Explainability AI: Introduction

Marcos M. Raimundo
EMAp - Fundação Getúlio Vargas

Summer School on Data Science
February 4th, 2020 - Rio de Janeiro - Brazil

Motivation

What is the motivation to study interpretability?

For me (scientists): Ways to have insights about what is being learned (curiosity, need new insights).

In general: With the increase in high-stakes decisions (e.g., credit and justice systems), it raises a lot of questions such as fairness, trust, and robustness.

Regulation laws: The right to have an explanation (law enforcement, might incur in penalty).

Material based on the course CS282BR: Topics in Machine Learning Interpretability and Explainability at Harvard. Lectured by Hima Lakkaraju and Ike Lage. `canvas.harvard.edu/courses/68154`

When do we need interpretability?

”Ad servers, postal code sorting, air craft collision avoidance systems—all compute their output without human intervention. Explanation is not necessary either because (1) there are no significant consequences for unacceptable results or (2) the problem is sufficiently well-studied and validated in real applications that we trust the system's decision, even if the system is not perfect.” [Doshi-Velez and Kim, 2017]

”The demand for interpretability arises when there is a mismatch between the formal objectives of supervised learning (test set predictive performance) and the real world costs in a deployment setting.” [Lipton, 2018]

”We argue that the need for interpretability stems from an incompleteness in the problem formalization, creating a fundamental barrier to optimization and evaluation.” [Doshi-Velez and Kim, 2017]

Why do we need interpretability?

Hard to measure and quantify properties – often subjective.

- Trust - A person might feel at ease with a well-understood model, even if this understanding has no purpose.
- Causality - Researchers hope to infer properties (beyond correlational associations) from interpretations/explanations.
- Informativeness/Scientific Knowledge - understanding the characteristics of a large dataset.

Why do we need interpretability?

- Fair and ethical decision making - Guard against certain kinds of discrimination which are too abstract to be encoded. No idea about the nature of discrimination beforehand. How can we be sure algorithms do not discriminate based on race?
- Privacy - The model might reveal individual information.
- Mismatched objectives - Often, we only have access to proxy functions of the ultimate goals
- Multi-objective trade-offs - Competing objectives - Even if the objectives are fully specified, trade-offs are unknown, decisions have to be case by case.

Why do we need interpretability?

- Reliability/robustness/safety - End to end system is never completely testable.
- Transferability/Training and deployment objectives diverge - Humans exhibit a richer capacity to generalize, transferring learned skills to unfamiliar situations
- Environment might even be adversarial - Changing pixels in an image tactically could throw off models but not humans

Properties

Properties of a interpretable model

- Transparency - How exactly does the model work? Details about its inner workings, parameters, etc.
- Post-hoc explanations - What else can the model tell me? Eg., visualizations of learned model, explaining by example

The explanation for our actions/decisions relies on a transparent or on a post-hoc explanation?

Properties of an interpretable model - Transparency

The capability of understanding the model itself.

- Simulatability - Is a user capable of understanding the model to calculate its prediction to a given sample? (e.g., Sparse linear models, Rule lists, Decision trees)
- Decomposability - Is a user capable of understanding each part of the model? (each node of a tree, the weight of each linear parameter).
- Algorithmic Transparency - Is a user capable of understanding, trusting, or predict the behavior of the learning algorithm? (e.g., quadratic optimization in SVM vs. heuristical gradient in neural networks)

Properties of an interpretable model - Post-hoc

The capability to explain the behavior of the model with other, post-hoc, processing of the learning process.

- Textual explanation - Learn a textual explanation (given by humans) of the predictions (humans do that, often after the decision making).
- Visualization - Usage of visualization tools to see predictions similar to the studied ones generate perturbations to observe the outcome.
- Local Explanations - Create explanations near to the studied sample to explain the prediction.
- Example Explanations - Give examples of ground truth samples to explain the predictions.

