Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead

Cynthia Rudin, Duke University

Presenters : Sreya Dutta Roy, Ziqian Lin

Overview

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

- ❖ Introduction
- ❖ Explainable ML Vs Interpretable ML
- ❖ Explainable ML Issues
- ❖ Interpretable ML Issues
- ❖ Encouraging Responsible ML Governance: Two proposal
- ❖ Algorithmic Challenges in Interpretable ML: Three challenges
- ❖ Assumption of Interpretable Models Might Exist
- ❖ Advantage of Lacking Algorithm Stability
- ❖ Conclusion and Questions

Introduction

Department of Computer Sciences
UNIVERSITY OF WISCONSIN-MADISON

- Black-box ML Models are being deployed in High-stakes decision Making

Some examples of High Stakes domains :

- *Criminal Justice*
- *Healthcare*
- *Energy Reliability*
- *Financial Risk Assessment*

NEED FOR INTERPRETABILITY !!

Types of Black Box Models

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

Black Box Models

Tough for Humans to Comprehend

Proprietary (Eg. COMPAS)

Some are Both !

Explainable ML Vs Interpretable ML

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

Explainable ML : Problems ?!?

❖ Post-hoc Model to explain first Blackbox model

Interpretable ML : Challenges ?!?

❖ Inherently Interpretable, provides own explanations !

Especially needed for High Stakes domains and cases where Troubleshooting is important

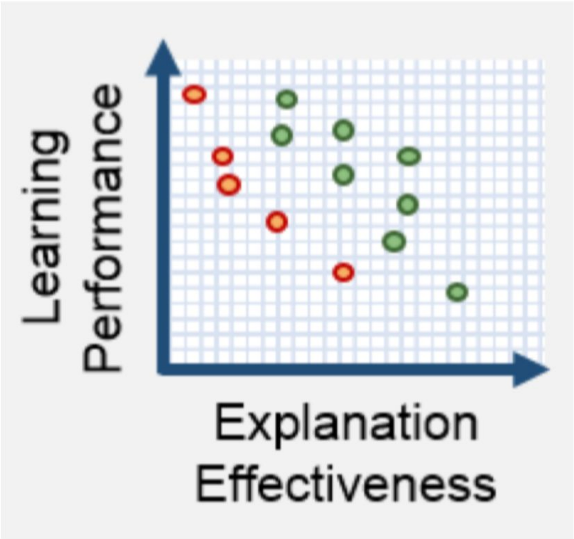
Explainable ML Issues

Department of Computer Sciences
UNIVERSITY OF WISCONSIN-MADISON

❖ <u>Common Myth of Trade-off between Accuracy and Interpretability</u>

Role of Data ?

- Structured Data an ally to Interpretability
- Repeated Iterations in Processing Data Leads to a more Accurate Model



DARPA XAI (Explainable AI) Board Agency Announcements

Is this Meaningful , Fair, Represenattive ?

- Using some Static Data?
- Comparing 1984 CART to 2018 Deep Learning Models ?

❖ Explainable ML Faithfulness to Original Model Computations

Why Explain ?

To Trust The Black Box Model

But

Explanation Model $\neq$ Original Model

Notion of **Distrust** on the Black Box Model due to Incorrect Explanation

Consider the case of Criminal Recidivism

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

ProPublica Analysis :

- Accused COMPAS of racism
- Showed Linear Dependency of Criminal Recidivism decision conditioned on Race

COMPAS :
Proprietary model that is used widely in the U.S. Justice system for parole and bail decisions

Explanation of COMPAS :

"This person is predicted to be arrested because they are black."

IS it correct to call it an explanation ?

- Features might not be same as in original COMPAS
- Primary **Features** in Criminal Recidivism Decisions are **Age, Criminal History** which could have **correlation with Race**
- **COMPAS is actually a nonlinear model**
- **Wouldn't bias / unbias be clearer if this was an Interpretable Model ?**

❖ Do Explanations always Make Sense ?

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

Suppose :

- Original Model Predicted correctly
- Explanation Model Approximated Predictions of Black Box Correctly

What about explanation's **Informativeness** or Enoughness to Make Sense ?

