In Algorithms We Trust

Interpretability, robustness and bias in machine learning

Louis Abraham

March 21th 2019

ACPR

A word about trust in decision making

About me

Louis Abraham

- Education: École polytechnique, ETH Zurich
- Experience:
  - Quant @ BNP Paribas
  - Deep learning @ EHESS / ENS Ulm
  - Data protection @ Qwant Care

What this talk is about

- Machine Learning
- Supervised learning
- Practical tools
- Humans

What this talk is *not* about

- Mathematics
- Deep Learning
- AI Safety
- Fairness in AI

Bias vs bias

- Oxford dictionary: *Inclination or prejudice for or against one person or group, especially in a way considered to be unfair.*
- Wikipedia: *In statistics, the bias (or bias function) of an estimator is the difference between this estimator's expected value and the true value of the parameter being estimated.*

Is this bias?

SELF-DRIVING CARS MORE LIKELY TO DRIVE INTO BLACK PEOPLE, STUDY CLAIMS

New study suggests autonomous vehicles might be racist

Anthony Cuthbertson | @ADCuthbertson |
Wednesday 6 March 2019 13:58 | |

Click to follow
The Independent Tech

Technology used in self-driving cars has a racial bias that makes autonomous vehicles more likely to drive into black people, a new study claims.

Researchers at the Georgia Institute of Technology found that state-of-the-art detection systems, such as the sensors and cameras used in self-driving cars, are better at detecting people with lighter skin tones.

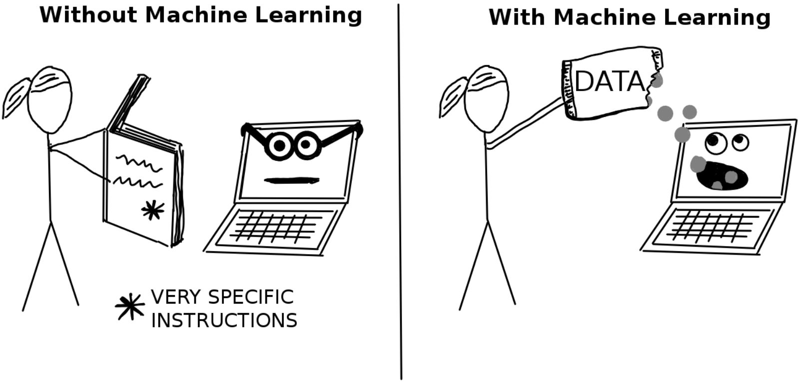
That makes them less likely to spot black people and to stop before crashing into them, the authors note.

source: The Independent

What bias really is

https://www.youtube.com/embed/lfpjXcawG60?rel=0

The difference between programming and ML



credits: Christoph Molnar

How developers explain their programs

credits: CommitStrip

How data scientists explain their programs

credits: xkcd

Do we need interpretability?

Interpretability is useful for:

- **Compliance**: *Right to explanation* in the GDPR (Goodman and Flaxman 2017; Wachter, Mittelstadt, and Russell 2017)
- Privacy
- Fairness
- Robustness
- Trust

Risks of interpretability

- Corporate secrecy
- Performance drop
- Manipulation
- Public relations

Different concepts

Quick survey

One will protect you, the other 2 will try to kill you.
Choose wisely.

- Interpretability
- Explainability
- Justifiability

Definition

(Biran and Cotton 2017)

> *Explanation is closely related to the concept of interpretability: systems are interpretable if their operations can be **understood** by a **human**, either through **introspection** or through a **produced explanation**.*
>
> *In the case of machine learning models, **explanation is often a difficult task** since most models are not **readily** interpretable.*

Different concepts

Quick survey

One will protect you, the other 2 will try to kill you. Choose wisely.

- **Interpretability**: *why* did the model do that
- Explainability: *how* the model works
- ~~Justifiability~~: justice, morals

