Explainable Machine Learning

Interpretable models

Bojan Mihaljević

ETSIINF
Universidad Politécnica de Madrid
b.mihaljevic@upm.es

Table of Contents

1. Introduction
2. Interpretability
3. Predictive power

Black-box

Roughly, a function that is too complicated for any human to comprehend.

Examples

- Random forest;
- Boosting;
- Neural networks;
- Nonlinear SVMs;
- . . .

Explainable AI

Explaining a black box model, via: an approximation, variable importance measures, . . .

Problem: How faithful?

- If the explanation was completely faithful to the original model, the explanation would equal the original model, and one would not need the black-box model in the first place.
- Can provide misleading or false characterizations.
- May add unjustified authority to the black box.

Inherently interpretable model

- Not a black box.
- Provide their own explanations, which are faithful to what the model actually computes.
- Alternative terms: transparent, glass-box, intelligible, explainable.

"Enhance human decision making, while black box AI replaces it."

Example

Pneumonia risk

- Large project, in the mid 90's, to predict the probability of death for patients with pneumonia so that high-risk patients could be admitted to the hospital.
- A goal was to perform a clinical trial to determine if machine learning could be used to predict risk prior to hospitalization and inform the decision about hospitalization.
- Neural nets clearly outperformed more traditional methods (AUC=0.86 versus 0.77 for logistic regression).
- After careful consideration they were considered **too risky for use on real patients and logistic regression was used instead**.

Example

Pneumonia risk

- A rule-based system, less accurate that the neural net, learned the rule "HasAsthma(x) ⇒ LowerRisk(x)": patients with pneumonia who have a history of asthma have lower risk of dying from pneumonia than the general population.
- A true pattern in the data: patients with a history of asthma were usually admitted directly to the ICU, and the aggressive care was so effective that it lowered their risk of dying from pneumonia compared to the general population.
- Because the prognosis for these patients is better than average, models trained on the data incorrectly learn that asthma lowers risk, when in fact asthmatics have much higher risk (if not hospitalized).

Example

Pneumonia risk

- If the rule-based system had learned that asthma lowers risk, certainly the neural nets had learned it, too.
- The rule-based system was intelligible and modular, making it easy to recognize and remove dangerous rules like the asthma one.
- Decision: **not use the neural nets**. Not because the asthma problem could not be solved, but because the **lack of intelligibility made it difficult to know what other problems might also need fixing**.
- For example, perhaps pregnant women with pneumonia also receive aggressive treatment that lowers their risk compared to the general population. The neural net might learn that pregnancy lowers risk, and thus recommend not admitting pregnant women, thus putting them at increased risk.

Confounding

- Example: a neural network turned out to be picking up on the word "portable" within an x-ray image, representing the type of x-ray equipment rather than the medical content of the image.
- With an interpretable model, this issue would never have gone unnoticed.
- Many other examples: cow on the beach, etc.
- "Clever Hans" effect: The classifier works "well", but is arriving at conclusions looking at the wrong features.
- If the training data is not representable of the population (e.g., not iid), we could be using it in cases that are very different from what it was trained on.

Black-box versus interpretable models

Black-box disadvatanges

- Often not easy to combine with expert knowledge.
- Harder to troubleshoot in real-time.
- When explaining them, one expects to lose accuracy.

High-stakes decisions

- Healthcare, justice system, credit scoring, self-driving cars, . . .
- XAI can be dangerous to a much higher degree.

Table of Contents

1 Introduction

2 Interpretability

3 Predictive power

Interpretability

Some characteristics

- Simulatability (a human is able to contemplate and reason about the entire decision-making process at once);
- Sparsity;
- Decomposability (modularity);
- Algorithmic transparency;
- Domain-specifics: e.g., causality, monotonicity, additivity.

Interpretability

Model families

- Logical: trees, rule lists/sets, ...
- Linear (logistic/linear regression, linear discriminant analysis)
- Additive: generalized aditive models, naive Bayes, ...
- k-nearest neighbors
- Other: Domain-specific, look up tables, ...