Consider Saliency Maps (for Low Stakes problems) :

	Test Image	Evidence for Animal Being a Siberian Husky	Evidence for Animal Being a Transverse Flute
Explanations Using Attention Maps			

❖ Black Box Compatibility with new Information based Decision Revision

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

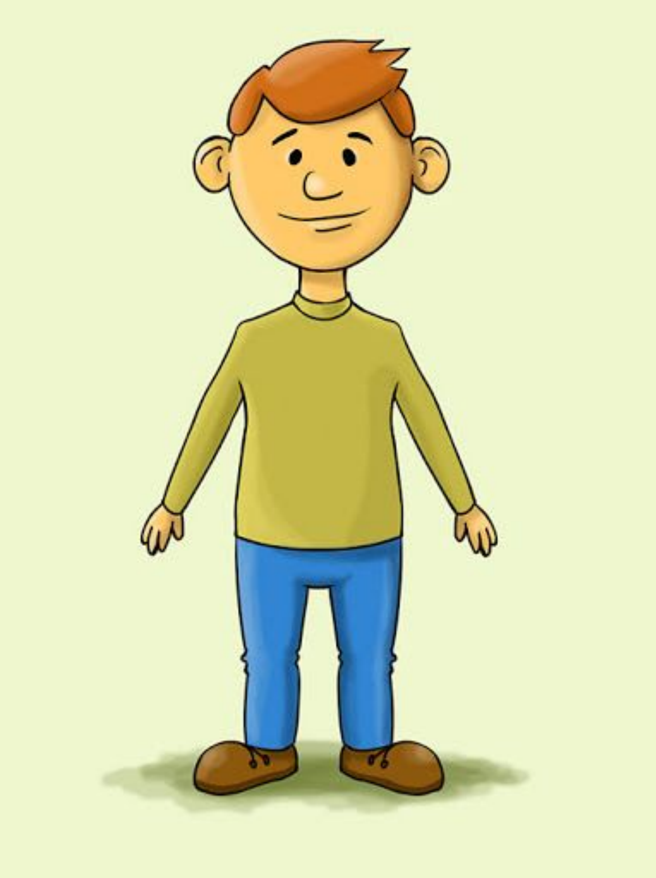
- An Interpretable model could clearly show the reasons for decision

- So if the new information received by say, a Judge was not factored, it could be easily included

- However with Black-Box Models, this could be fairly tricky.

Eg. Factoring in Seriousness of Crime in the Compas Decision.

To introduce the next issue Let’s meet Tim and Harry !!



- They have same age and similar criminal history
- However one is denied bail and one isn’t

WHY?!?!

❖ Overly Complicated Decision Pathway ripe to Human Error

Department of Computer Sciences
UNIVERSITY OF WISCONSIN-MADISON

COMPAS depends on ~130+ factors and Human Surveys

Human Surveys have High Chances of Typographical Errors

These Errors sometimes lead to random Parole / Bail Decisions

- PROCEDURAL UNFAIRNESS !!
- Troubleshooting Nightmare

Why Advocate for Extra Explainable Model and Not Interpretable Models ?

❖ Profit Afforded to Black Box Intellectual Property

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

But would one pay for such a simple if-else model ?

CORELS (Certifiably Optimal Rule Lists) :

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

COMPAS Accuracy ⟺ CORELS Accuracy

Qualitative Differences

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

COMPAS	CORELS
black box	full model is in Figure 3
130+ factors	only age, priors, (optional) gender
might include socio-economic info	no other information
expensive (software license),	free, transparent
within software used in U.S. Justice System	

Environmental & Health

BreezoMeter, used by Google during the California wildfires of 2018, which predicted air quality as "good – ideal air quality for outdoor activities,"

- ***Confounding Issues*** *haunt Datasets (Mainly Medical)*
- *Leading to* ***Fragile Models*** *with serious errors, even with change of an xray equipment.*
- ***Interpretable Models would have helped in early detections***

Medical Datasets, Automations

Zech et al. noticed that their neural network was 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.