Evaluating

Evaluating explainability [Doshi-Velez and Kim, 2017]

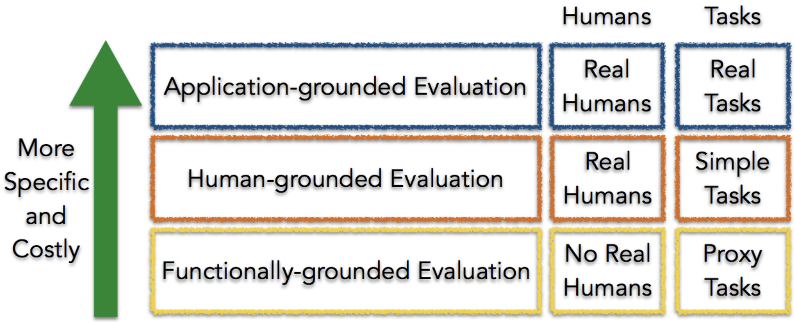


Figura 1: Taxonomy to evaluation of explainable systems [Doshi-Velez and Kim, 2017].

Application-grounded evaluation

Real humans (domain experts), real tasks.

Can be tested in real applications with help of a domain expert.

Typical evaluation in HCI and visualization communities.

Human-grounded evaluation

Real humans, simplified tasks.

The evaluation can be done by real, lay, humans.

It evaluates more general notions of explainability.

Potential experiments:

- Pairwise comparisons.
- Simulate the model output.
- What changes should be made to input to change the output?

Functionally-grounded evaluation

No humans, proxy tasks.

Appropriate for a class of models already validated. Eg., decision trees, sparse linear models.

We can do this when a method is not yet mature, or human subject experiments are unethical.

Potential experiments:

- Complexity (of a decision tree) compared to other other models of the same (similar) class.
- How many levels? How many rules? How many weights.

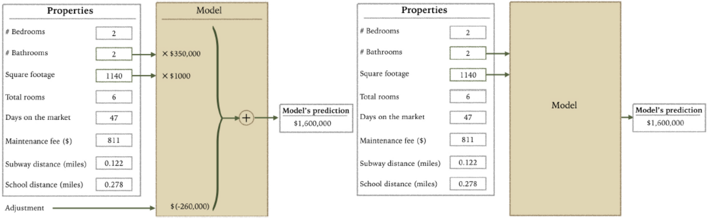
Evaluating - Experiments

How can we evaluate interpretability? Experiments!

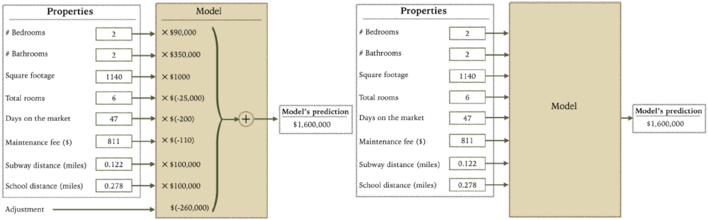
Experiments evaluating the quality of human simulating, trusting, and detecting on mistakes is a linear regressor [Poursabzi-Sangdeh et al., 2018].

Experiment evaluating the impact of the number of Lines, Terms, Cognitive Chunks, and Repetitions in Response Time, Accuracy, and Subjective Difficult (attested by the user) [Lage et al., 2017].

Linear models [Poursabzi-Sangdeh et al., 2018]



(a) Clear, two-feature condition (CLEAR-2). (b) Black-box, two-feature condition (BB-2).



(c) Clear, eight-feature condition (CLEAR-8). (d) Black-box, eight-feature condition (BB-8).

Figura 2: Illustration of the system to evaluate explainability in linear systems [Poursabzi-Sangdeh et al., 2018].

Linear models [Poursabzi-Sangdeh et al., 2018]

Factors studied: number of features and the transparency.

Evaluation:

- Capability of simulating the model.
- Trusting the model – how much the prediction deviates when the model's response is presented.
- Detection of mistakes – how capable was the user of adjusting the prediction in extreme cases.

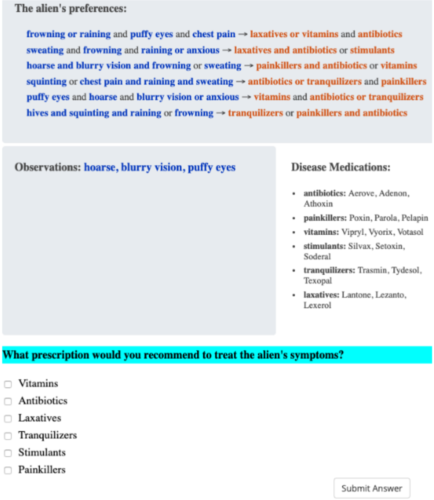
Factors and tasks were chosen based on the literature.

