Interpretable Machine Learning

Dr. Alex Williams

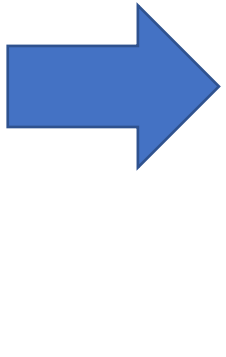
November 9, 2020

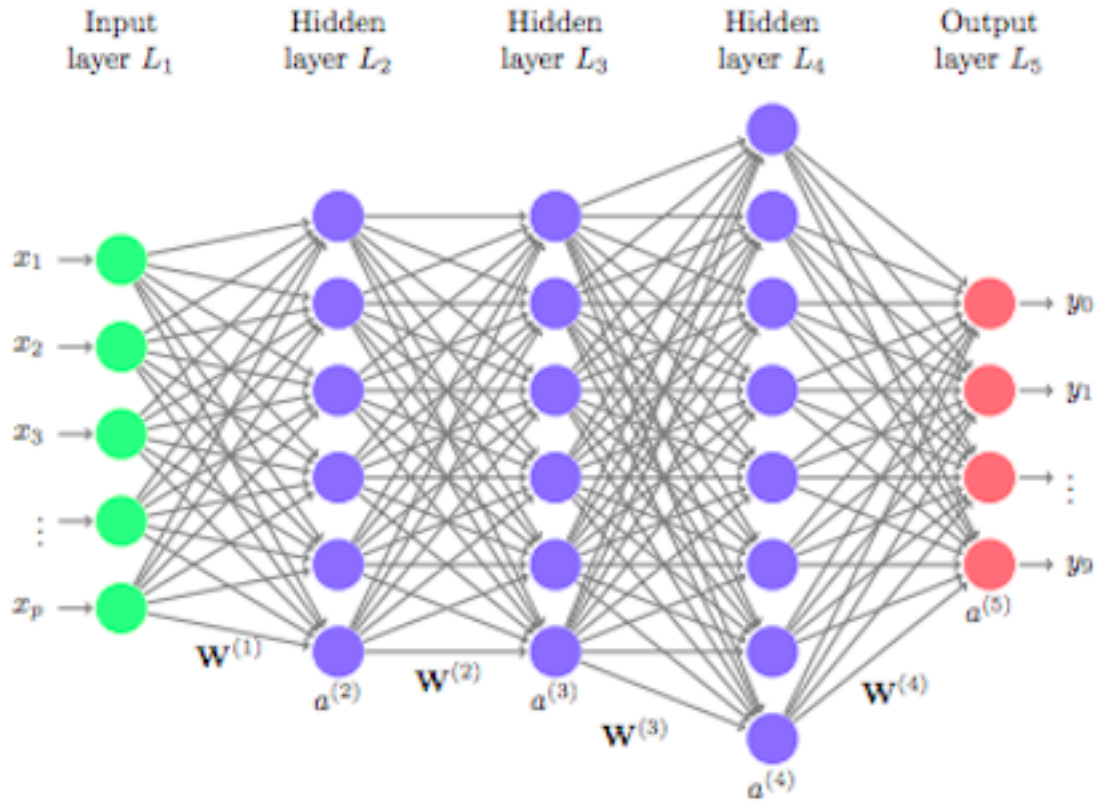
COSC 425: Introduction to Machine Learning
Fall 2020 (CRN: 44874)

Today's Agenda

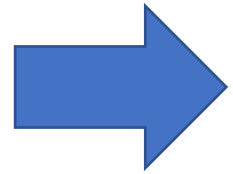
We will address:

1. Interpretability in Machine Learning



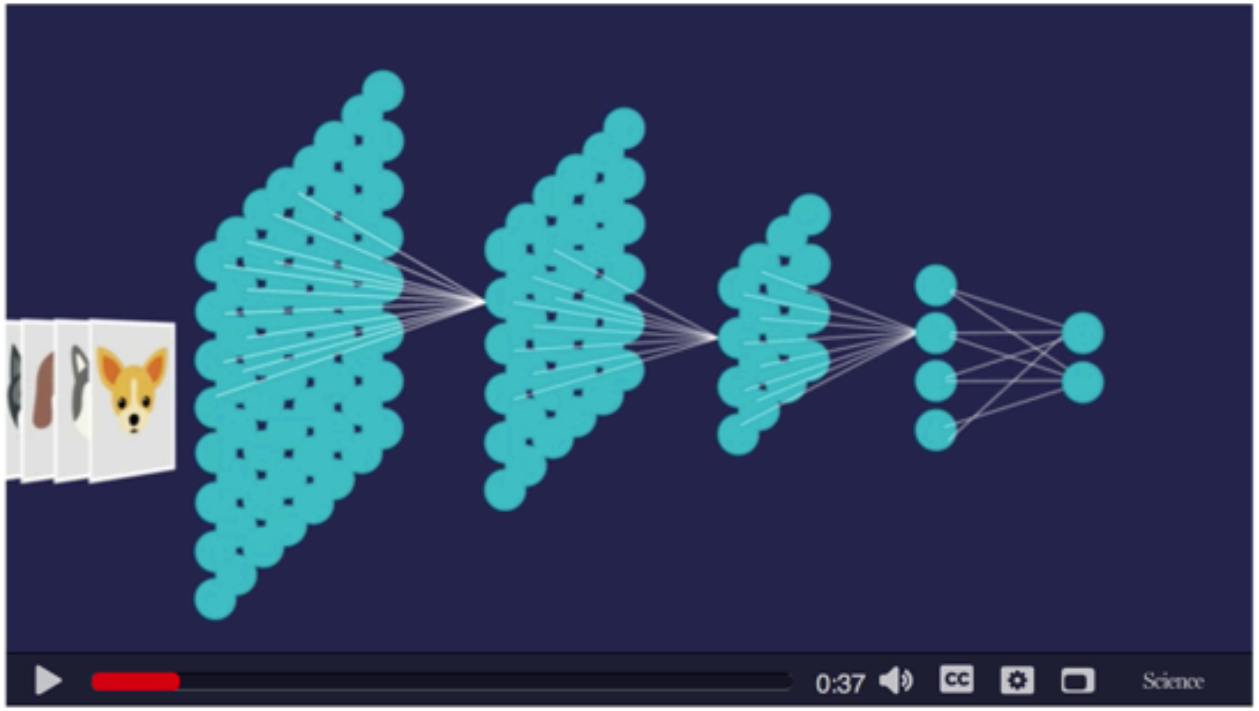


A Deep Neural Network for Object Detection



Label:
Boat

… Why?



How AI detectives are cracking open the black box of deep learning

By Paul Voosen | Jul. 6, 2017 , 2:00 PM

Jason Yosinski sits in a small glass box at Uber's San Francisco, California, headquarters, pondering the mind of an artificial intelligence. An Uber research scientist, Yosinski is performing a kind of brain surgery on the AI running on his laptop. Like many of the AIs that will

Defining Interpretability

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

→ The definition suggests that interpretability is a function of both the user and the machine.

Interpretability may not be necessary.

→ If you don't care how a decision is made. (Just accuracy is enough!)

→ When the model is used in low-risk environments where a mistake has no severe consequences, e.g. Netflix Recommendations

→ For a model that's been extensively tested and evaluated, e.g. OCR

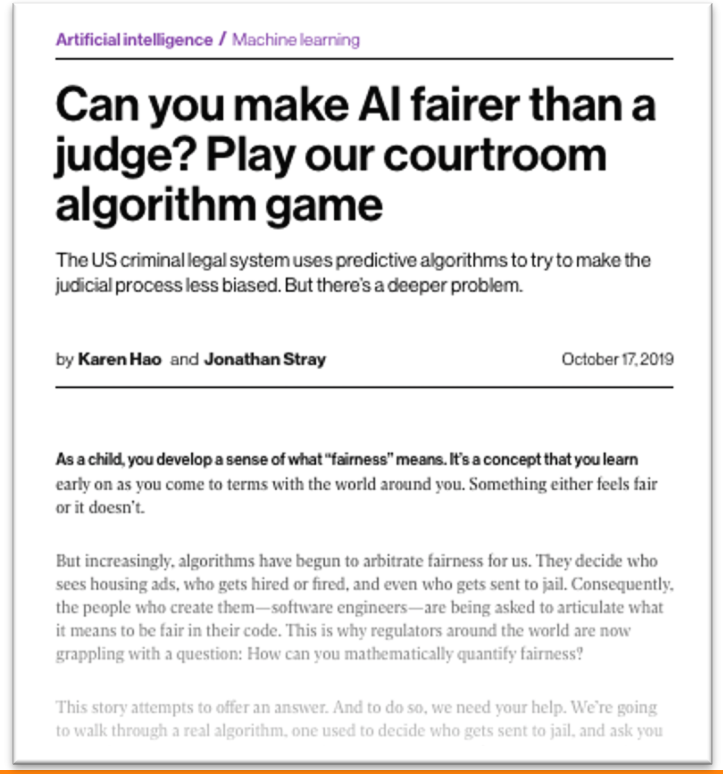
→ When exposing information increases the risk of users gaming a system, e.g. credit scoring.