Notice : CONFLICT OF INTEREST :

"The companies that profit from these models are not necessarily responsible for the quality of individual predictions "

They are not directly affected if an applicant is denied loan or if a prisoner stays in prison for long due to their mistake

Some “Debatable” Arguments in Favour of Black Box Models:

- Keeping Models as Black Boxes / Hidden helps prevent them from being gamed or Reverse-Engineered

Is Reverse Engineering always bad ?
Building a higher credit score => more creditworthiness

- Belief that “counterfactual explanations” are sufficient (Minimal Change in input to get opposite Result)

Eg. Save $1000 more to get loan or
Get a new job with $1000 more salary to get loan

“*Minimal*” depends on circumstances / individual.

★ Black boxes are bad at factoring in new information

High Efforts to Construct Interpretable Models

- Need for more Domain Expertise : Definition for Interpretability for the Domain

- Interpretability Constraints (like Sparsity) -> Computationally hard Optimization Problems in worst case

Might be worthwhile in high stakes problems to invest here

❖ Black box Seem to uncover “hidden patterns”

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

- Black boxes are seen to uncover hidden patterns the user was unaware of

- If the pattern was important enough for the Blackbox to leverage it for predictions, an interpretable model might also locate and use it

- Depends on Researcher’s ability to construct accurate-yet-interpretable models

Encouraging Responsible ML Governance: Two Proposals

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

regulation

European Union’s revolutionary General Data Protection Regulation and other AI regulation

× an interpretable model

√ an explanation

it is not clear whether the explanation is required to be accurate, complete, or faithful to the underlying model

Two Proposal

Encouraging Responsible ML Governance: Two Proposals

(1) For certain high-stakes decisions, no black box should be deployed when there exists an interpretable model with the same level of performance.(stressful)

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

COMPAS	CORELS
black box 130+ factors might include socio-economic info expensive (software license), within software used in U.S. Justice System	full model is in Figure 3 only age, priors, (optional) gender no other information free, transparent

Opacity is viewed as essential in protecting intellectual property, so it’s still a long way.

Encouraging Responsible ML Governance: Two Proposals

(2) Let us consider the possibility that organizations that introduce black box models would be mandated to report the accuracy of interpretable modeling methods. (less stressful)

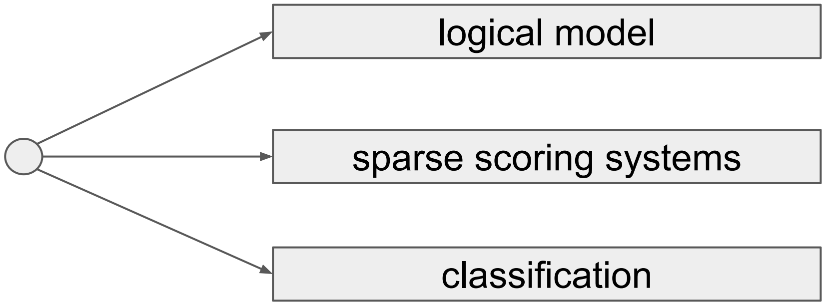
accuracy ⟸ trade off ⟹ **interpretability**

× solve all problems

√ rule out companies selling recidivism prediction models, possibly credit scoring models, and other kinds of models where we can construct accurate yet-interpretable alternatives.