Table of Contents

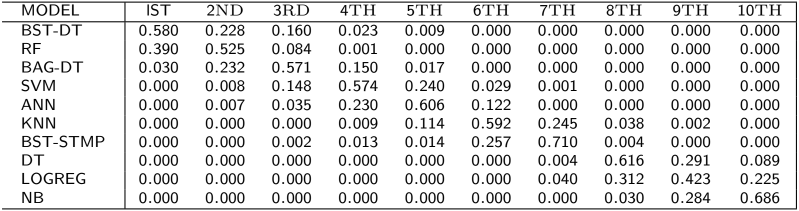
1 Introduction

2 Interpretability

3 Predictive power

Empirical comparisons

Example



MODEL	IST	2ND	3RD	4TH	5TH	6TH	7TH	8TH	9TH	10TH
BST-DT	0.580	0.228	0.160	0.023	0.009	0.000	0.000	0.000	0.000	0.000
RF	0.390	0.525	0.084	0.001	0.000	0.000	0.000	0.000	0.000	0.000
BAG-DT	0.030	0.232	0.571	0.150	0.017	0.000	0.000	0.000	0.000	0.000
SVM	0.000	0.008	0.148	0.574	0.240	0.029	0.001	0.000	0.000	0.000
ANN	0.000	0.007	0.035	0.230	0.606	0.122	0.000	0.000	0.000	0.000
KNN	0.000	0.000	0.000	0.009	0.114	0.592	0.245	0.038	0.002	0.000
BST-STMP	0.000	0.000	0.002	0.013	0.014	0.257	0.710	0.004	0.000	0.000
DT	0.000	0.000	0.000	0.000	0.000	0.000	0.004	0.616	0.291	0.089
LOGREG	0.000	0.000	0.000	0.000	0.000	0.000	0.040	0.312	0.423	0.225
NB	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.030	0.284	0.686

Empirical comparisons

- The results are somewhat typical: black-box models on top, interpretable ones below.
- Results like this may be the basis of the belief that there is necessarily a trade-off between predictive power and interpretability.
- However, this is not necessarily so.
- For example, it is unclear how to fairly compare methods that are sensitive to preprocessing, such as naive Bayes, with, for example, tree-based methods which barely need preprocessing.
- Tree-based ensembles on top: barely require preprocessing, while reduce bias/variance. Excellent off-the-shelf models.

Simple models in practice

- Often, simple models can account for a large fraction (e.g., over 90%) of the predictive power of “the best” model.
- In some cases, this laboratory difference might not be realized in practice, due to uncertainties arising from training/serving skew.

Practical issues

- Training data might not be representative of the population (not iid).
- Concept drift: changes in the relation between the features and/or in the definition of the class.
- Uncertainty about the class labels.
- Data leakage: the doctors’ notes may reveal the patients’ outcome before it is officially recorded.

Interpretability is useful for troubleshooting, which can lead to better accuracy.

Predictive power

- Often, it is much easier to train a black box model; for data that are unconfounded, complete, clean, and iid, this works well.
- Example: image recognition.
- When data is noisy, relatively simple white-box methods tend to be effective.
- There is no conclusive evidence for a general tradeoff between accuracy and interpretability when one considers the full data science process, and iteratively refining the model.

Applicability

Rules of thumb (Rudin)

Models	Data type
Decision trees / decision lists (rule lists) / decision sets	Somewhat clean tabular data with interactions including multiclass problems. Particularly useful for categorical data with complex interactions (i.e., more than pairwise).
Scoring systems	Somewhat clean tabular data, typically used in medicine and criminal justice because they are small enough that they can be memorized by humans.
Generalized additive models (GAMs)	Continuous data with at most quadratic interactions, useful for large-scale medical record data.
Case-based reasoning	Any data type, (different methods exist for different data types), including multiclass problems.

Summary

- Low-stakes decisions may not require interpretable models.
- However, for high-stakes decisions, much safer to use interpretable models, rather than "explained" black box models.
- Often, one aims for explaining a black box without considering whether there is an interpretable model of the same accuracy.
- A globally interpretable model was among the winners of the FICO challenge Recognition Prize for the competition: http://dukedatasciencefico.cs.duke.edu/.