Necessary Scenarios

Interpretability is becoming increasingly necessary.
→ When the prediction is associated with some serious consequences
→ If users need to debug a poorly performing system

Example: COMPAS (2017)
→ A computer system that predicts a defendant's likelihood to re-offend.

→ How can the system explain itself reliably?
→ **Related**: Should such a system exist? Why?



Artificial intelligence / Machine learning

Can you make AI fairer than a judge? Play our courtroom algorithm game

The US criminal legal system uses predictive algorithms to try to make the judicial process less biased. But there's a deeper problem.

by **Karen Hao** and **Jonathan Stray** October 17, 2019

As a child, you develop a sense of what "fairness" means. It's a concept that you learn early on as you come to terms with the world around you. Something either feels fair or it doesn't.

But increasingly, algorithms have begun to arbitrate fairness for us. They decide who sees housing ads, who gets hired or fired, and even who gets sent to jail. Consequently, the people who create them—software engineers—are being asked to articulate what it means to be fair in their code. This is why regulators around the world are now grappling with a question: How can you mathematically quantify fairness?

This story attempts to offer an answer. And to do so, we need your help. We're going to walk through a real algorithm, one used to decide who gets sent to jail, and ask you

Explanations

If machine learning models can explain decisions, the following traits can be checked more easily (Doshi-Velez and Kim, 2017).

1. Fairness: Making sure that predictions are unbiased and not discriminating against protected groups.

2. Privacy: Ensuring that sensitive information in the data is protected.

3. Reliability and Robustness: Test that small changes in the input don't lead to big changes in the prediction.

4. Causality: Check if only causal relationships, and not chance correlation between input and output, are picked up.

5. Trust: Check if humans trust a system that explains the decisions compared to a black box.

Interpretability Methods

One can make a model more interpretable by:

→ **Showing feature summary statistics / visualizations**

  → Describing how each feature affects model predictions

→ **Showing model internals (e.g. learned weights)**

→ **Showing examples**

  → e.g. Counterfactuals: To explain the prediction of a data point, find a similar data point by changing some of the features for which the predicted outcomes flip.

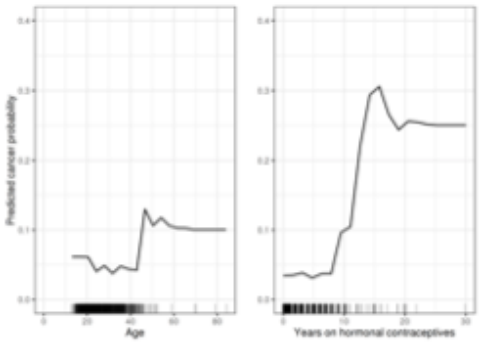
→ Approximating the black box model using an interpretable model.

Note: These methods can be:

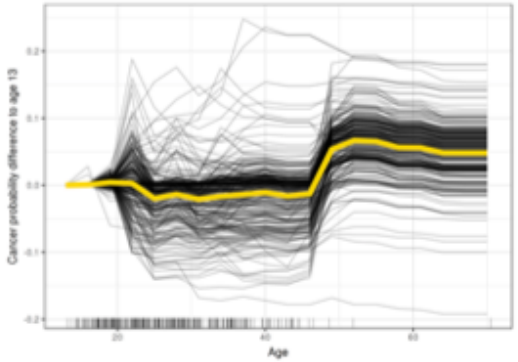
→ Model-specific or model-agnostic

→ Local (explaining a single prediction) or global (explaining the entire model behavior)