Algorithmic Challenges in Interpretable ML: Three cases

interpretability is domain-specific => a large toolbox => design’s skills

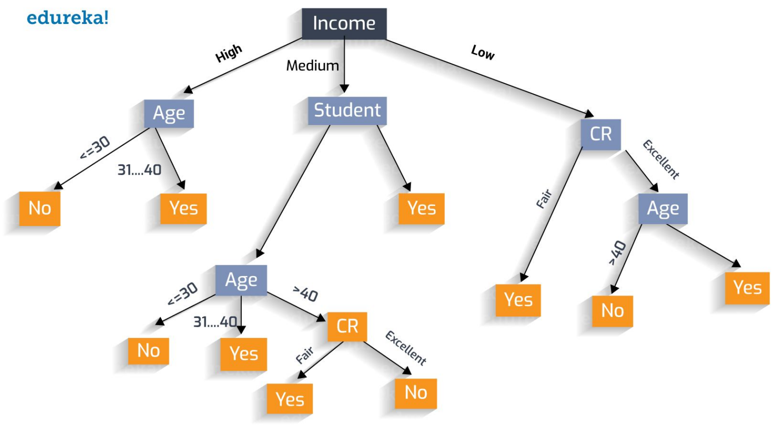


three cases’ common => human-designed models by ML

Algorithmic Challenges in Interpretable ML: (1) logical models

Definition: A logical model consists of statements involving “or,” “and,” “if-then,” etc.

Example: Decision trees



Training observations are indexed from i = 1, .., n;
F is a family of logical models such as decision trees.
The optimization problem is:

$$\min_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{[\text{training observation } i \text{ is misclassified by } f]} + \lambda \times \text{size}(f) \right)$$

Algorithmic Challenges in Interpretable ML: (1) logical models

$$\min_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} 1_{[\text{training observation } i \text{ is misclassified by } f]} + \lambda \times \text{size}(f) \right)$$

the **size** of the model can be measured by the **number of logical conditions** in the model

IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

computationally hard
The challenge is whether we can solve (or approximately solve) problems like this in practical ways by leveraging new theoretical techniques and advances in hardware.

Algorithmic Challenges in Interpretable ML: (1) logical models

Department of Computer Sciences
UNIVERSITY OF WISCONSIN-MADISON

$$\boxed{\text{CORELS}} \xrightarrow{?} \boxed{\text{all possible models}}$$

(i) a set of theorems allowing massive reductions in the search space of rule lists;

(ii) a custom fast bit-vector library that allows fast exploration of the search space;

(iii) specialized data structures that keep track of intermediate computations and symmetries.

https://www.jmlr.org/papers/volume18/17-716/17-716.pdf

Challenges in Interpretable ML: (2) sparse scoring systems

Department of Computer Sciences
UNIVERSITY OF WISCONSIN-MADISON

Definition: A scoring system is a sparse linear model with integer coefficients – the coefficients are the point scores.

Example: a scoring system for criminal recidivism:

1.	Prior Arrests ≥ 2	1 point		$\cdots$
2.	Prior Arrests ≥ 5	1 point	+	$\cdots$
3.	Prior Arrests for Local Ordinance	1 point	+	$\cdots$
4.	Age at Release between 18 to 24	1 point	+	$\cdots$
5.	Age at Release ≥ 40	-1 points	+	$\cdots$
		SCORE	=	$\cdots$

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

Challenges in Interpretable ML: (2) sparse scoring systems

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

1.	Prior Arrests ≥ 2	1 point		$\cdots$
2.	Prior Arrests ≥ 5	1 point	+	$\cdots$
3.	Prior Arrests for Local Ordinance	1 point	+	$\cdots$
4.	Age at Release between 18 to 24	1 point	+	$\cdots$
5.	Age at Release ≥ 40	-1 points	+	$\cdots$
		SCORE	=	$\cdots$

SCORE	-1	0	1	2	3	4
RISK	11.9%	26.9%	50.0%	73.1%	88.1%	95.3%

The problem is hard
mixed-integer-nonlinear program (MINLP)

the second challenge is to create algorithms for scoring systems that are computationally efficient

$$\min_{b_1,b_2,..,b_p \in \{-10,-9,...,9,10\}} \frac{1}{n}\sum_{i=1}^{n} \log\left(1+\exp\left(-\sum_{j=1}^{p} b_j x_{i,j}\right)\right) + \lambda \sum_j 1_{[b_j \neq 0]},$$

The first term is the logistic loss used in logistic regression (sigmoid)

