Harish Guruprasad Ramaswamy
IIT Madras and RBCDSAI

1. Need for interpretability in AI

2. Neural Models
   - Linear
   - Trees
   - Multi-layer Perceptrons
   - CNNs

- Attention networks
- RNNs

3. Interpretability Paradigms

- Intrinsic
- Reverse Engg.
- Feature visualization
- Attribution maps
- Influence functions
- Other

4. Caveat Emptor

You are likely to have a heart attack within a month

What?? Why?

Because
“BP*top_right_X_ray+
(Sugar-sin(BP^2)) > 45”

?!*#!!!

□□□

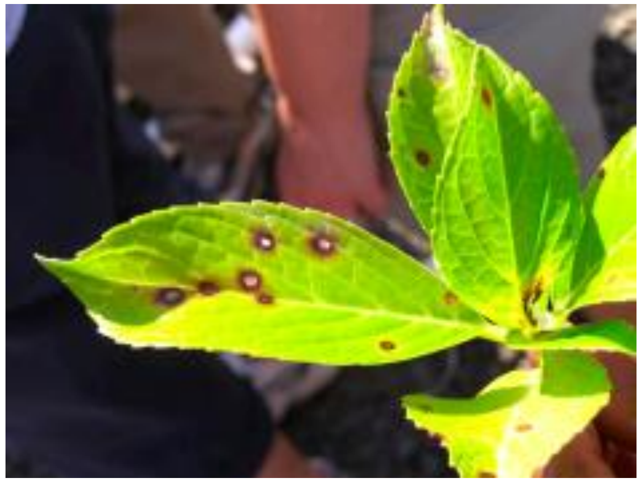
MAKE YOUR RIGHTS
16°C
Mantenha
Sua Cidade Limpa

Training data

Healthy

Healthy Diseased Diseased

Testing data



- Accuracy - Interpretability tradeoff
- General philosophy of the model — Global methods •

Precise explanations of a prediction —Local Methods

Simple fully observed models are often weak, e.g. Markov Chain models for speech.

Latent/Hidden variable models improve performance significantly. e.g. Hidden Markov models.

1. Need for interpretability in AI

2. Machine Learning Models
   - Linear
   - Trees
   - Multi-layer Perceptrons
   - CNNs
   - Attention networks
   - RNNs

3. Interpretability Paradigms
   - Intrinsic
   - Reverse Engg.

- Feature visualization
- Attribution maps
- Influence functions
- Other

4. Caveat Emptor

- Output is a linear function of input.
- e.g. Heart Risk = (BP-120)+(Sugar-100)+10(Cholesterol)
- Ultimate interpretability

SF NY NY SF

Cat

1. Need for interpretability in AI

2. Neural Models
   - Linear
   - Trees
   - Multi-layer Perceptrons

- CNNs
- Attention networks
- RNNs

3. Interpretability Paradigms
    - Intrinsic
    - Reverse Engg.
    - Feature visualization
    - Attribution maps
    - Influence functions
    - Training for Interpretability
    - Other

4. Caveat Emptor

- Linear and small depth tree models are intrinsically interpretable.
- Simple attention models are also interpretable to an extent.

Xu et al. (2015)

Figuring out the meaning of each element of the state vector:

Karpathy et al. 2016

Local Syntactic

Block

Delimiter

Pande et al. 2020

Block	CLS	SEP

dobj	Amod	advm
Local	Syntax	Nsubj

Pande et al. 2020

1. Black Box Approaches.
    1. Saliency
    2. Occlusion
    3. Class Activation Maps.
2. Optimising the Model
    1. Feature visualisation
    2. Other

Original
image GradCAM: cat GradCAM: dog

Selvaraju et al. (2017)

Importance of

channel ‘k’ for class ‘c’ =

Selvaraju et al. (2017)

Desai et al. (2020)

Desai et al. (2020)

Visualize a learned filter by finding an artificial image that triggers it.

Olah et al. 2017

Visualize a learned filter by finding an artificial image that triggers it.

Olah et al. 2017

Olah et al. 2017

Why optimize over hallucinated images?

Olah et al. 2017

https://distill.pub/2018/building-blocks/

Olah et al. 2018

- Can we explain predictions using training data?
- “ How would the model’s predictions change on a

given test point, if we did not have a given training point?”

- Remove/Perturb/Repeat a sample and retrain!

- Influence functions: A more efficient approach for the same.

Koh and Liang (2017)

Most harmful training image for a wrong prediction

Most useful training image for a right prediction

Koh and Liang (2017)

1. Select a dataset X. This can be the same dataset that was used for training the black box model or a new dataset from the same distribution. You could even select a subset of the data or a grid of points, depending

on your application.

2. For the selected dataset X, get the predictions of the black box model.

3. Select an interpretable model type (linear model, decision tree, …). 4.

Train the interpretable model on the dataset X and its predictions.

5. Congratulations! You now have a surrogate model.

6. Measure how well the surrogate model replicates the predictions of the black box model.

7. Interpret the surrogate model.

Molnar (2020)

— Simple model prediction

— Complex model prediction

R^2 captures how much better the simple model is at explaining the complex model, when compared to a constant.

Molnar (2020)

Example: Explain a complex SVM model for predicting daily number of rented bikes using a regression tree.

R^2=0.77 Molnar (2020)

Dhurandhar et al (2018)

Pertinent negatives:

Pertinent positive:

Dhurandhar et al (2018)

1. Need for interpretability in AI

2. Neural Models
   - Linear
   - Trees
   - Multi-layer Perceptrons
   - CNNs
   - Attention networks
   - RNNs

3. Interpretability Paradigms
   - Intrinsic
   - Reverse Engg.
   - Feature visualization
   - Attribution maps
   - Influence functions
   - Other

4. **Caveat Emptor**

• Chris Olah’s blog and Distill posts: colah.github.io •

Christoph Molnar. *Interpretable Machine Learning.*

• Karpathy et al. *Visualizing and understanding RNNs.*ICLR 2016.

• Kian Katanfaroosh. Stanford Interpretability Lecture. •

Xu et al. *Show, Attend and Tell.* ICML 2015.

• Koh and Liang. *Understanding Black-box Predictions via Influence Functions.* ICML 2017.

• Selvaraju et al. GradCAM. ICCV 2017.