FYS: AI in Healthcare

Interpretability in Machine Learning

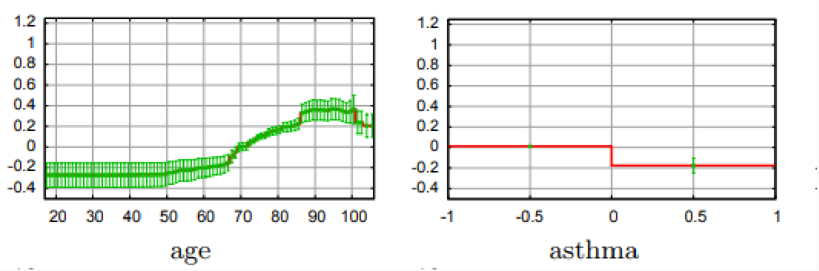
John Lalor

October 23, 2018

Admin

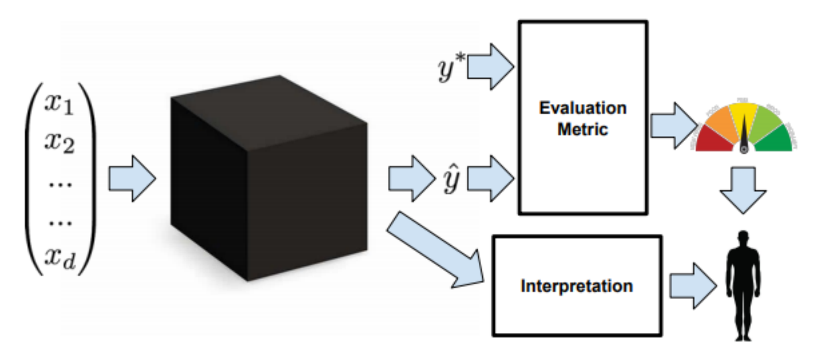
Midterm questions?

Why is interpretability important?



Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

Evaluation-Interpretability Relationship



Lipton, Zachary C. "The Mythos of Model Interpretability." Queue 16.3 (2018): 30.

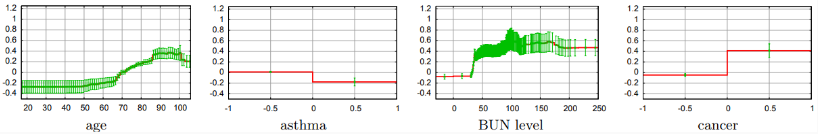
Why do we want interpretability?

Trust

https://waymo.com/ontheroad/

Why do we want interpretability?

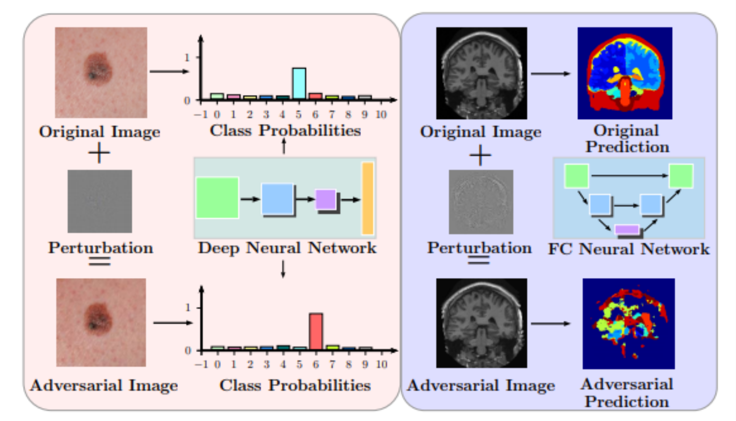
Causality



Caruana, Rich, et al. "Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission." Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2015.

Why do we want interpretability?

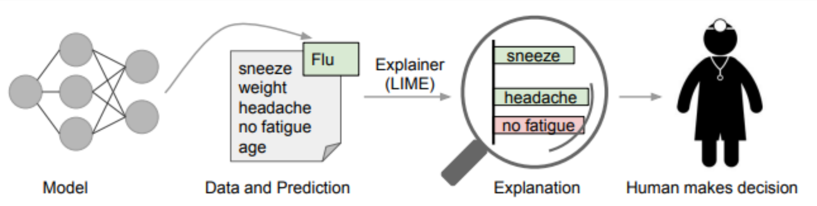
Transferability



Paschali, Magdalini, et al. "Generalizability vs. Robustness: Adversarial Examples for Medical Imaging." arXiv:1804.00504 (2018).

