BIOS 835: Interpretable Machine Learning

November 29, 2023

What is Interpretable Machine Learning?

- Interpretability in machine learning (IML) refers to the degree to which a human can understand the cause of a decision made by a machine learning model.
- It's a critical aspect, especially in fields that affect human lives.
- IML is a big topic, there a whole book about it.
- Interpretability is not required if the model has no significant impact.

"Imagine someone named Mike working on a machine learning side project to predict where his friends will go for their next holidays based on Facebook data. Mike just likes to surprise his friends with educated guesses where they will be going on holidays. "

Healthcare Diagnosis

- Machine learning models are increasingly used to diagnose diseases from medical images, like MRI scans or pathology slides. Interpretability is essential here to:
  - Understand the model's reasoning for its diagnosis.
  - Gain trust from medical professionals who will act on these diagnoses.
  - Identify if the model is using sensible medical indicators or if it's being misled by irrelevant artifacts in the data.

Personalized Medicine

- In personalized medicine, interpretability helps in:
  - Understanding why certain treatments are recommended based on a patient's genetic makeup.
  - Ensuring that it is actually genetics that are driving the results.
  - Allowing doctors to combine machine-generated insights with their clinical expertise.
  - Ensuring that patients receive treatments that are effective for their specific conditions.

IVF

- In modern IVF, black box models that have not been subjected to randomized controlled trials determine which embryos to transfer to the woman.
- AI is used as an embryo selection tool to improve the success rate per transfer.
- The ethical issues in this process are significant, partly because it involves creating new people.
- It could lead to:
  - Misrepresentation of Patient Values (maybe the model prefers females)
  - Health and Well-Being of Future Children
  - Impacts of Disvaluing Diversity (e.g., disabilities)

COMPAS

- Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) provides a black box proprietary algorithm for predicting criminal recidivism.
- The COMPAS algorithm has been accused of being racially biased (link), though this has never been proven.
- Wang et al. (2020) compared the COMPAS model with a very simple decision tree and found that it predicted recidivism more accurately.

Interpretable Machine Learning

- Here's a definition of IML from Rudin et al. (2021):

  "An interpretable machine learning model obeys a domain-specific set of constraints to allow it (or its predictions, or the data) to be more easily understood by humans. These constraints can differ dramatically depending on the domain."

- A typical interpretable supervised learning setup, with data $\{z_i\}_i$, and models chosen from function class $\mathcal{F}$ is:

$$\min_{f \in \mathcal{F}} \sum_i L(f, z_i) + C \cdot \text{InterpretabilityPenalty}(f)$$

  such that f is within an interpretability constraint.

Interpretable Machine Learning

- These constraints aim to make the resulting model f – or its predictions – more interpretable.
- The interpretability penalties or constraints depend on the model and application.
  - For most regression/classification: sparsity of the model, monotonicity with respect to a variable,
  - Computer vision: case-based reasoning or disentanglement, and visual understanding of intermediate computations
  - Generative constraints (e.g., laws of physics)

Interpretable Machine Learning

- Creating interpretable models can sometimes be much more difficult than creating black box models for many different reasons
  1. Solving the optimization problem may be computationally hard, depending on the choice of constraints and the model class $\mathcal{F}$.
  2. When one does create an interpretable model, one invariably realizes that the data are problematic and require troubleshooting, which slows down deployment (but leads to a better model).
  3. It might not be initially clear which definition of interpretability to use.

Interpretability and accuracy

- Rudin et al (2021) state that:

  "It is important not to assume that one needs to make a sacrifice in accuracy in order to gain interpretability. In fact, interpretability often begets accuracy and not the reverse. Interpretability versus accuracy is, in general, a false dichotomy in machine learning."

- **One clear example:** sparse linear models can be more accurate than dense linear models.

Interpretability and the scientific process

- Knowledge Discovery process

$$\text{Raw Data} \rightarrow \text{Target Data} \rightarrow \text{Preprocessed data} \rightarrow \text{Transformed data} \rightarrow \text{(modeling) Patterns} \rightarrow \text{Knowledge}$$

  How can we tune the processing of the data, the loss function, or the evaluation metric if we can't understand how the model works?
- Where was Q1 in HW5 in this processes? (static vs. raw data)
- Further, even for the most difficult and complicated computer vision problems, there are examples of models that gain substantial interpretability and do not sacrifice accuracy.