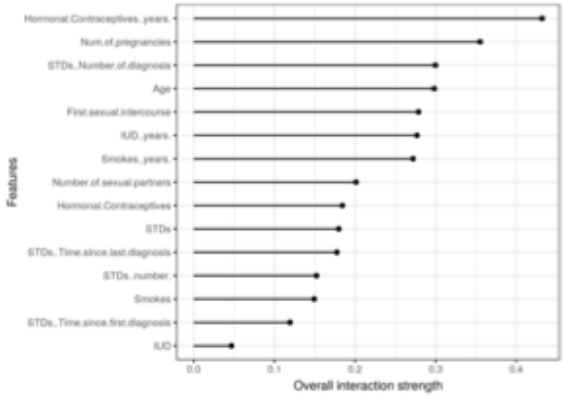
Feature Summary and Visualization



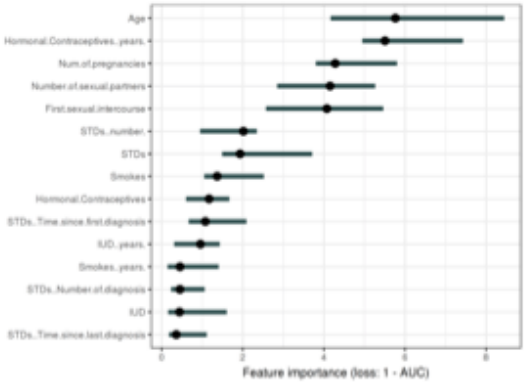
Partial Dependence Plots



Individual Conditional Expectation



Feature Interaction



Feature Importance

Global Surrogate Model

Choose a dataset ***X'***
→ Same dataset from training black box model.
→ A new dataset from the same distribution

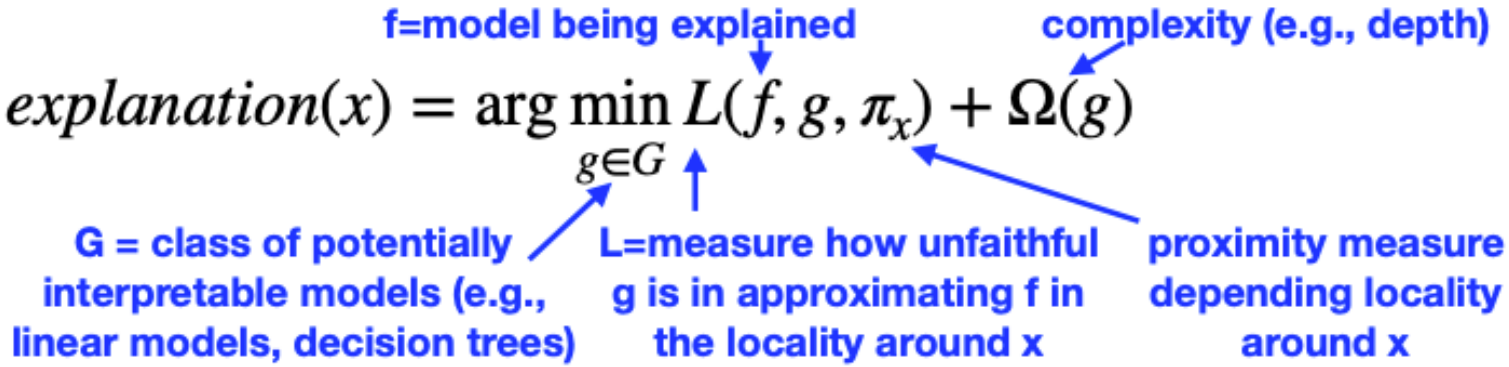
After dataset is chosen / generated, do the following:
1. For ***X'***, get predictions ***Y'*** of the black box model.
2. Choose an interpretable model (linear model, decision tree)
3. Train an interpretable model on ***X'*** and its prediction ***Y'***.
4. Using this "surrogate" model:
    → Measure how well it replicates the prediction of the black box model.
    → Interpret and visualize the surrogate model.

(Ribeiro et al., 2016)

Local Surrogate Models (LIME)

Local Surrogate Models are interpretable models used to explain individual predictions of the black box machine learning models.

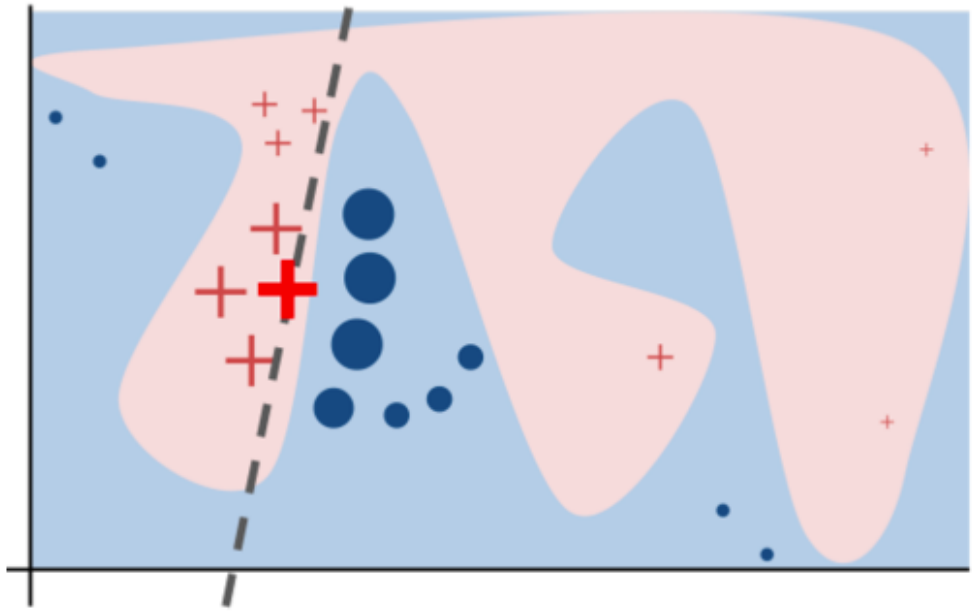
1. Choose the instance for which you want to have an explanation of its black box prediction.
2. Perturb your dataset and get the black box predictions for these new points.
3. Weight the new samples by their proximity to the instance of interest.
4. Fit a weighted, interpretable model on the dataset with the variations.