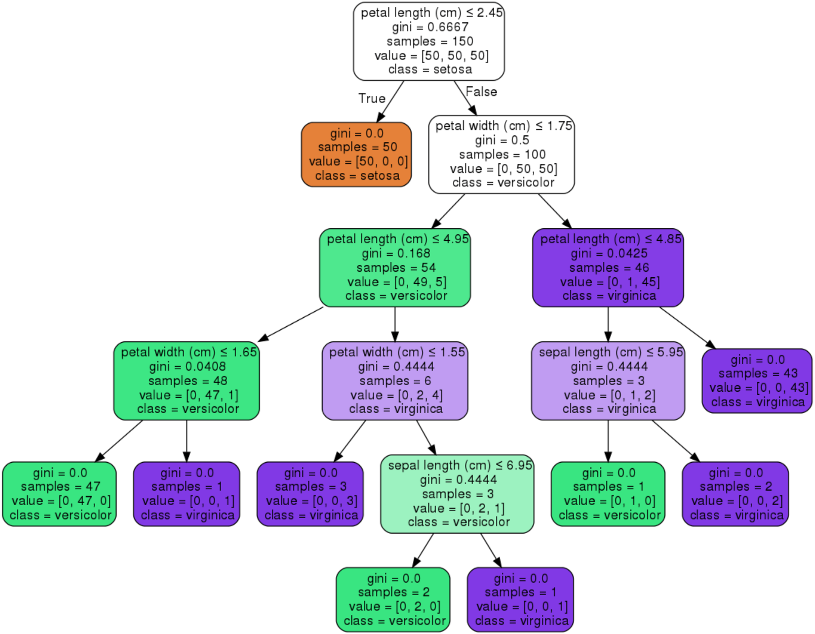
Interpretability of the whole process

- model selection
- training
- evaluation

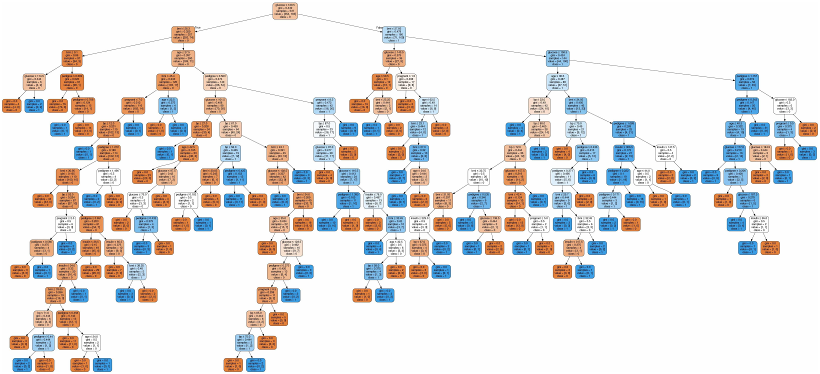
3 options:

- readily interpretable models
- feature importance
- example based explanations

Is this an interpretable model?



Is this an interpretable model?



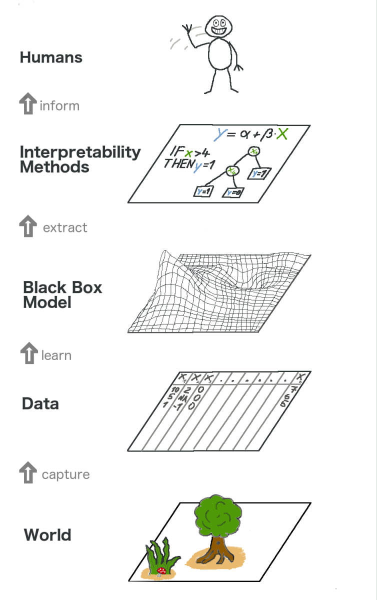
Interpretable models

- *sparse* or *low-dimensional* linear models (regression, logistic regression, SVM)
- *small* decision trees (~~forests~~)
- decision rules, for example *falling rule lists* (Wang and Rudin 2015)
- naive Bayes classifier
- k-nearest neighbors

Make them more powerful!

- preprocessing / normalization
- feature engineering

Model agnostic methods



credits: Christoph Molnar

Model agnostic methods

Why you want model-agnostic methods

(Ribeiro, Singh, and Guestrin 2016a)

- Use more powerful models
- Produce better explanations
- Representation flexibility
- Lower cost to switch models
- Explanation coherence
- Compare models and explanations independently

The 10 best model-agnostic methods

1. plots
2. plots
3. plots
4. plots
5. plots
6. plots
7. plots
8. Counterfactual explanations (Wachter, Mittelstadt, and Russell 2017)
9. LIME (Ribeiro, Singh, and Guestrin 2016b)
10. Shapley Values (Lundberg and Lee 2017)

Counterfactual explanations

(Wachter, Mittelstadt, and Russell 2017)

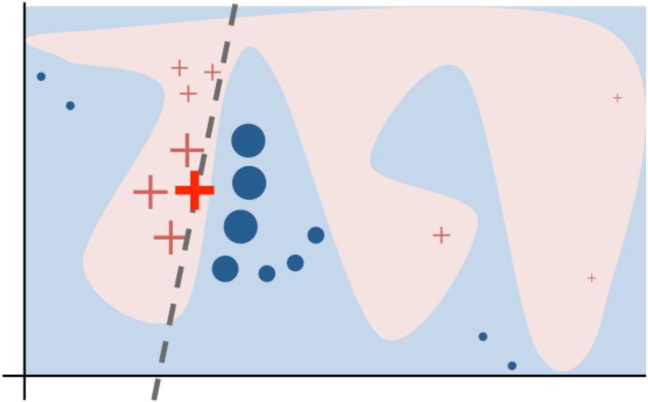
$$\arg\min_{x'} \max_{\lambda} \lambda \cdot (\hat{f}(x') - y')^2 + d(x, x')$$