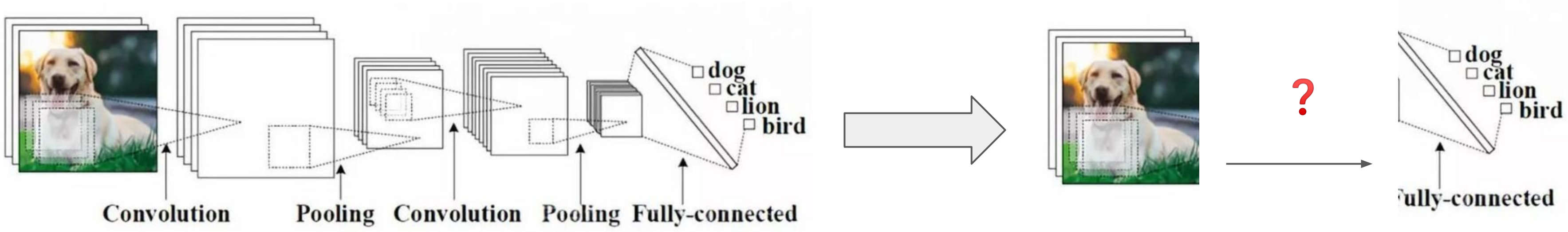
RiskSLIM (Risk-Supersparse-Linear-Integer-Models)

Department of Computer Sciences
UNIVERSITY OF WISCONSIN-MADISON

Challenges in Interpretable ML: (3) Classification

Even for classic domains of machine learning, where latent representations of data need to be constructed, there could exist interpretable models that are as accurate as black box models. Using classification as example:

The network must then make decisions by **reasoning about parts of the image** so that the explanations are real, and **not posthoc**.



Challenges in Interpretable ML: (3) Classification

a special prototype layer to the end of the network by Chaofan Chen
https://arxiv.org/pdf/1806.10574.pdf

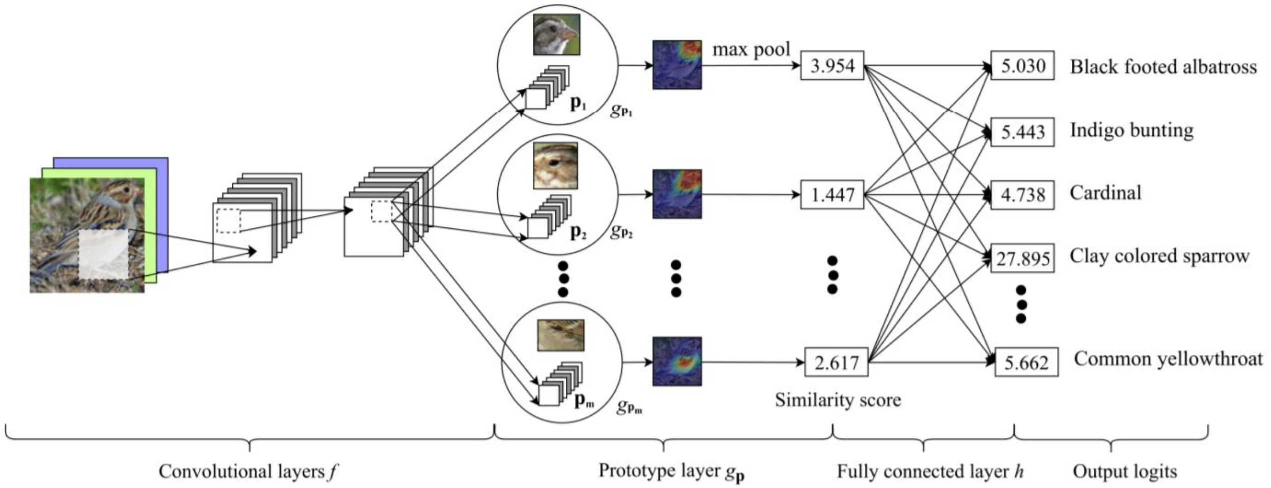


Table 1: Top: Accuracy comparison on cropped bird images of CUB-200-2011
Bottom: Comparison of our model with other deep models