5. Explain the prediction by interpreting the local model.

(Ribeiro et al., 2016)

Local Surrogate Models (LIME)



Toy example to present the intuitions for LIME: The black box model's complex decision function ***f*** (unknown to LIME) is presented by the blue/pink background, which cannot be approximated well with a linear model. The bright bold red cross is the instance being explained. LIME samples instances, get predictions using f, weighs them by the proximity to the instance being explained (represented here by size). The dashed line is the learned explanation that is locally (but not globally) faithful.

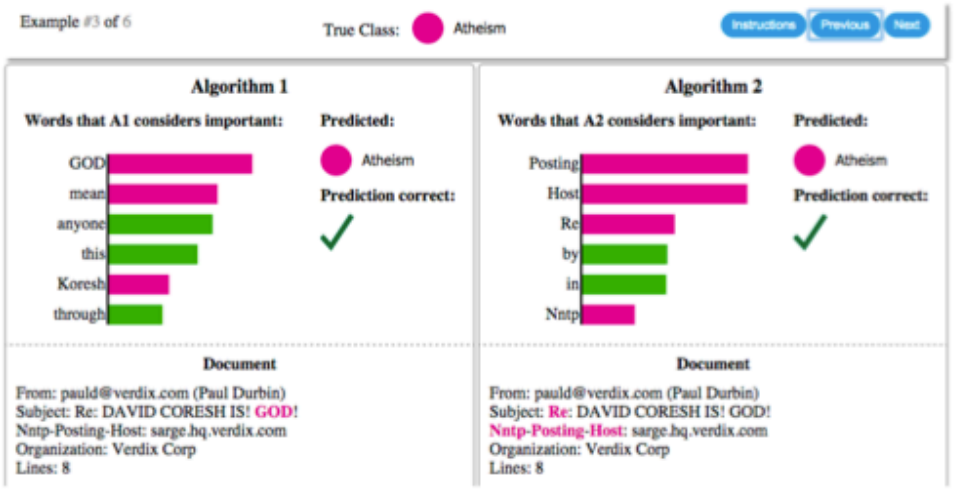
(Ribeiro et al., 2016)

LIME for Images

Google's Inception v3
Pre-Trained Object Classification

(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

LIME
A Tool for Explaining Object Classification



Example-Based Explanations

Explaining a model by selecting instances of the dataset.
→Instead of creating summaries of the features.

On makes sense if the examples themselves make sense to people immediately.
→ Similar to how people make decisions based on similar cases they've seen before.

Counterfactual Explanations
→ A counterfactual explanation of a prediction describes the smallest change to the feature values that changes the predictions of a predefined output.

Counterfactual Explanations

"How would input features need to be changed to get a predicted score of 0?"

Score	GPA	LSAT	Race	GPA x'	LSAT x'	Race x'
0.17	3.1	39.0	0	3.1	34.0	0
0.54	3.7	48.0	0	3.7	32.4	0
-0.77	3.3	28.0	1	3.3	33.5	0
-0.83	2.4	28.5	1	2.4	35.8	0
-0.57	2.7	18.3	0	2.7	34.9	0