- simply: find a neighbor with a different prediction
- is this useful?
- preserves secrecy
- related to adversarial examples

LIME (Local Interpretable Model-agnostic Explanations)

(Ribeiro, Singh, and Guestrin 2016b)

- given a point x, trains surrogate model g on neighbors
- $\xi(x) = \arg\min_{g \in G} L(f, g, \pi_x) + \Omega(g)$
- complete framework: categorical data, text, images...
- open-source Python library

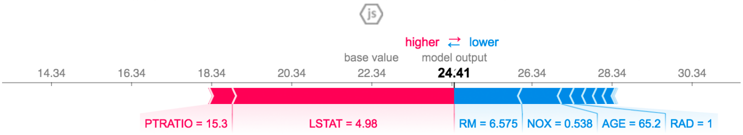


SHAP (SHapley Additive exPlanations)

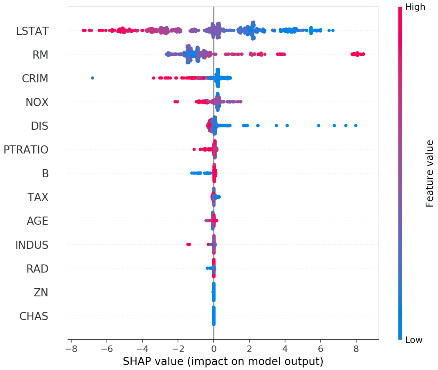
(Lundberg and Lee 2017)

- find feature importance by ablation
- generalizes LIME, Quantitative Input Influence and others
- relies on economic theory and is consistent with humans
- open-source Python library

SHAP (SHapley Additive exPlanations)



Explanation of one instance



Summary over the dataset

Evaluation of interpretability

(Doshi-Velez and Kim 2017)

- Application-grounded Evaluation: Real humans, real tasks
- Human-grounded Metrics: Real humans, simplified tasks
- Functionally-grounded Evaluation: No humans, proxy tasks

The beginning...

References I

Biran, Or, and Courtenay Cotton. 2017. "Explanation and Justification in Machine Learning: A Survey." In *IJCAI-17 Workshop on Explainable Ai (Xai)*. Vol. 8.

Burrell, Jenna. 2016. "How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms." *Big Data & Society* 3 (1). SAGE Publications Sage UK: London, England: 2053951715622512.

Datta, Anupam, Shayak Sen, and Yair Zick. 2016. "Algorithmic Transparency via Quantitative Input Influence: Theory and Experiments with Learning Systems." In *2016 Ieee Symposium on Security and Privacy (Sp)*, 598–617. IEEE.

Doshi-Velez, Finale, and Been Kim. 2017. "Towards a Rigorous Science of Interpretable Machine Learning." *arXiv Preprint arXiv:1702.08608*.

References II

Goodman, Bryce, and Seth Flaxman. 2017. "European Union Regulations on Algorithmic Decision-Making and a 'Right to Explanation'." *AI Magazine* 38 (3): 50–57.

Guidotti, Riccardo, Anna Monreale, Salvatore Ruggieri, Franco Turini, Fosca Giannotti, and Dino Pedreschi. 2018. "A Survey of Methods for Explaining Black Box Models." *ACM Computing Surveys (CSUR)* 51 (5). ACM: 93.

Lipton, Zachary C. 2016. "The Mythos of Model Interpretability." *arXiv Preprint arXiv:1606.03490*.

Lundberg, Scott M, and Su-In Lee. 2017. "A Unified Approach to Interpreting Model Predictions." In *Advances in Neural Information Processing Systems*, 4765–74.

Miller, Tim. 2018. "Explanation in Artificial Intelligence: Insights from the Social Sciences." *Artificial Intelligence*. Elsevier.

References III

Molnar, Christoph. 2019. *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. https://christophm.github.io/interpretable-ml-book/.

Ribeiro, Marco Tulio, Sameer Singh, and Carlos Guestrin. 2016a. “Model-Agnostic Interpretability of Machine Learning.” *arXiv Preprint arXiv:1606.05386*.

———. 2016b. “Why Should I Trust You?: Explaining the Predictions of Any Classifier.” In *Proceedings of the 22nd Acm Sigkdd International Conference on Knowledge Discovery and Data Mining*, 1135–44. ACM.

Vellido, Alfredo, José David Martín-Guerrero, and Paulo JG Lisboa. 2012. “Making Machine Learning Models Interpretable.” In *ESANN*, 12:163–72. Citeseer.

References IV

Wachter, Sandra, Brent Mittelstadt, and Chris Russell. 2017. “Counterfactual Explanations Without Opening the Black Box: Automated Decisions and the Gdpr.” *Harvard Journal of Law & Technology* 31 (2): 2018.

Wang, Fulton, and Cynthia Rudin. 2015. “Falling Rule Lists.” In *Artificial Intelligence and Statistics*, 1013–22.