Base	ProtoPNet	Baseline	Base	ProtoPNet	Baseline
VGG16	76.1 ± 0.2	74.6 ± 0.2	VGG19	78.0 ± 0.2	75.1 ± 0.4
Res34	79.2 ± 0.1	82.3 ± 0.3	Res152	78.0 ± 0.3	81.5 ± 0.4
Dense121	80.2 ± 0.2	80.5 ± 0.1	Dense161	80.1 ± 0.3	82.2 ± 0.2

Assumption of Interpretable Models Might Exist

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

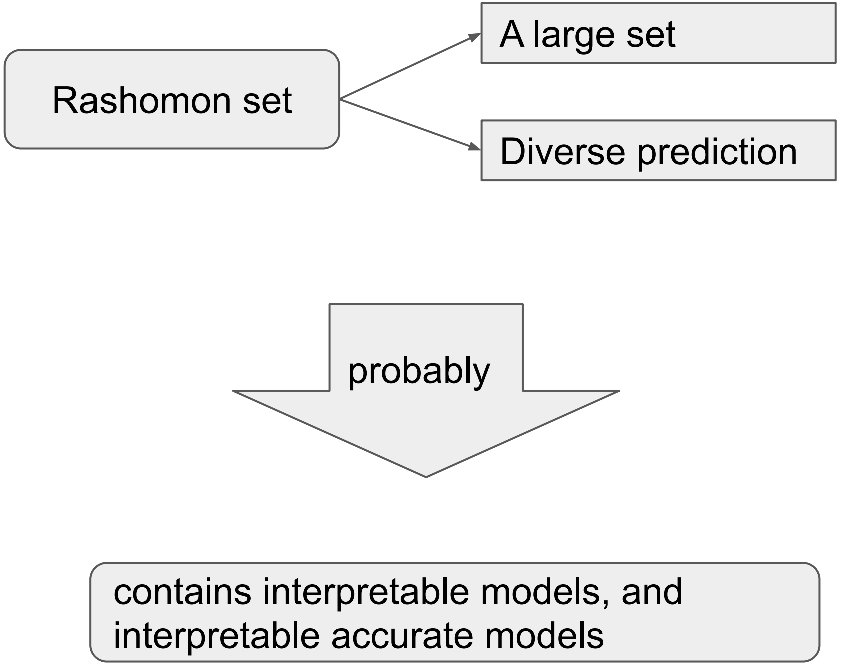
Rashomon set

definition: the set of reasonably accurate predictive models (say within a given accuracy from the best model accuracy).

data finite => many close-to-optimal models that predict differently from each other, e.g. RF, NN, SVM

A large set

Assumption of Interpretable Models Might Exist

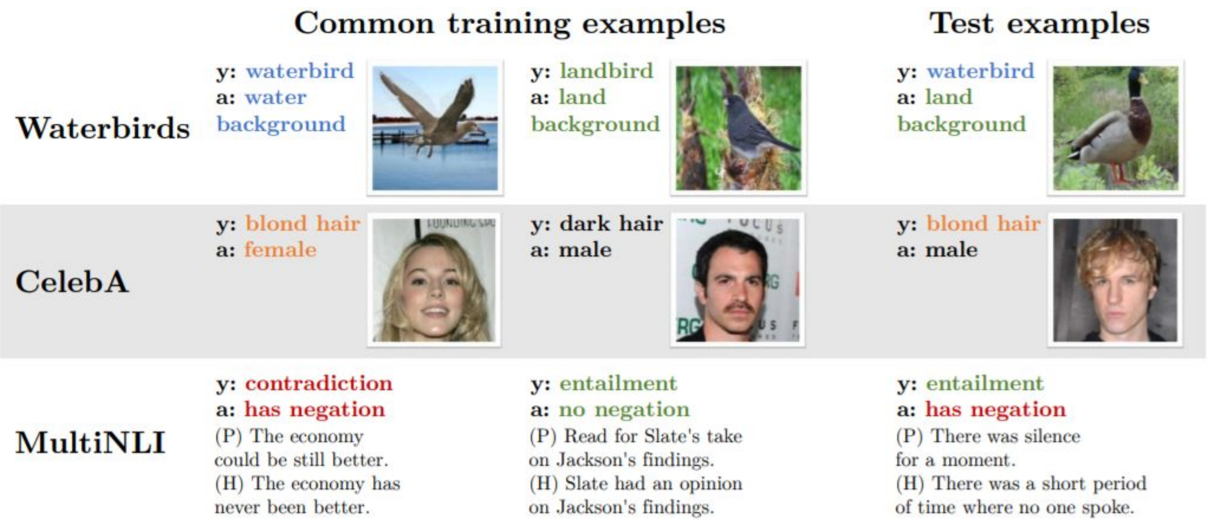


Algorithm Stability

A common criticism of decision trees: They are not stable.