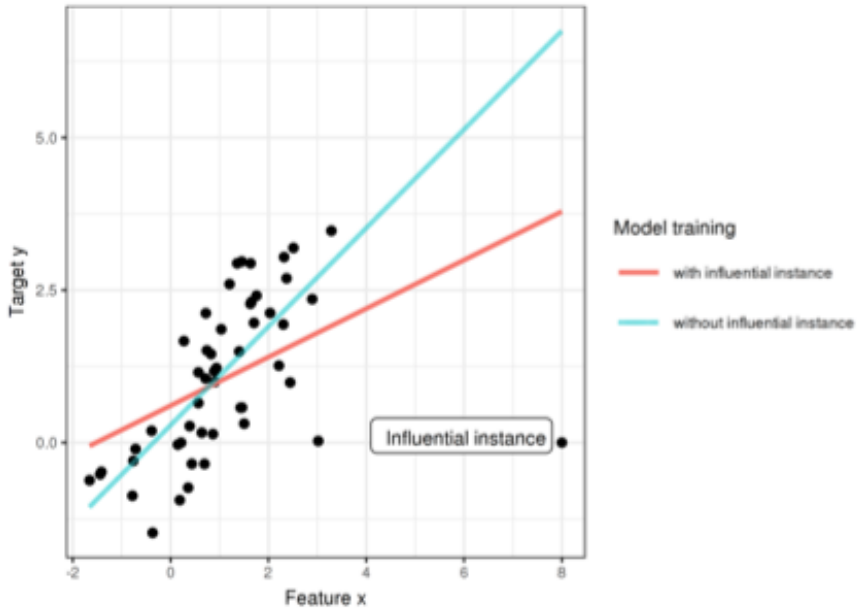
The first column contains the predicted score, the next 3 columns the original feature values and the last 3 columns the counterfactual feature values that result in a score close to 0. The first two rows are students with above-average predictions, the other three rows below-average. The counterfactuals for the first two rows describe how the student features would have to change to decrease the predicted score and for the other three cases how they would have to change to increase the score to the average. The counterfactuals for increasing the score always change the race from black (coded with 1) to white (coded with 0) which shows a racial bias of the model. The GPA is not changed in the counterfactuals, but LSAT is.

Influential Instances

Machine learning models are a product of training data.
→ Deleting one of the training instances can affect the resulting model.

→ An instance is **"influential"** when its deletion from the training data considerably changes the parameters of the predictions of the model.

So… what makes a good explanation?



Defining "Good" Explanations (Miller, 2017)

Contrastive

→ People don't ask why a certain prediction was made, but why this prediction was made instead of another prediction.

→ Counterfactual: How would the Y look like, if example X were different?

Selective

→ People do not expect explanations to cover the actual and complete list of the cause.

→ Make explanations short, e.g. 1-3 reasons.

User-Centered:

→ Understandable to the target audience.

Coherent with Prior Beliefs

→ Humans tend to ignore information that is not coherent with their prior beliefs

→ Confirmation Bias!

Defining "Good" Explanations (Miller, 2017)

General and Probable

→ A feature that can explain a lot of examples.

Focus on the Abnormal

→ People focus on the abnormal causes (i.e. causes that had a small likelihood, but happened anyways) to explain events.

→ e.g., if one of the input features was abnormal (a rare category) and the feature influenced the prediction, it should be included in the explanation.

Truthful

→ Sometimes referred to as "high fidelity".

→ i.e., the explanation must correspond to how the model behaves at least in the vicinity of the instance being predicted and explained.

Evaluating Interpretability

Application-Level Evaluation (Real Task with Targeted/Expert Users):
→ e.g. Having radiologists test detection software.

Human-Level Evaluation (Simple Task with General Applicability to People at Large):
→ e.g. Show users different explanations and the human chooses the best.

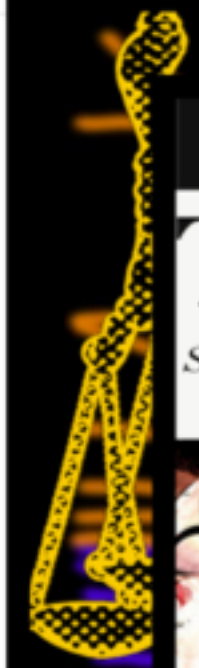
Function-Level Evaluation (Proxy Task with Machine-driven Technique):
→ e.g. use depth of the tree (e.g. preference for shorter trees is better) as a proxy.

Many open research questions in interpretability!
→ Consider a physician working with your Decision Tree implementation in Project 1.
    → How interpretable is it out of the box? What would you change?

Can an Algorithm Hire Better Than a Human?

Claire Cain Miller

The Algorithm That Beats Your Bank Manager



The Marshall Project Nonprofit journalism about criminal justice SEARCH ABOUT DONATE

The New Science of Sentencing

Should prison sentences be based on crimes that haven't been committed yet?

