A (Very) Short Introduction to Interpretable Machine Learning

Mohamed Siala
homepages.laas.fr/msiala

INSA-Toulouse & LAAS-CNRS

December 20, 2021

Context

- The purpose is to introduce the notion of interpretable machine learning and why it matters
- This short course is for non CS students. I will present the algorithms in a high level way without technical details.
- Many notions are introduced in a non-formal way
- Recent personnel interest in ethical AI
- There is a practical session next week
- Three parts: Why? How ? What's next?

WHY ?

The COMPAS Tool

MIT Technology Review

Featured Topics Newsletters Events Podcasts Sign in Subscribe

TECH POLICY

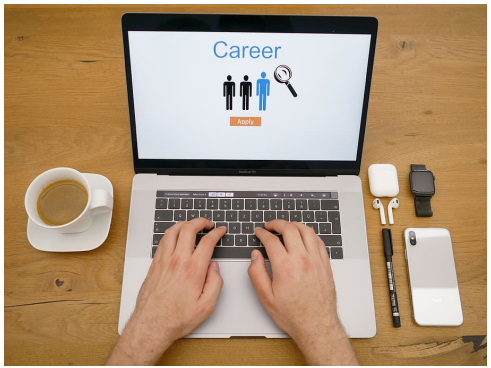
AI is sending people to jail—and getting it wrong

Using historical data to train risk assessment tools could mean that machines are copying the mistakes of the past.

By Karen Hao January 21, 2019

IAN WALDIE/GETTY IMAGES

Increasing Number of Real Life and Social AI Applications



AI: Increasing Number of Real Life Applications Of Machine Learning

- The diverse applications of AI raised many ethical issues and questions
  - Job applications: AI that parses CVs for software engineers and recommends to hire mostly men
  - Credit scoring: AI that gives a credit score (for bank loans and credit applications) that recommends people from a particular geographical region, specific gender, social class, etc
  - Compass tool: (2016) used by judges in the US to predict which criminals are likely to re-offend is found to be biased by the skin color ("race" African-American/Caucasian).

COMPASS data and Rule-based Predictions

Sex	Age	Priors	Juvenile Felonies	Juvenile Crimes	Race
Male	15	1	0	1	Caucasian
Male	15	1	0	1	African-American
Female	33	1	0	1	African-American
Female	27	0	1	0	Caucasian
Male	41	0	1	0	Caucasian
...	...	...	...	...	...

The problem is to predict recidivism. That is, the tendency of a convicted criminal to re-offend.

A Step Back : Computationally Speaking, what is a Prediction Problem?

- **Input:** Some (historical) data coming from an unknown distribution
- **Question (Hard Version)** Predict the underling function
- **Question (Soft Version)** Predict a function belonging to certain family (neural networks, svm, decision tree, etc) that minimises the prediction error

A Step Back : Machine Learning

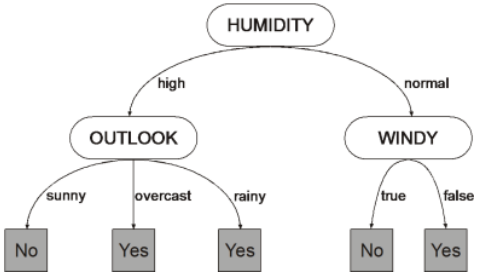
- Supervised ML: data is labelled (associated to a value). For instance:
  - Time expectancy of a device
  - Prediction of device periodic maintenance
  - Weather forecast
- Unsupervised ML: data is unlabelled. For instance:
  - Trace COVID density in a geographical zone
  - Clustering of population to build transportation

