An Introduction to Interpretable Machine Learning

Sael Lee
BK Associate Professor
SNU Computer Science and Engineering
Seoul National University, Korea
My Research

Machine Learning for Bio-medical Data Analysis

- **Interpretable** Method
- **Integrative** Method
- Utilize **Prior Knowledge**

Omics Data + Clinical Data

Interpretable Machine Learning

Prior Bio-clinical Knowledge

Better Healthcare
Overview

- Part 1: Introduction
  - What and why interpretable ML?
  - What is evaluated in interpretable ML?
  - How can we categorize interpretable ML?

- Part 2: Interpretable ML overview
  - Intrinsically interpretable models
  - Post-hoc interpretable methods
  - Transparent box models

- Part 3: Interpretable deep learning
Part 1: Introduction

- **What and why interpretable ML?**
- What is evaluated in interpretable ML?
- How can we categorize interpretable ML?

Most of the contents in Part 1 comes from

- CVPR’18 Tutorial by W. Samek, G. Montavon and K.R. Müller [CVPR’18 Tutorial]
What is Interpretable Machine Learning?

- “Interpretability is the degree to which a human can understand the cause of a decision.” (Miller 2017)

- “Interpretability is the degree to which a human can consistently predict the model’s result.” (Kim et al. 2016)

- “Interpretable Machine Learning refers to methods and models that make the behavior and predictions of machine learning systems understandable to humans.” (C. Molnar 2018)

- “Making a machine learning interpretable can, but does not necessarily have to, imply providing a (human-style) explanation of a prediction.”
DARPA XAI Project

Today

Training Data → Machine Learning Process → Learned Function

Task

• Why did you do that?
• Why not something else?
• When do you succeed?
• When do you fail?
• When can I trust you?
• How do I correct an error?

User

XAI

Training Data → New Machine Learning Process → Explainable Model + Explanation Interface

Task

• I understand why
• I understand why not
• I know when you succeed
• I know when you fail
• I know when to trust you
• I know why you erred

User
Explainable Models

New Approach

Create a suite of machine learning techniques that produce more explainable models, while maintaining a high level of learning performance.

Deep Explanation
- Modified deep learning techniques to learn explainable features

Interpretable Models
- Techniques to learn more structured, interpretable, causal models

Model Induction
- Techniques to infer an explainable model from any model as a black box

Lee Sael - SNU
Why Interpretability?

1. Gain trust by **verifying** that classifier works as expected.

Suggested treatment of patient A is 
because the model detected/learned
features in the patient.

Suggested treatment of patient B is
because the model detected/learned
features in the patient.
Why Interpretability?

2. **Compliance** to legislation
   - EU GDPR’s “**Right to Explanation**”
   - Retain human decision in order to **assign** responsibility!
   - Protect privacy by determining whether sensitive information in the data is revealed.
   - **Unbiased** and not discriminating against protected groups
Why Interpretability?

- 3. Gain **insights** for advance in science
  - Ex> Detecting causality, detecting significant features

[D. Choi & L. Sael (in submission)]
Why Interpretability?

4. **Improve** the learning machine by

- Finding the cause of low accuracy (**debug** and **audit**) or cause of reduction in **robustness**

![Diagram showing standard ML vs. interpretable ML]
When Interpretability is Not Needed

- Model has no significant impact
  – for fun models

- When problem is well-studied and validated.
Part 1: Introduction

- What and why interpretable ML?
- **What is evaluated in interpretable ML?**
- How can we categorize interpretable ML?
Evaluating Interpretability

- **Application level evaluation**
  - Tested by domain experts

- **Human level evaluation**
  - Tested by non-experts

- **Function level evaluation**
  - Computed

**Note:** Currently there are no real consensus on what interpretability in machine learning is and how to measure it.
Application Level Evaluation

- Explanation is included in the software product and end user (domain expert) tests it.
Human Level Evaluation

- A simplified application level evaluation where the evaluator is not a domain expert
Function Level Evaluation (FLE)

Function that has already been shown via human level evaluation to correlate with interpretability

- Example measures
  - Model sparsity
  - Conservation
  - Continuity (Monotonicity)
  - Selectivity
  - Positivity
  - Uncertainty
  - Cognitive processing time
  - Etc.
FLE 1: Model Sparsity

- Small number of features are easier to explain

Before pruning:
- Many synapses and neurons
- Complexity

After pruning:
- Reduced synapses and neurons
- Simplified structure

Deep Compression

Lee Sael - SNU
FLE2: Conservation

- Total attribution on the input features should be proportional to the amount of (explainable) evidence at the output.