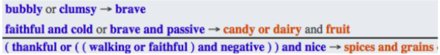
Linear models [Poursabzi-Sangdeh et al., 2018]

Findings:

- Sparse, transparent models are better to simulate the outcome. And Dense, transparent models are worse than dense black-box models.
- No significant difference in trusting the model.
- Clear models are worse to detect mistakes.
- User's prediction errors have no significant difference.

Rule sets [Lage et al., 2017]





(a) Overall representation of the system. **(b)** Representation of the explicit (top) and an implicit (bottom) cognitive chunk.

Rule sets [Lage et al., 2017]

Study the impact of complexity in the properties of the model.

Evaluated complexity:

- Lines.
- Terms.
- Cognitive Chunks.
- Repretitions.

Properties:

- Response Time.
- Accuracy.
- Subjective Difficult (attested by the user).

Rule sets [Lage et al., 2017]

Findings:

"Greater complexity results in longer response times, with the most marked effects for cognitive chunks, followed by model size, then number of variable repetitions."

"Consistency across metrics: subjective difficulty of use follows response time, less clear trends in accuracy."

Creating explainability

Rule Based Approaches [Lakkaraju et al., 2016]

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

Else if Risk-Depression=Yes **then** Depression

Else if BMI $\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

Else if Headaches=Yes **and** Dizziness=Yes, **then** Depression

Else if Doctor-Visits$\geq$ 0.3 **then** Diabetes

Else if Disposition-Tiredness=Yes **then** Depression

Else Diabetes

(a) Original

If Respiratory-Illness=Yes **and** Smoker=Yes **and** Age$\geq$ 50 **then** Lung Cancer

If Risk-LungCancer=Yes **and** Blood-Pressure$\geq$ 0.3 **then** Lung Cancer

If Risk-Depression=Yes **and** Past-Depression=Yes **then** Depression

If BMI$\geq$ 0.3 **and** Insurance=None **and** Blood-Pressure$\geq$ 0.2 **then** Depression

If Smoker=Yes **and** BMI$\geq$ 0.2 **and** Age$\geq$ 60 **then** Diabetes

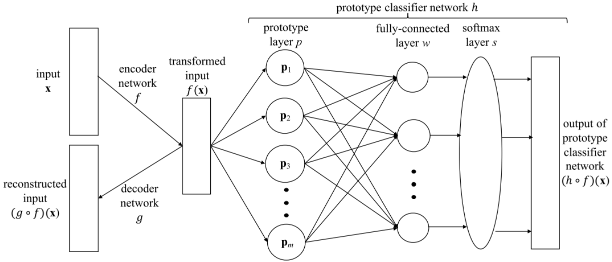
If Risk-Diabetes=Yes **and** BMI$\geq$ 0.4 **and** Prob-Infections$\geq$ 0.2 **then** Diabetes

If Doctor-Visits $\geq$ 0.4 **and** Childhood-Obesity=Yes **then** Diabetes

(b) Interpretable

Figura 5: Represetation of the same rule set depicted with two distinct approaches.

Prototype Based Approaches [Li et al., 2018]



(a) Representation of the neural network

(b) Original samples

(c) Prototypes

Figura 6: Representation of prototype based interpretation.