This article is more than **4 years old**

A beauty contest was judged by AI and the robots didn't like dark skin

The first international beauty contest decided by an algorithm has sparked controversy after the results revealed one glaring factor linking the winners

▲ One expert says the results offer 'the perfect illustration of the problem' with machine bias. Photograph: Fabrizio Bensch/Reuters

The first international beauty contest judged by "machines" was supposed to use objective factors such as facial symmetry and wrinkles to identify the

Is an algorithm any less racist than a human?

Employers trusting in the impartiality of machines sounds like a good plan to eliminate bias, but data can be just as prejudiced as we are

▲ A wealth of startups have sprung up in recent years to address the appetite for more diverse workforces by utilising algorithms. Photograph: Alamy Stock Photo

We would all like to fancy ourselves as eminently capable of impartiality, able to make decisions without prejudices - especially at work. Unfortunately, the reality is that human bias, both conscious and unconscious, can't help but come into

Machine Bias

There's software used across the country to predict future criminals. And it's biased against blacks.

by Julia Angwin, Jeff Larson, Surya Mattu and Lauren Kirchner, ProPublica
May 23, 2016

ON A SPRING AFTERNOON IN 2014, Brisha Borden was running late to pick up her god-sister from school when she spotted an unlocked kid's blue Huffy bicycle and a silver Razor scooter. Borden and a friend grabbed the bike and scooter and tried to ride them down the street in the Fort Lauderdale suburb of Coral Springs.

Just as the 18-year-old girls were realizing they were too big for the tiny conveyances — which belonged to a 6-year-old boy — a woman came running after them saying, "That's my kid's stuff." Borden and her friend immediately dropped the bike and scooter and

Why does Interpretability matter?
→ Because data is biased.

Data reflects the biases of the labeler.
→ If the labeler maintains prejudices in their labeling process, the data will reflect these prejudices.

Machine learning is becoming a statistical tool to support a growing number of decisions that affect our success and well-being.

→ COMPAS is a leading example of bias.

ProPublica Analysis of COMPAS Algorithm

	White	Black
Wrongly Labeled High-Risk	23.5%	**44.9%**
Wrongly Labeled Low-Risk	**47.7%**	28.0%

Bias isn't just about Race.

Video here.

Beyond COSC425

You should now:

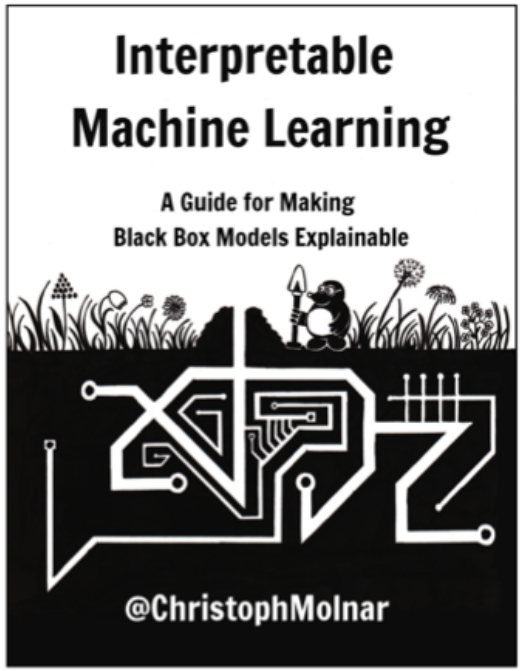
1. Understand the need for interpretable models.
2. Understand the importance of bias in machine learning.

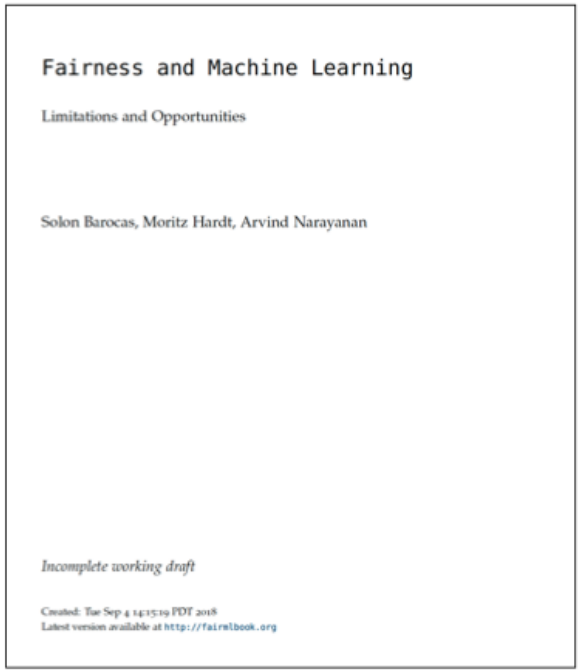
More Reading

→ Interpretable ML.

