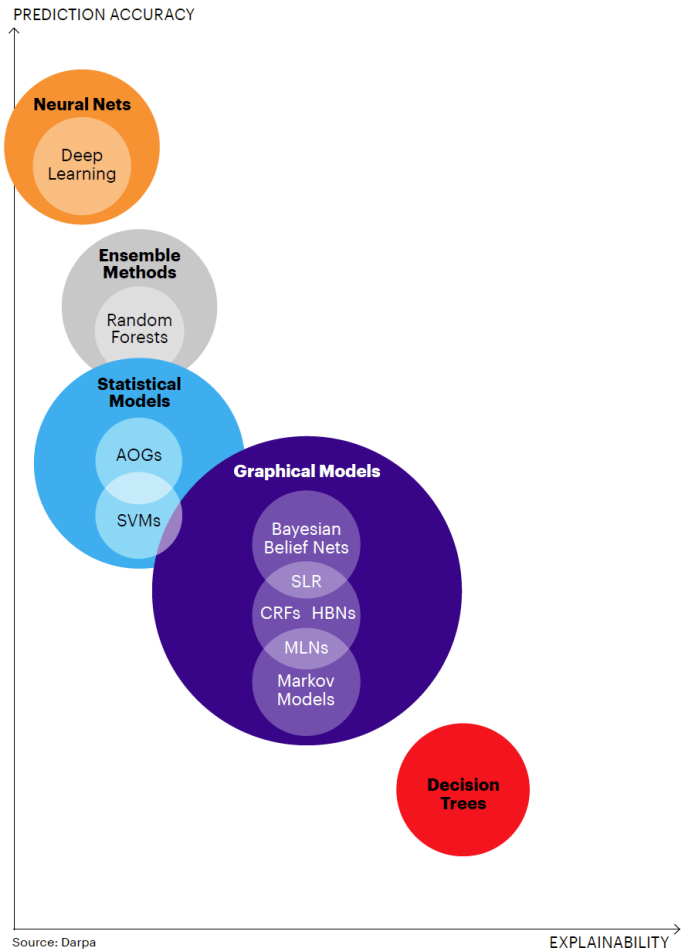


Image source: Explainable AI: The next stage of human-machine collaboration by Accenture

Explainability is essential for ...

- Model debugging - Why did my model make this mistake?
- Feature engineering - How can I improve my model?
- Detecting fairness issues - Does my model discriminate?
- Human-AI cooperation - How can I understand and trust the model's decisions?
- Regulatory compliance - Does my model satisfy legal requirements?
- High-risk applications - Healthcare, finance, judicial, ...

Growing global AI regulation

- **GDPR**: Article 22 empowers individuals with the right to demand an explanation of how an automated system made a decision that affects them
- **Algorithmic Accountability Act 2019**: Requires companies to provide an assessment of the risks posed by the automated decision system to the privacy or security and the risks that contribute to inaccurate, unfair, biased, or discriminatory decisions impacting consumers
- **California Consumer Privacy Act**: Requires companies to rethink their approach to capturing, storing, and sharing personal data
- **Washington Bill 1655**: Establishes guidelines for the use of automated decision systems to protect consumers, improve transparency, and create more market predictability
- **Massachusetts Bill H.2701**: Establishes a commission on automated decision-making, transparency, fairness, and individual rights
- **Illinois House Bill 3415**: States predictive data analytics determining creditworthiness or hiring decisions may not include information that correlates with the applicant race or zip code

Source: AAAI 2020 Tutorial on Explainable AI

Importance of explainability

XAI enables **model governance** for fairness, accountability, and transparency

Model explainability is critical for data scientists, auditors, and business decision makers alike to ensure compliance with company policies, industry standards, and government regulations:

- Data scientists need the ability to explain their models to executives and stakeholders, so they can understand the value and accuracy of their findings. They also require explainability to debug their models and make informed decisions about how to improve them.
- Legal auditors require tools to validate models with respect to regulatory compliance and monitor how models' decisions are impacting humans.
- Business decision makers need peace-of-mind by having the ability to provide transparency for end users. This allows them to earn and maintain trust.