Input feature $\mathbf{X}$

$\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_d$

ML

$f(\mathbf{x}) = \text{“cat”}$

Responsibility $\mathbf{R}$

$\sum_{i=1}^{d} R_i \propto f(\mathbf{x})$

Lee Sael - SNU
FLE 3: Continuity (Monotonicity)

- If two inputs are similar and the prediction is similar, then the explanation should be similar.
FLE 4: Selectivity

- If input features are relevant to output decision, removing them should reduce evidence at the output.

\[ f(x) = \max(x_1, x_2) \]

Method 1
- \( x_2 > x_1 \) remove \( x_1 \)

Method 2

[Bach’15, Samek’17]

[CVPR’18 Tutorial]
FLE 5: Positivity

- If the model is certain about its prediction, input features are either relevant (positive) or irrelevant (zero).

Nonnegative Matrix Factorization (NMF) of facial images [Lee and Seung (1999)]
FLE 6: Uncertainty

- How uncertain the model is about the output.
FLE 7: Cognitive Processing Time

- How long does it take to understand the explanation?
e.g., Functional evaluation of intrinsically interpretable models

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Linear</th>
<th>Monotone</th>
<th>Interaction</th>
<th>Task</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear models</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Regr.</td>
</tr>
<tr>
<td>Logistic regression</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Class.</td>
</tr>
<tr>
<td>Decision trees</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>Class. + Regr.</td>
</tr>
<tr>
<td>RuleFit</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>Class. + Regr.</td>
</tr>
<tr>
<td>Naive Bayes</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>Class.n</td>
</tr>
<tr>
<td>k-nearest neighbours</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>Class. + Regr.</td>
</tr>
</tbody>
</table>

Lee Sael - SNU
**e.g., Functional Evaluation of DNN**

<table>
<thead>
<tr>
<th>Explanation techniques</th>
<th>Uniform</th>
<th>(Gradient)$^2$</th>
<th>(Guided BP)$^2$</th>
<th>Gradient x Input</th>
<th>Guided BP x Input</th>
<th>LRP$_{α, β_0}$</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Properties</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1. Conservation</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>2. Positivity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>3. Continuity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>4. Selectivity</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Lee Sael - SNU
Part 1: Introduction

- What and why interpretable ML?
- What is evaluated in interpretable ML?
- How can we categorize interpretable ML?
Criteria for Interpretability 1

Intrinsic vs post hoc

- **Intrinsically** interpretable models
  - Decision trees
  - Linear models
  - Logistic regression
  - Decision rules
  - RuleFit
  - Other interpretability integrated machine learning

- **Post hoc** interpretable models
  - Selecting and training a black box model and applying interpretability methods after the training
Criteria for Interpretability 2

Model-specific vs model-agnostic

- Model-specific interpretation tools
  - Limited to specific model classes
  - E.g. interpretation of regression weights in a linear model
  - *Interpreting neural networks is also model-specific

- Model-agnostic tools
  - Independent of machine learning model and are usually post hoc
  - Assumes that model internals, e.g., weights or structural information, are not accessible

[C. Molnar 2018]
Note: Model-Agnostic Methods (MAM)

Pros of MAM [Ribeiro et al. 2016]

- **Model flexibility**
  - Not tied to any models

- **Explanation flexibility**
  - Not tied to a form of explanation

- **Representation flexibility**
  - Need not use the same feature representation as the model being explained

We will focus more on the Agnostic Methods

Figure from [C. Molnar, 2018]
### Criteria for Interpretability 3

#### focus on model (global)  
1. Interpreting Learned Models

#### focus on data (local)  
2. Explaining Decisions

**Model**

“What **pattern** describes a **class** according to the model?”

“What **prototypes** does the model learn?”

“How do **parts of the model** influence predictions?”

**Prediction**

“Explain **why** a instance (set of instances) has been classify in a certain way \( f(x) \)?”

**Data**

“Which part of the instance are **relevant** or **predictive** of the task?”
From Model Analysis to Decision Analysis

Discriminative models
- \( p(C|X) \)

Generative models
- \( p(X,C) = p(C)p(X|C) \)
- \( p(C|X) = p(C)p(X|C)/p(X) \)

Lee Sael - SNU
Criteria for Interpretability 4

Outcome of the interpretable methods

- **Intrinsically interpretable model**
  - *Approximate* the black box model via interpretable model
  - E.g. model compression, global surrogate methods

- **Model internals**
  - *Applicable for intrinsically interpretable models*
  - E.g. learned weights in linear model

- **Data points**
  - E.g. identification of *prototypes* of predicted classes

- **Features**
  - *Feature* summary statistics or visualization
Reference