Interpretability, Explainability, and Fairness in Machine Learning Models

Gonzalo Rivero

September, 24 2020

Introduction

Goals for today

- Discuss general concepts with Rudin (2019) as starting point
- Explore two/two-and-a-half core ideas:
  1. Interpretability
  2. Explainability
  3. Fairness (+ causality!)
- Question: **Should we use ML for high-stakes decisions?**
- Follow-up: **How to make sure it is not doing a bad thing?**
- Goal is to raise awareness of this literature.
- No technical details or how-to.

Interpretability and explainability

Context

- Two standard types of cases have brought back the literature:

1. COMPAS:
   - A “decision support tool” used to predict recidivism.
   - Scores can be considered by judges during sentences.
   - Proprietary software which cannot be examined by the public.
   - Evaluation by ProPublica showed that it depended on race.
   - “Blacks are almost twice as likely to be labelled as higher risk”
2. Algorithmic hiring:
   - ML models through all the pipeline
     - Who sees the ads
     - Expected performance of the applicant
     - Which applicants will receive more screening
     - Forecast salary
   - Evidence of reproduction of human biases

- “Right to explanation” in the GDPR

Interpretability vs. explainability

- The *tentative* definitions are:
  - “High stakes” $\rightarrow$ “Impact on society”
  - “The bad thing” $\rightarrow$ “automation of discrimination”
- Problem is that models are “black boxes”
- This suggests two/three approaches:
  1. Simplify (*constraint*) the model to make it easier to *interpret*
  2. Build tools to make a complicated models easier to *explain*
  3. (Maybe) develop criteria to assess the model predictions
- Rudin (2019) suggests that 1. is strictly superior *and feasible*

Why explainability

- We want intepretability because it:
  1. Engenders trust
  2. May uncover causal relations
  3. It addresses the right to explanation
- An interpretable model is
  1. Open (as opposed to proprietary)
  2. Intelligible
- Not what ML optimizes for!
  - Performance metrics do not sufficiently characterize the model
- Alternative is to do post-hoc interpretation

Why not explainability?

1. There is no trade-off between accuracy and interpretability
   - At least not with structured data and meaningful features
   - Lasso performs as well as RF/GBM in many domains
   - Interpretability allows us to improve feature construction
2. Explanations are not faithful to the model
   - By construction (otherwise, the explanation is the model)
   - No guarantee that the explanation is correct
     - Should we trust the explanation or the model?
   - Does COMPAS really base the predictions on race?
3. Ripe for procedural error
   - Models are hard to troubleshoot

Why not interpretable models?

- Why haven't we seen more interpretable models?

1. Companies cannot benefit from them
   - Transparency is a property of them *and the ownership*
   - Disconnect between profits and responsibility for predictions
2. Interpretable models are harder
   - Unlike off-the-shelf ML approach
   - May require domain expertise
3. Black box lead to discovery
   - Reverse direction from data to theory

Why are interpretable models harder?

- We associate interpretability with
  1. Linear models
     - Weighted combinations are nice
     - Integer combinations are even nicer
     - Very easy as scoring
     - NP-complete
  2. If-then rules
     - Current trees use greedy heuristics
     - We would like globally optimal trees
     - That also minimize complexity (f.i. leaves)
     - NP-complete
  3. Case-based reasoning
     - Interpretability is domain specific
     - Not clear what it means in general
     - We may have different heuristics for different problems
     - Interpretability is like performance

But what is interpretability?

- We could have several goals all of them challenging.
  - Trust. We want models that we can trust. But is that...
    - ... confidence in performance? How often is right
    - ... willingness to relinquish control? When is it right
  - Causality. We want relations to have causal meaning
    - Impossible without a theoretical model
  - Transferability. Models should not depend on the environment
    - Possibility of gaming
    - Limited generalizability beyond train distribution
  - Informativeness
    - Provide insights to decision makers
    - Learn the structure of the data
    - Isn't a explainable model better?
  - Fair and ethical decision making
    - Not the metric we use

Is interpretability undesirable?

- Let's consider the properties of interpretable models
  - Simulatability
    - The model can be contemplated at once (i.e., simplicity)
    - "Lasso is more interpretable than an NN"
    - Is that about model size or computation for inference?
    - A subjective statement about limits of cognition
    - How many dimensions before a tree is not interpretable?
  - Decomposability
    - Input, parameters and calculations are interpretable
    - Manual feature is more interpretable than automatic ones
    - But they are much less robust
  - Transparency
    - Are models less transparent than humans?
    - Any model is replicable
- Interpretability is a subjective goal

Fairness

Fairness in Machine Learning

- Ignore the model, think about the predictions
  - The predictions should be "fair"
  - Standard question since the 1960s in education research
- Related to the notion of non-discrimination
  - Procedural fairness
  - Outcome equalization
  - These two principles may enter in conflict
  - See Ricci vs. DeStefano
- "No fairness through unawareness"
  - Removing the protected attribute is not enough
  - Formal vs. intentional disparate treatment
  - Ignore biases in data vs. Induce biases in data

Advantages of fairness as a concept

- Suitable for a operational definition
  - Covariates/features
  - A protected attribute
  - A decision rule
- Suitable for guiding model corrections
  - In pre-processing (uncorrelate feature space)
  - In training (customize loss function)
  - In post-processing (adjust the predictions)

But what is fairness?

Standard formal criteria (attribute A, score R, target Y)

- Independence: $R \perp A$
  - Acceptance rate should be equal across groups
  - Condition can be met without fairness
  - Accept at same rate but use different procedures
  - Easy to satisfy and verify
- Separation: $R \perp A|Y$
  - Correlation between R and A is justified by Y
  - Equalization of FPR and FNR across groups
  - $FPR = \Pr\{R = 1|Y = 1\}; FNR = \Pr\{R = 1|Y = 0\}$
  - You can choose which one to relax
  - Dealt with in ROC
- Sufficiency: $Y \perp A|R$
  - Parity of positive/negative predicted values
  - The score is calibrated *by group*
  - Usually does not require intervention

Challenges

- Impossibility theorems
  - No two criteria can be simultaneously satisfied
  - COMPAS debate between Northpointe and ProPublica
    - ProPublica: COMPAS violates separation
    - Northpointe: COMPAS satisfies sufficiency
  - Guidance cannot come only from data
- Observability and inference
  - All the criteria are observational (no *what-if*)
  - Depends on a causal graph
  - Can build identical joint distributions with different fairness
  - Observational definition cannot distinguish them
  - Answers cannot depend only on observational data

Conclusions

Conclusions

- “New” fields that have grown considerably
  - Complicated to navigate
  - Sometimes very technical and inaccessible
  - Disparate languages across fields
- Many new available tools
  - Will discuss some in the third session
- Clear business impact
  - Increasing application of ML for social outcomes
  - Performing ML or auditing ML
  - Leverage the company’s SME
- More practical motivation
  - What is your model capturing?
  - Does it do what you think it does?
  - Intepretation/explanation matters for validity

References

- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence, 1(5), 206-215.
- Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012, January). Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (pp. 214-226).
- Chouldechova, A. (2017). Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big data, 5(2), 153-163.
- Lipton, Z. C. (2018). The mythos of model interpretability. Queue, 16(3), 31-57.
- Kleinberg, J., Mullainathan, S., & Raghavan, M. (2016). Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807.