→ Fairness and ML.

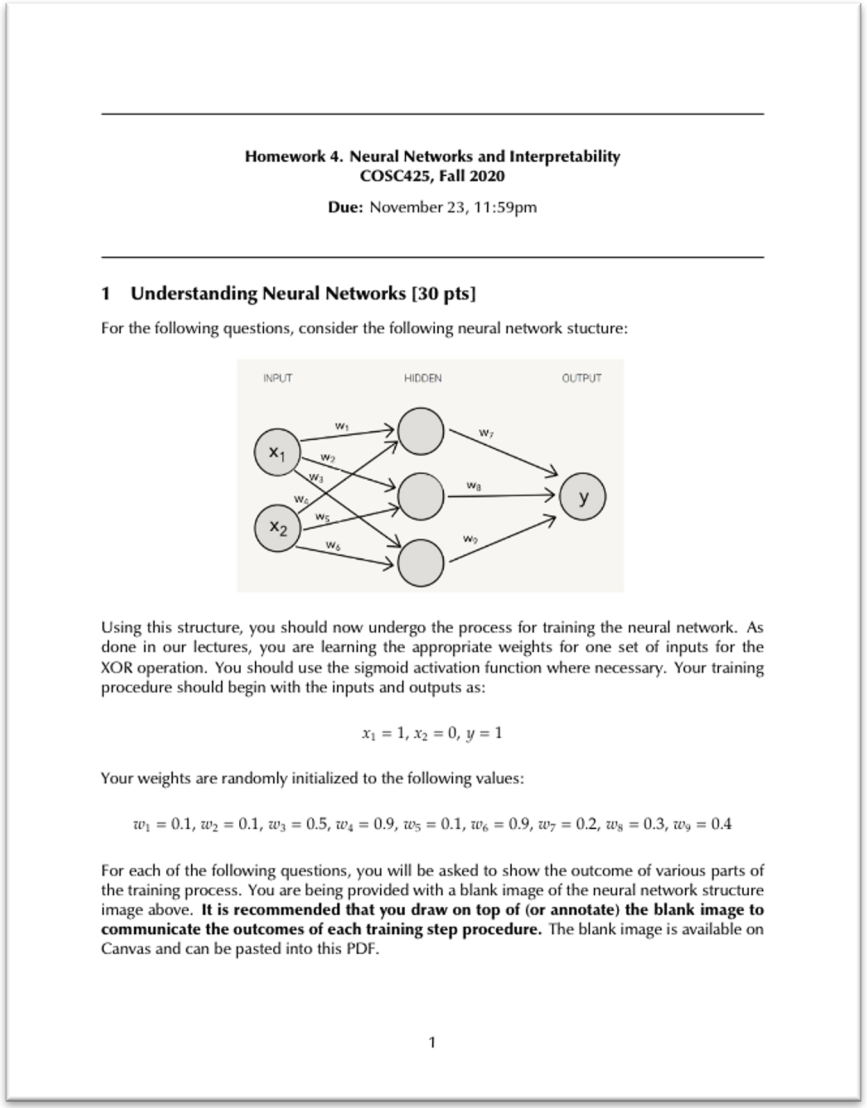
Both have my endorsement!





Homework 4

Due: November 23rd @ 11:59pm



Homework 4. Neural Networks and Interpretability
COSC425, Fall 2020

Due: November 23, 11:59pm

1 Understanding Neural Networks [30 pts]

For the following questions, consider the following neural network stucture:

Using this structure, you should now undergo the process for training the neural network. As done in our lectures, you are learning the appropriate weights for one set of inputs for the XOR operation. You should use the sigmoid activation function where necessary. Your training procedure should begin with the inputs and outputs as:

$$x_1 = 1, x_2 = 0, y = 1$$

Your weights are randomly initialized to the following values:

$$w_1 = 0.1, w_2 = 0.1, w_3 = 0.5, w_4 = 0.9, w_5 = 0.1, w_6 = 0.9, w_7 = 0.2, w_8 = 0.3, w_9 = 0.4$$

For each of the following questions, you will be asked to show the outcome of various parts of the training process. You are being provided with a blank image of the neural network structure image above. **It is recommended that you draw on top of (or annotate) the blank image to communicate the outcomes of each training step procedure.** The blank image is available on Canvas and can be pasted into this PDF.

1

Next Time

We will address:

1. Reinforcement Learning