Why do we want interpretability?

Informativeness

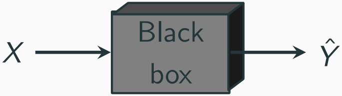


Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. "Why should i trust you?: Explaining the predictions of any classifier." Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. ACM, 2016.

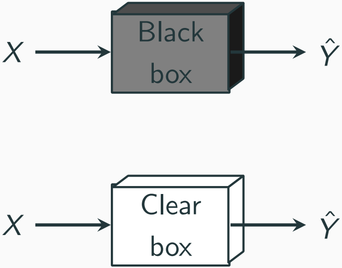
Why do we want interpretability?

Fairness (e.g. *the right to an explanation*)

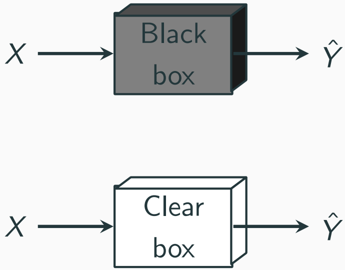
Transparency



Transparency



Transparency



Model transparency (simulatability)

Parameter transparency (decomposability)

Training transparency (algorithmic transparency)

Model transparency

Given the input data and model parameters:

Could a human produce an output?

In some reasonable amount of time?

Model transparency

Given the input data and model parameters:

- Could a human produce an output?
  - In some reasonable amount of time?
- Example: BOW logistic regression

Model transparency

Given the input data and model parameters:

- Could a human produce an output?
  - In some reasonable amount of time?
- Example: BOW logistic regression
- Example: fully-connected DNN with hidden layer size 10

Parameter transparency

Each part of the model is intuitive

- Inputs
- Parameters
- Calculations

Parameter transparency

Each part of the model is intuitive

- Inputs
- Parameters
- Calculations

Ex.: Descriptive decision tree nodes

Parameter transparency

Each part of the model is intuitive

- Inputs
- Parameters
- Calculations

Ex.: Descriptive decision tree nodes

Ex.: Linear model parameters

Parameter transparency

Each part of the model is intuitive

- Inputs
- Parameters
- Calculations

Ex.: Descriptive decision tree nodes

Ex.: Linear model parameters

Caveat: Can be fragile depending on pre-processing

Algorithmic transparency

Insight into the decision-making process

Algorithmic transparency

Insight into the decision-making process

- Linear models?

Algorithmic transparency

Insight into the decision-making process

- Linear models?
- DNNs?

Algorithmic transparency

Insight into the decision-making process

- Linear models?
- DNNs?
- Humans?

Explainability

Human interpretability

After the fact interpretation

- Not part of model training

Text

a beer that is not sold in my neck of the woods , but managed to get while on a roadtrip . poured into an imperial pint glass with a generous head that sustained life throughout . nothing out of the ordinary here , but a good brew still . body was kind of heavy , but not thick . the hop smell was excellent and enticing . very drinkable

very dark beer . pours a nice finger and a half of creamy foam and stays throughout the beer . smells of coffee and roasted malt . has a major coffee-like taste with hints of chocolate . if you like black coffee , you will love this porter . creamy smooth mouthfeel and definitely gets smoother on the palate once it warms . it 's an ok porter but i feel there are much better one 's out there .

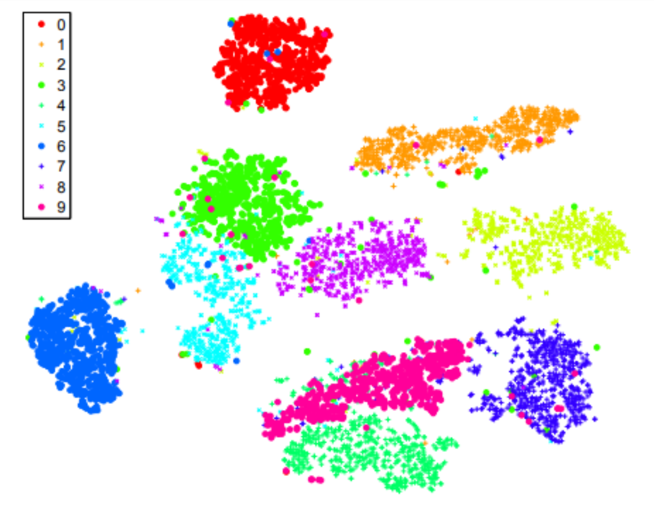
i really did not like this . it just seemed extremely watery . i dont ' think this had any carbonation whatsoever . maybe it was flat , who knows ? but even if i got a bad brew i do n't see how this would possibly be something i 'd get time and time again . i could taste the hops towards the middle , but the beer got pretty nasty towards the bottom . i would never drink this again , unless it was free . i 'm kind of upset i bought this .