small changes in the training data => completely different trees

which tree to choose? ~~ linear models when there are highly correlated features



Algorithm Stability

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

drawbacks? → advantages?

Not stable → Large Rashomon set → Great skills to choose interpretable model

Adding regularization to an algorithm increases stability, but also limits flexibility of the user to choose which element of the Rashomon set which would be more desirable.

Conclusion

The paper appeals that we should pay more attention and give more efforts to interpretability rather than explanation in both academic and industrial fields.

Hoping everyone will have Interpretable Models with High Accuracies!

Questions

Q1

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

What could be some issues with “Explanations” of Black Box Models ?

A. Lack of Confounding Issues in Data while generating “Explanations”
B. Lack of Informativeness of “Explanations”
C. Lack of Faithfulness to Original Model Computations
D. Issues with Counterfactual Explanations

Ans : B,C, D

Q2

Department of Computer Sciences
UNIVERSITY OF WISCONSIN–MADISON

What is the size of the model by CORELS in page 6 figure 3 based on the paper?

A.3

B.4

C.5

D.6

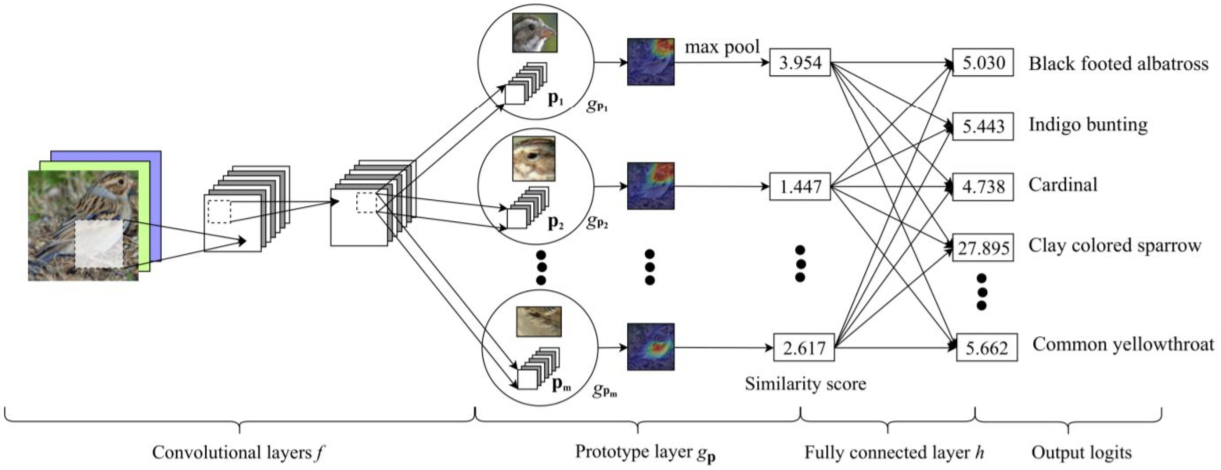
IF	age between 18-20 and sex is male	THEN predict arrest (within 2 years)
ELSE IF	age between 21-23 and 2-3 prior offenses	THEN predict arrest
ELSE IF	more than three priors	THEN predict arrest
ELSE	predict no arrest.	

Ans: A

Q3

What's the main idea of Chen, Li work on classification?

A. prototype layer to find similarity with prototype to get Interpretability
B. Multi-process to classify from roughly to precisely to get Interpretability
C. Self-attention to get saliency map without supervision to get Interpretability
D. All above.



Ans: A