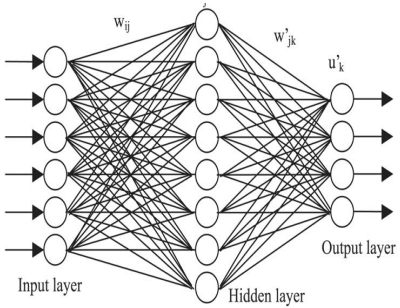
Interpretable vs. Black-box Models

Supervised Machine Learning

- Regression : The outcome is a continuous value (cost, volume, space, salary, etc)
- Classification: The outcome is a discrete value ('class') : Blue/red/green, True/False, Cancer/No Cancer, No COVID/Original COVID/Omicron/Delta/Alpha etc

Learning Models : Black-Box vs Interpretable Models





Definitions [Rudin et al., 2021]

- **Black-box model** : A formula that is either too complicated for any human to understand, or proprietary, so that one cannot understand its inner workings
- **Interpretable 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.

Why Interpretable Models?

- Transparent
- Trustworthy
- Inherently Explainable
- Well adapted for troubleshooting and diagnosis
- **Mandatory criteria in high-stake decision making**

HOW?

Different Models Depending on the Data At Hand

The tip of the iceberg

- **Tabular data** : Decision trees, decision lists, decision rules, scoring systems, ...
- **Raw inputs** : Disentangled neural networks ...
- **Continuous data** : Generalised additive models, ...
- Many others: case-based reasoning

Restrictions in this course

- Here, we consider only tabular data and models with decision trees/sets/rules

Decision rules & Decision Sets

- They are defined as If-Condition-Then-Prediction rules
- **Decision sets**: no specific order is given between the rules. Ties are broken by majority votes
- **Decision rules**: rules are ordered by priority

Example of Rule List found by FairCORELS

- Data : `https://www.kaggle.com/danofer/compass`
- FairCORELS: `https://github.com/ferryjul/fairCORELS`

```
if [priors:>3] then [recidivism]
else if [age:21-22 && gender:Male] then [recidivism]
else if [age:18-20] then [recidivism]
else if [age:23-25 && priors:2-3] then [recidivism]
else [no recidivism]
```

Rule list 5. Example of an unconstrained rule list found by `FairCORELS` on COMPAS dataset, with Accuracy $= 0.681$, $\text{UNF}_{\text{EOdds}} = 0.217$ and $\text{UNF}_{\text{CUAE}} = 0.046$

Decision Trees

Decision trees are rooted directed graph where each internal node corresponds to a feature; every child corresponds to a specific value of it's parent feature. Leaf nodes are the classes/values of the prediction model.

How to construct A Decision Tree?

Different Criteria

- Depth
- Number of nodes (questions?)
- Minimum number of support per leaf node
- Sparsity
- But also Fairness (avoid explicit sensitive features) and privacy, etc
- ...

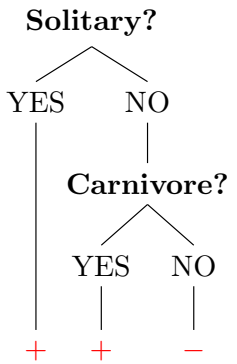
How to Construct a Decision Tree?

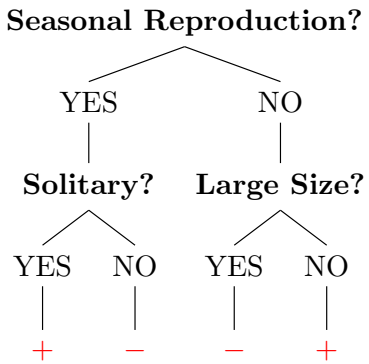
Different approaches

- Greedy Algorithms (CART, C4.5, etc)
- Exact Algorithms :
  1. Branch and bound algorithms
  2. Mixed Integer Programming
  3. Constraint programming and Boolean Satisfiability
  4. Dynamic Programming

Toy Example: DTs to Predict The Likelihood of Animal Extinction

Big Size	Carnivore	Seasonal Reproduction	Solitary	Extinct
0	1	0	1	yes
1	0	0	1	yes
0	0	0	1	no
1	1	1	0	no
0	0	1	0	yes





- Find A decision tree with 100% accuracy
- Find the best decision tree with depth 1
- Find the best decision tree with depth 2
- Find the best decision tree with depth 3
- Find the best decision tree with depth 3 using the minimum number of nodes
- Why the number of features (i.e. nodes) is important?