a : poured a nice dark brown with a tan colored head about half an inch thick , nice red/garnet accents when held to the light . little clumps of lacing all around the glass , not too shabby . not terribly impressive though s : smells like a more guinness-y guinness really , there are some roasted malts there , signature guinness smells , less burnt though , a little bit of chocolate … … m : relatively thick , it is n't an export stout or imperial stout , but still is pretty hefty in the mouth , very smooth , not much carbonation . not too shabby d : not quite as drinkable as the draught , but still not too bad . i could easily see drinking a few of these .

Figure 3: Examples of extracted rationales indicating the sentiments of various aspects. The extracted texts for appearance, smell and palate are shown in red, blue and green color respectively. The last example is shortened for space.

Lei, Tao, Regina Barzilay, and Tommi Jaakkola. "Rationalizing Neural Predictions." EMNLP 2016.

Visualizations



Maaten, Laurens van der, and Geoffrey Hinton. "Visualizing data using t-SNE." Journal of machine learning research 9.Nov (2008): 2579-2605.

Local explanations

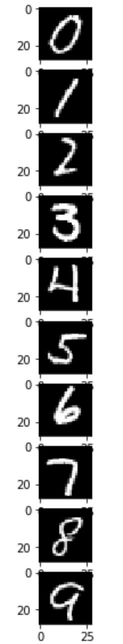
Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv:1312.6034 (2013).

Explanations by example

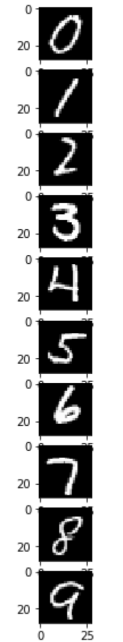
Newspapers			
New York	New York Times	Baltimore	Baltimore Sun
San Jose	San Jose Mercury News	Cincinnati	Cincinnati Enquirer
NHL Teams			
Boston	Boston Bruins	Montreal	Montreal Canadiens
Phoenix	Phoenix Coyotes	Nashville	Nashville Predators
NBA Teams			
Detroit	Detroit Pistons	Toronto	Toronto Raptors
Oakland	Golden State Warriors	Memphis	Memphis Grizzlies
Airlines			
Austria	Austrian Airlines	Spain	Spainair
Belgium	Brussels Airlines	Greece	Aegean Airlines
Company executives			
Steve Ballmer	Microsoft	Larry Page	Google
Samuel J. Palmisano	IBM	Werner Vogels	Amazon

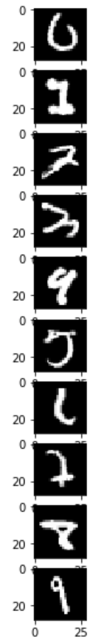
Mikolov, Tomas, et al. "Distributed representations of words and phrases and their compositionality." Advances in neural information processing systems. 2013.