Tech giants are embracing XAI

Microsoft

- Model interpretability in Azure Machine Learning
- Fairlearn: A toolkit for assessing and improving fairness in AI

Google

- Explainable AI on Google Cloud
- The What-If Tool: Code-free probing of ML models

Facebook

- Captum - A model interpretability library for PyTorch
- Introducing explainability to ad and news feed algorithms

IBM

- AI Explainability 360

Taxonomies of explainability methods

- **Intrinsic or post hoc?**
  - This criteria distinguishes whether interpretability is achieved by restricting the complexity of the ML model (intrinsic) or by applying methods that analyze the model after training (post hoc).
- **Model-specific or model-agnostic?**
  - Model-specific interpretation tools are limited to specific model classes, e.g., xgboostExplainer, randomForestExplainer, and DeepLIFT
  - Model-agnostic tools can be used on any ML model and are applied after the model has been trained (post hoc). These agnostic methods usually work by analyzing feature input and output pairs. By definition, these methods cannot have access to model internals such as weights or structural information.
- **Local or global?**
  - Does the interpretation method explain an individual prediction or the entire model behavior? Or is the scope somewhere in between?

Different scope of explainability

Algorithm Transparency: *How does the algorithm create the model?*

- Algorithm transparency is about how the algorithm learns a model from the data and what kind of relationships it can learn.

Global, Holistic Model Interpretability: *How does the trained model make predictions?*

- You could describe a model as interpretable if you can comprehend the entire model at once (Lipton 2016). Global model interpretability is very difficult to achieve in practice. Any model that exceeds a handful of parameters or weights is unlikely to fit into the short-term memory of the average human.

Global Model Interpretability on a Modular Level: *How do parts of the model affect predictions?*

- For linear models, the interpretable parts are the weights, for trees it would be the splits (selected features plus cut-off points) and leaf node predictions. You can easily understand a single weight.

Local Interpretability for a Group of Predictions: *Why did the model make specific predictions for a group of instances?*

- Model predictions for multiple instances can be explained either with global model interpretation methods (on a modular level) or with explanations of individual instances.