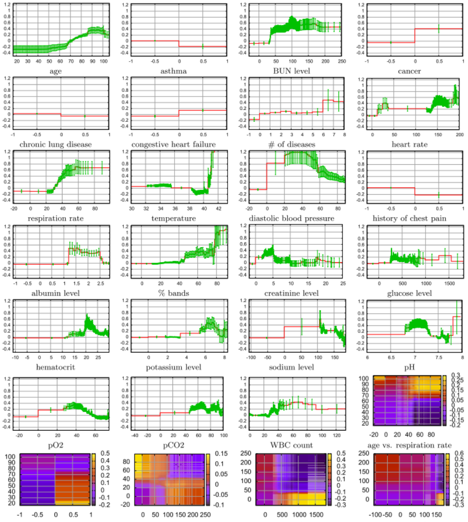
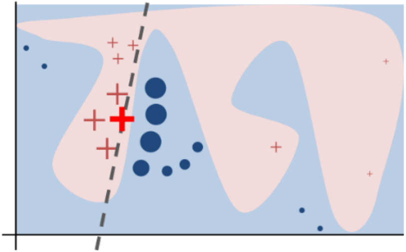
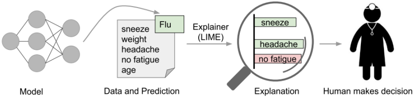


Figura 7: Outcome represented by line graphs for single features, and heat maps for pairwise interaction terms.

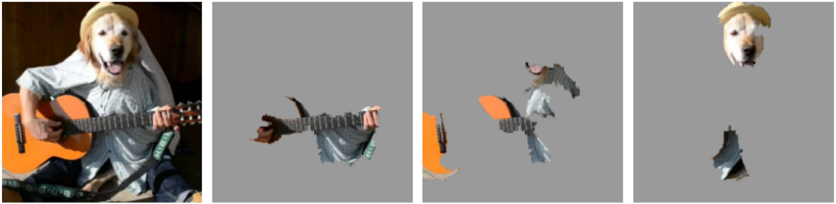
Explaining Black-Box Models [Ribeiro et al., 2016]



(a) Local explanation



(b) System behavior



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

(c) Example

Visualizing Model Behavior

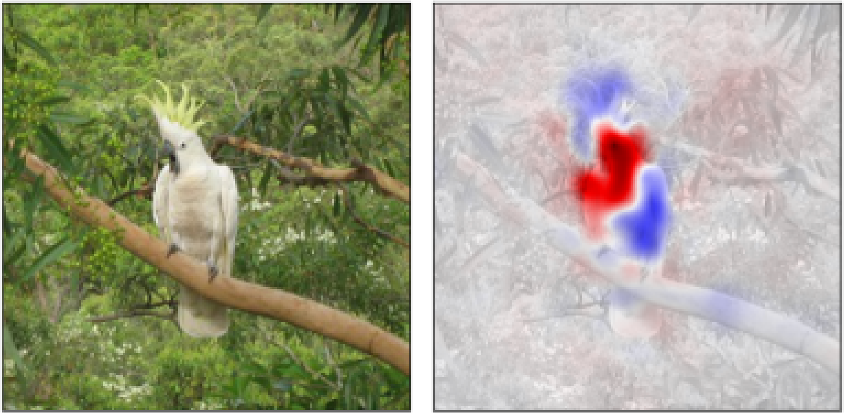


Figura 8: Example of our visualization method: explains why the DCNN (GoogLeNet) predicts ”cockatoo”. Shown is the evidence for (red) and against (blue) the prediction.

Feature Importance Based Explanations [Kim et al., 2018]

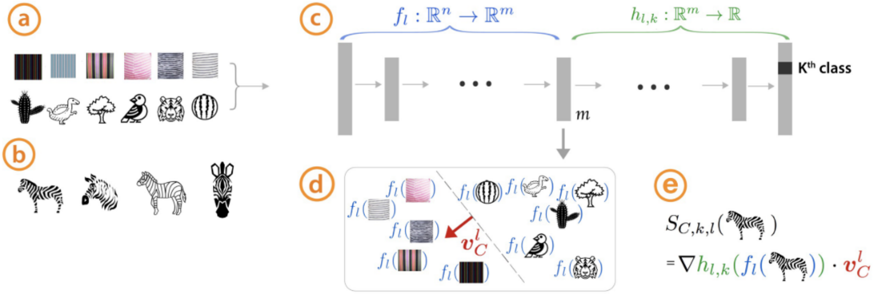
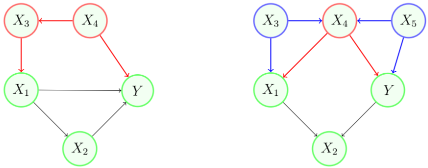


Figura 9: Quantitative Testing with Concept Activation Vectors: (a) concept vs random samples, (b) studied class (zebras), (c) trained network, (d) linear classifier, (e) evaluation.