Decision Trees

- Inherently interpretable
- Tree structured model
- Simple yet powerful model
- Yes, it can be used for classification AND regression
- Can be used with multi-valued/binary input/output

Algorithm 1 GREEDY (with binary features)

Require: $data$, $Nodes$,$Height$

$Tree \leftarrow \emptyset$;

$stop \leftarrow check\_stop(data, Nodes, Height)$;

if $!stop$ **then**

$f \leftarrow section(data)$

$dataL \leftarrow data$ such that f ;

$dataR \leftarrow data$ such that $!f$;

$LeftTree \leftarrow GREEDY(dataL, Nodes, Height)$;

$RightTree \leftarrow GREEDY(dataR, Nodes, Height)$;

$Nodes \leftarrow Nodes + 2$

$Height \leftarrow Height + 1$

$Tree \leftarrow (f, LeftTree, RightTree)$

end if;

return Tree ;

How to Select?

- Best feature is the one that gives data with one label on each side
- if not possible picks the one that tends to do so
- Different criteria : Information Gain (ID3), Gain Ration (C4.5), Gini Index (CART), etc

CART Algorithm: Classification And Regression Trees

Selection based on Gini Index

- Gini index: Proposed by Corrado Gini in statistics to estimate inequality among a society group/nation
- Values Between $[0, 1]$
  - 0 corresponds to perfect equality (same presence of all possible outcome among the distribution)
  - 1 corresponds to perfect inequality (dominance of a specific outcome among the distribution)
- CART select features that tend to favour inequalities (best case to have one class per branch)
- Can be computed efficiently
- Very fast, however, no guarantee of optimality

What is Next?

Interpretability vs. Explainability

- Iterpretability and explainability are similar but not exactly the same
- Iterpretability is a property of the model itself
- Explainability is a post-prediction step
- Interpretable models are explainable by default
- One can have a black-box model associated to an explanation algorithm

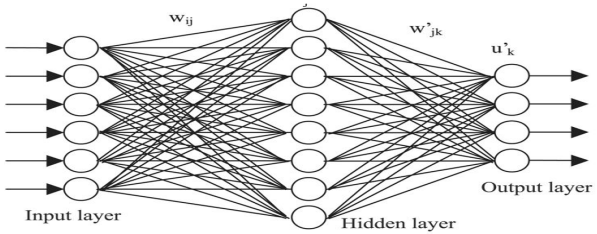
Explainability

- The GDPR (General Data Protection Regulation) explicitly mentions that, in the context of algorithmic decision-making, every user has the right to **explanation**.
  `Link:https://gdpr-info.eu/`
- First question: What does it mean to "explain"?
- Second question: What if we cannot "explain" (black-box models)?

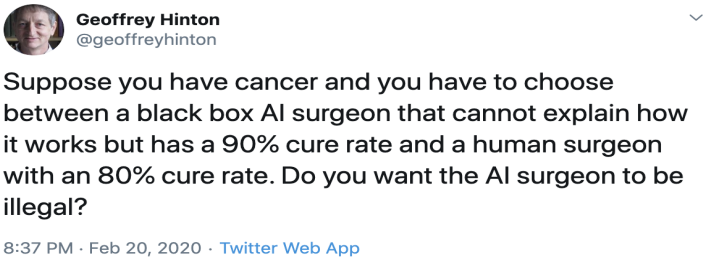
What does it mean to "explain"?

- It depends on the context : The nature of explanation can differ from application to application
- Explain to who? See an interview of Richard Feynman on the why question `https://www.youtube.com/watch?v=Q1lL-hXO27Q&feature=emb_title`

What if we cannot "explain"?