Interpretabe ML Models

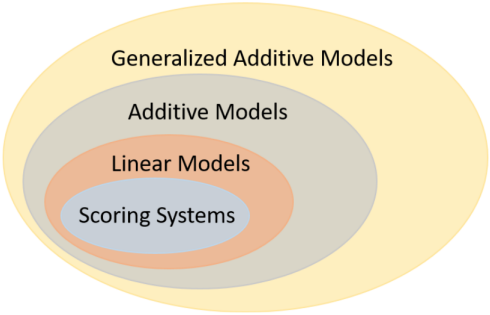
- **Transparent Models:** These are models that are inherently interpretable.
- Examples include linear regression, logistic regression, or decision trees.
- Their decision-making process is relatively straightforward and can be followed by humans.
- Nearest-neighbor methods are also relatively transparent.

Scoring Systems

Patient screens positive for obstructive sleep apnea if Score >1			
1.	age ≥ 60	4 points	.....
2.	hypertension	4 points	+ .....
3.	body mass index ≥ 30	2 points	+ .....
4.	body mass index ≥ 40	2 points	+ .....
5.	female	-6 points	+ .....
	Add points from row 1-6	Score	= .....

Table 2: A scoring system for sleep apnea screening (Ustun et al., 2016). Patients that screen positive may need to come to the clinic to be tested.

Linear models and extensions

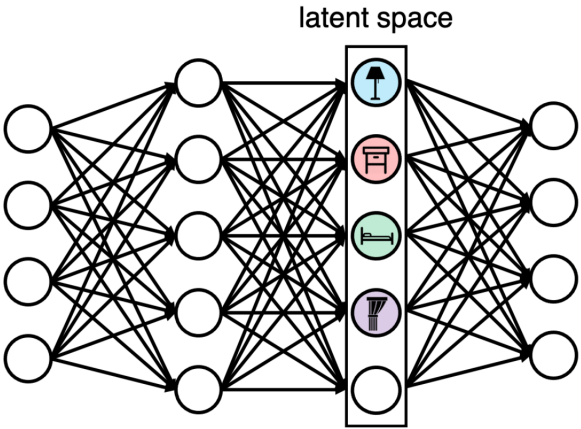


$$g\{E(Y)\} = f(X_1) + f(X_2) + \ldots + f(X_P)$$

Interpretabe ML Models

- Deep neural networks (DNNs) are the quintessential "black box" because the computations within its hidden layers are typically uninterpretable.
- The concept of "complete supervised disentanglement" in DNNs relates to the idea of separating or disentangling the different factors of variation in the data in a manner that's aligned with specific, known labels.
- It's about making the representations within a neural network reflect distinct and interpretable features of the input data, with supervision (i.e., constraints) to ensure alignment with relevant labels or factors.
- For example, say we are trying to determine if a picture is a lamp, bed, nightstand, or curtain.

Disentanglement of DNNs



Disentanglement of DNNs

- Supervision (i.e., constraints) involves using labeled data to guide the disentanglement process.
- The model learns to isolate factors of variation that correspond to the labels provided in the training data.
- By aligning the disentangled factors with known labels, it becomes easier to interpret what each part of the model's representation is capturing.
- Complete supervised disentanglement in neural networks represents a significant step towards more interpretable and fair AI systems. However, achieving true disentanglement is non-trivial and is an active area of research within the AI community.

Interpretabe ML through DR

- Earlier in the class, we discussed many different dimension reduction (DR) methods, such as PCA, t-SNE, UMAP, and PaCMAP.
- DR can help us gain insight and build hypotheses through data visualization of the underlying structure.
- Biases or pervasive noise may be illuminated.
- They take a high-dimensional vector and project it in a low-dimensional space (commonly 2-dimensional). The constraint of mapping to 2 dimensions is an interpretability constraint.
- Thus, even though we don't understand how DR methods reduce the dimensions, they are constrained and interpretable (by our earlier definition).