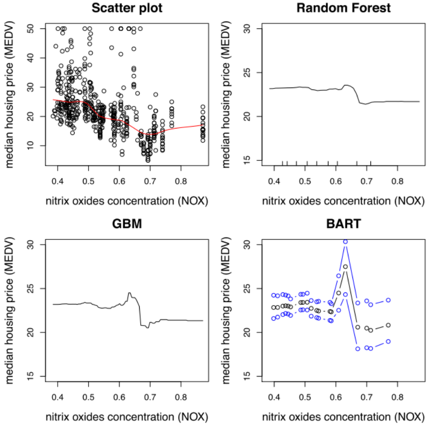
Actionable Explanations (Recourse) [Ustun et al., 2019]

Features to Change	Current Values		Required Values
n_credit_cards	5	⟶	3
current_debt	$3,250	⟶	$1,000
has_savings_account	FALSE	⟶	TRUE
has_retirement_account	FALSE	⟶	TRUE

Figura 10: Changes in a sample feature set that change the outcome.



(a) Representation of causal networks



(b) Graph for marginalization for the impact of a feature in the outcome

Figura 11: Representation of causal interpretations.

Conclusion

Open Problems: Design Issues

What proxies are best for real-world applications?

What factors to consider when designing simpler tasks in place of real-world tasks?

Claims about interpretability must be qualified

If a model satisfies a form of transparency, highlight that clearly.

For post-hoc interpretability, fix a clear objective and demonstrate evidence.

Transparency may be at odds with broader objectives of AI

Choosing interpretable models over accurate ones to convince decision makers .

Short term goal of building trust with doctors might clash with long term goal of improving health care.

Post-hoc interpretations can mislead

Do not blindly embrace post-hoc explanations!

Post-hoc explanations can seem plausible but be misleading.

They do not claim to open up the black-box.

They only provide plausible explanations for its behavior. Eg., text explanations.

Explainability AI: Introduction

Marcos M. Raimundo
EMAp - Fundação Getúlio Vargas

Summer School on Data Science
February 4th, 2020 - Rio de Janeiro - Brazil

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., and Elhadad, N. (2015).
Intelligible Models for HealthCare.
pages 1721–1730.

Doshi-Velez, F. and Kim, B. (2017).
Towards A Rigorous Science of Interpretable Machine Learning.
(MI):1–13.

References ii

Kim, B., Wattenberg, M., Gilmer, J., Cai, C., Wexler, J., Viegas, F., and Sayres, R. (2018).
Interpretability beyond feature attribution: Quantitative Testing with Concept Activation Vectors (TCAV).
35th International Conference on Machine Learning, ICML 2018, 6:4186–4195.

Lage, I., Chen, E., He, J., Narayanan, M., Kim, B., Gershman, S. J., and Doshi-velez, F. (2017).
Human Evaluation of Models Built for Interpretability.

References iii

- Lakkaraju, H., Bach, S. H., and Leskovec, J. (2016).
Interpretable decision sets: A joint framework for description and prediction.
Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu:1675–1684.

- Li, O., Liu, H., Chen, C., and Rudin, C. (2018).
Deep learning for case-based reasoning through prototypes: A neural network that explains its predictions.
32nd AAAI Conference on Artificial Intelligence, AAAI 2018, pages 3530–3537.

Lipton, Z. C. (2018).
The mythos of model interpretability.
Communications of the ACM, 61(10):35–43.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Vaughan, J. W., and Wallach, H. (2018).
Manipulating and Measuring Model Interpretability.
pages 1–20.

Ribeiro, M. T., Singh, S., and Guestrin, C. (2016).
"Why should i trust you?"Explaining the predictions of any classifier.
Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 13-17-Augu:1135–1144.

Ustun, B., Spangher, A., and Liu, Y. (2019).
Actionable recourse in linear classification.
FAT 2019 - Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency*, pages 10–19.

Zhao, Q. and Hastie, T. (2019).
Causal Interpretations of Black-Box Models.
Journal of Business and Economic Statistics.