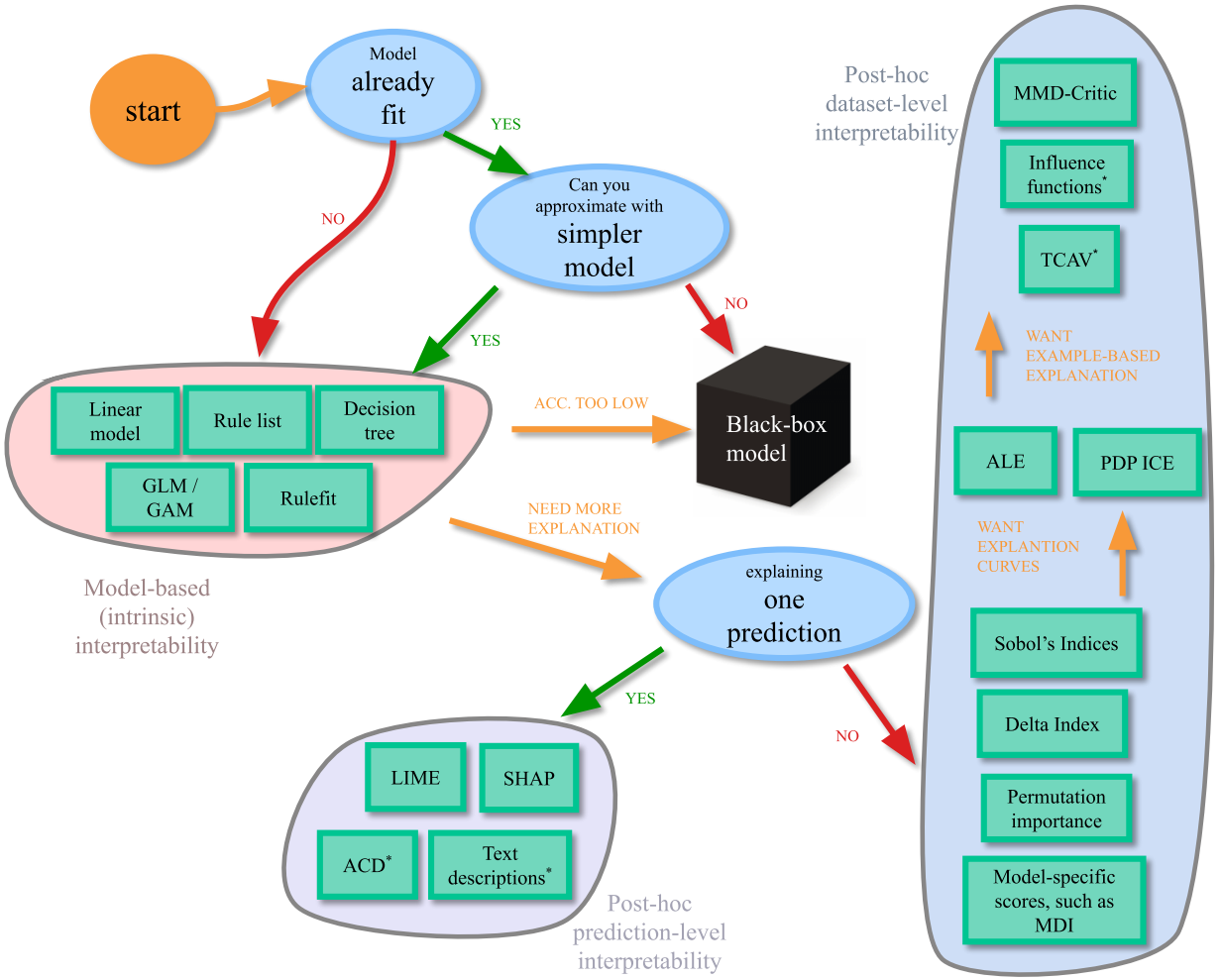
Local Interpretability for a Single Prediction: *Why did the model make a certain prediction for an instance?*

- You can zoom in on a single instance and examine what the model predicts for this input, and explain why. If you look at an individual prediction, the behavior of the otherwise complex model might behave more pleasantly. Locally, the prediction might only depend linearly or monotonically on some features, rather than having a complex dependence on them.

Properties of explanations

Comprehensibility		How much effort is needed for a human to interpret it?
Succinctness		How concise is it?
Actionability		How actionable is the explanation? What can we do with it?
Reusability		Could it be interpreted/reused by another AI system?
Accuracy		How accurate is the explanation?
Completeness		Does the "explanation" explain the decision completely, or only partially?

Image source: Explainable AI: The next stage of human-machine collaboration by Accenture



* Denotes that a method only works on certain models (e.g. only neural networks)

interpretability cheat-sheet

View on github

Based on this interpretability review and the sklearn cheat-sheet.
More in this book + these slides.

Summaries and links to code

RuleFit – automatically add features extracted from a small tree to a linear model

LIME – linearly approximate a model at a point

SHAP – find relative contributions of features to a prediction

ACD – hierarchical feature importances for a DNN prediction

Text – DNN generates text to explain a DNN's prediction (sometimes not faithful)

Permutation importance – permute a feature and see how it affects the model

ALE – perturb feature value of nearby points and see how outputs change

PDP ICE – vary feature value of all points and see how outputs change

TCAV – see if representations of certain points learned by DNNs are linearly separable

Influence functions – find points which highly influence a learned model

MMD-CRITIC – find a few points which summarize classes

Source: https://twitter.com/csinva_/status/1138811752769101825