Post-hoc Interpretability

- Explainable AI (XAI), is where one attempts to explain a black box using an approximation model, derivatives, variable importance measures, or other statistics.
- To some, XAI is a pointless endeavor as it doesn't adhere to the initial definition of interpretable ML we first discussed.

  "An interpretable machine learning model obeys a domain-specific set of constraints to allow it (or its predictions, or the data) to be more easily understood by humans. These constraints can differ dramatically depending on the domain."

- However, we'll review some of these approaches (we've many of these already).

Global vs Local Methods

- Global interpretability methods aim to provide an understanding of the model's behavior as a whole.
  - They offer an overview of the model's decision logic.
  - They can identify general trends and patterns the model has learned from the data.
  - They are useful for understanding the model's overall behavior, which can be important for debugging, model validation, and gaining stakeholder trust.
- Local interpretability methods focus on individual predictions and aim to explain the decisions made in specific instances.
  - They explain why the model made a certain prediction for an individual instance.
  - They can help to understand model behavior for particularly interesting or important single predictions.
  - They are crucial when individual accountability is important, such as in loan approvals or medical diagnoses.

Feature Importance Scores (Generally Global)

- **Permutation Feature Importance:** Evaluates the change in the model's performance when the values of a single feature are randomly shuffled.
- **Gini Importance or Mean Decrease in Impurity:** Used in tree-based models, it measures each feature's contribution to the homogeneity of the nodes and leaves.
- **SHAP (SHapley Additive exPlanations):** Utilizes game theory to attribute the contribution of each feature to each prediction. It is based on Shapley values, which use game theory to assign credit for a model's prediction to each feature or feature value **(big topic can be local or global)**.

Visualizations

- **Partial Dependence Plots** show the marginal effect of a feature on the predicted outcome, averaging out the effects of all other features.
- **Individual Conditional Expectation (ICE) Plots:** Local alternative to PDPs. ICE plots visualize the dependence of the prediction on a feature for individual instances, thus providing a finer-grained analysis than PDPs.

LIME, Counterfactuals, and Surrogates

- **Local Interpretable Model-agnostic Explanations (LIME):** generates an interpretable model (like a linear model or decision tree) around the prediction of an individual instance by perturbing the input data and fitting a simple model to the resultant local dataset.
- **Counterfactuals** explain model predictions by showing how slight changes to the input features could change the prediction to a desired outcome. They are "what-if" scenarios for individual predictions.
- **Global Surrogate Models:** A simpler, interpretable model is trained to approximate the predictions of the complex model. The surrogate model's behavior is then used to interpret the original model.

Model-Agnostic vs. Model-Specific

- **Model-Agnostic Methods:** These can be applied to any machine learning model (e.g., LIME, SHAP). They are flexible and can be used to compare different types of models.
- **Model-Specific Methods:** These are designed to explain a particular type of model (e.g., feature weights in linear models, decision paths in trees).

Ethical and legal considerations

- The ability to interpret a model is closely linked to stakeholders' trust in the model's decisions.
- Interpretable models do not necessarily create or enable trust – they could also enable distrust. They simply allow users to decide whether to trust them.
- Interpretability is especially important in areas with ethical implications, such as criminal justice, loan approval, or patient treatment.
- Interpretability is also driven by legal requirements, such as the General Data Protection Regulation (GDPR, an EU law) "*right to explanation*," which mandates that users have the right to be given explanations for algorithmic decisions that affect them.

References

- Himabindu Lakkaraju and Osbert Bastani. "How do I fool you?": Manipulating user trust via misleading black box explanations. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES), page 79–85, February 2020.
- Thibault Laugel, Marie-Jeanne Lesot, Christophe Marsala, Xavier Renard, and Marcin Detyniecki. The dangers of post-hoc interpretability: Unjustified counterfactual explanations. In Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19, pages 2801–2807, 7 2019.
- Molnar, C. (2020). Interpretable machine learning. Lulu. com.
- Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1:206–215, May 2019.
- Cynthia Rudin, Caroline Wang, and Beau Coker. The age of secrecy and unfairness in recidivism prediction. Harvard Data Science Review, 2(1), 2020.