- Deep learning models: Huge success in many applications: image recognition problems, autonomous cars, heath-care systems, etc
- Hard to explain the predictions (no obvious way to figure out with part of the input influenced the prediction)



Do we really need to explain?

FairWashing [Aïvodji et al., 2019]

- Imagine situations where it is possible to find multiple explanations for the same action. Which one to choose?
- What if some explanations are "more fair" than others?
- A company proposing an AI decision making service can always present the fairest explanations to the user..
- **Interpretable Models are well encouraged to avoid such a problem.**

In practice with existing tools

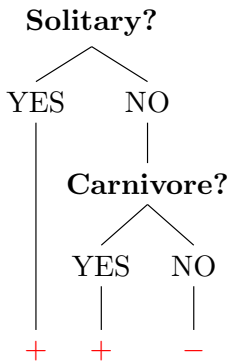
- The general procedure: Build a model, evaluate the model (eventually re-build), deploy the model
- How to evaluate the model given that the real test will happen once the model is deployed ?
- We can estimate the behaviour of the model on a subset of the data
- Sacrifice a part of the data for test

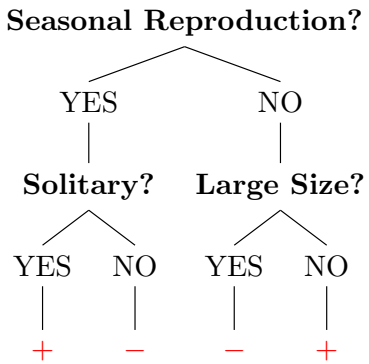
How it works in practice

- Split the data into train data (used to build the model) and a test data (to evaluate the model). For instance 80% for train and 20% for test
- Build the model
- Prediction error on train/test
- Evaluate the Generalisation, the confusion matrix (Ratio of True/False positive/negative)
- Repeat this process many times by randomly selecting the train/test data (k-fold evaluation, random splits)

Toy Example: DTs to Predict The Likelihood of Animal Extinction

Big Size	Carnivore	Seasonal Reproduction	Solitary	Extinct
0	1	0	1	yes
1	0	0	1	yes
0	0	0	1	no
1	1	1	0	no
0	0	1	0	yes





- Find A decision tree with 100% accuracy
- Find the best decision tree with depth 1
- Find the best decision tree with depth 2
- Find the best decision tree with depth 3
- Find the best decision tree with depth 3 using minimum number of nodes
- Why the number of features (i.e. nodes) is important?

Things to consider

- Equivalence between trees
- Useless splits
- Redundant splits
- How about symmetric/isomorphic/identical subtrees?
  - Two children of a parent are identical $--> $ remove the parent
  - Eventually consider other graphical models such as Binary Decision Diagram
- Compact representation
- Trade-off Sparsity vs. Depth
- Redundant subtrees (bad for memory consumption and CPU time)

References

- Interpretable Machine Learning A Guide for Making Black Box Models Explainable.
  `https://christophm.github.io/interpretable-ml-book/`
- Interpretable Machine Learning: Fundamental Principles and 10 Grand Challenges Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, Chudi Zhong
  `https://arxiv.org/abs/2103.11251`
- Anything from Cynthia Rudin:
  1. Homepage `https://users.cs.duke.edu/~cynthia`
  2. Youtube
     `https://www.youtube.com/channel/UCFAkntpj2aOBJ1q4o4FhVwQ`

References

Aïvodji, U., Arai, H., Fortineau, O., Gambs, S., Hara, S., and Tapp, A. (2019).
Fairwashing: the risk of rationalization.
arXiv preprint arXiv:1901.09749.

Rudin, C., Chen, C., Chen, Z., Huang, H., Semenova, L., and Zhong, C. (2021).
Interpretable machine learning: Fundamental principles and 10 grand challenges.
CoRR, abs/2103.11251.