Explanations by example



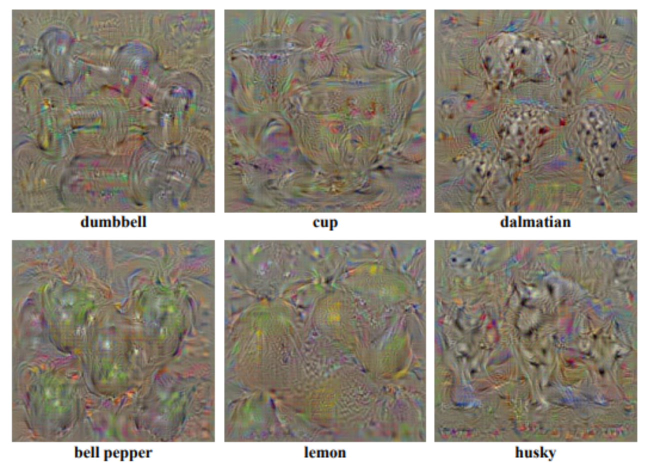
Explanations by example





Interpretability Examples

Saliency maps



$$\arg\max_{I} S_c(I) - \lambda ||I||_2^2$$

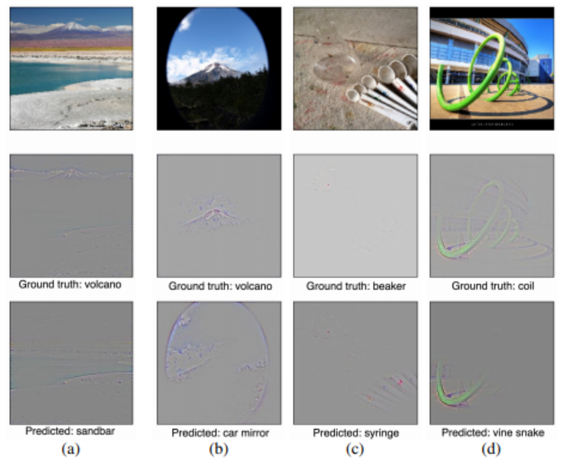
Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv:1312.6034 (2013).

Saliency maps

Simonyan, Karen, Andrea Vedaldi, and Andrew Zisserman. "Deep inside convolutional networks: Visualising image classification models and saliency maps." arXiv:1312.6034 (2013).

Gradient based localization: Grad-CAM

Analyzing failures



Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." ICCV. 2017.

Gradient based localization: Grad-CAM

Handling adversarial noise

Boxer: 0.40 Tiger Cat: 0.18 — (a) Original image

Airliner: 0.9999 — (b) Adversarial image

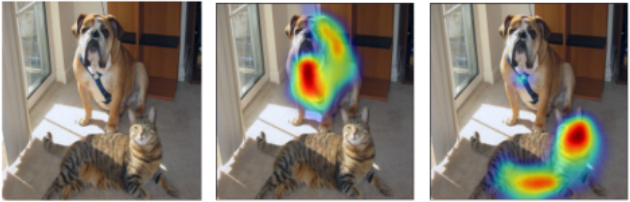
Boxer: 1.1e-20 — (c) Grad-CAM "Dog"

Tiger Cat: 6.5e-17 — (d) Grad-CAM "Cat"

Selvaraju, Ramprasaath R., et al. "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization." ICCV. 2017.

Gradient based localization: Grad-CAM

Counterfactuals



(a) Original Image (b) Cat Counterfactual exp (c) Dog Counterfactual exp

Selvaraju, Ramprasaath R., et al. “Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization.” ICCV. 2017.

Interpretability in healthcare

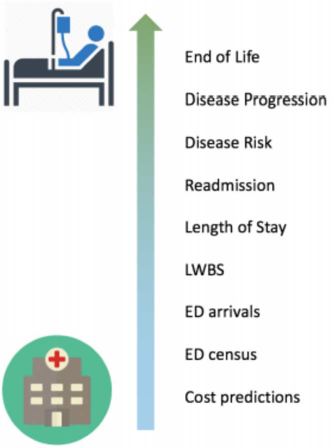


Fig. 1: Prediction Use Cases vs. Need for Interpretability (LWBS: left without being seen)

Ahmad, M. et al. "Interpretable Machine Learning in Healthcare." ACM-BCB 2018.

Activity: Interpretability

Is it interpretable?

- Food

Is it interpretable?

- Food

    Grocery store

Is it interpretable?

- Food

    Grocery store

    Restaurant

Is it interpretable?

- Food
  - Grocery store
  - Restaurant
- Travel

Is it interpretable?

- Food
  - Grocery store
  - Restaurant
- Travel
  - Airlines

Is it interpretable?

- Food
  - Grocery store
  - Restaurant
- Travel
  - Airlines
  - Google maps

Is it interpretable?

- Food
  - Grocery store
  - Restaurant
- Travel
  - Airlines
  - Google maps

- Learning

Is it interpretable?

- Food
  - Grocery store
  - Restaurant
- Travel
  - Airlines
  - Google maps
- Learning
  - In the classroom

Is it interpretable?

- Food
  - Grocery store
  - Restaurant
- Travel
  - Airlines
  - Google maps
- Learning
  - In the classroom
  - Learning by doing

Is it interpretable?

- Food
  - Grocery store
  - Restaurant
- Travel
  - Airlines
  - Google maps
- Learning
  - In the classroom
  - Learning by doing
- Law

Is it interpretable?

- Food
  - Grocery store
  - Restaurant
- Travel
  - Airlines
  - Google maps
- Learning
  - In the classroom
  - Learning by doing
- Law
- Taxes