Responsible Machine Learning*

Lecture 3: Fairness

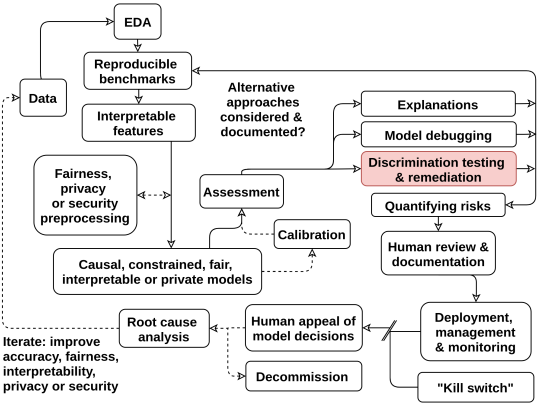
Patrick Hall

The George Washington University

June 2, 2020

Contents

A Responsible Machine Learning Workflow†



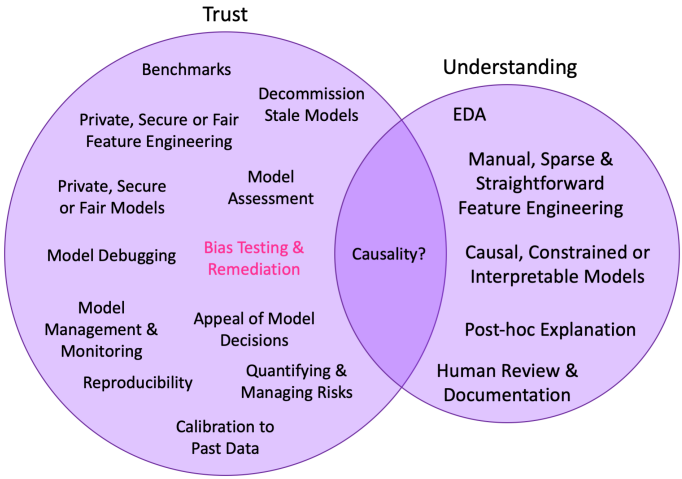
† *A Responsible Machine Learning Workflow*

Why Care About Discrimination in ML?

- **Reputational risk**:
  - Upon encountering a perceived unethical ML system, 34% of consumers are likely to, "stop interacting with the company."[‡]
- Non-compliance fines and litigation costs.
- Responsible practice of ML.

[‡]See: Why addressing ethical questions in AI will benefit organizations.

Elements of Responsible Practice of ML



What Is Bias?

- Almost all data, statistical models, and machine learning (ML) models encode different types of *bias*, i.e., systematic misrepresentations of reality.
- Sometimes, bias is helpful.
  - shrunken and robust β_j coefficients in penalized linear models
- Other types of bias might be unwanted, unhelpful, or illegal discrimination.

What is Discrimination in ML?

In some applications[§], model predictions should ***ideally*** be independent of demographic group membership.

In these applications, a model exhibits discrimination if:

1. Demographic group membership is not independent of the likelihood of receiving a favorable or accurate model prediction.
2. Membership in a *subset* of a demographic group is not independent of the likelihood of receiving a favorable or accurate model prediction (i.e., *local bias*).[2]

[§]e.g., Under the Equal Credit Opportunity Act (ECOA), as implemented by Regulation B, and the Fair Credit Reporting Act (FCRA)

What is Discrimination in ML?

Several forms of discrimination may manifest in ML, including:

- Overt discrimination, i.e. *disparate treatment*.
- Unintentional discrimination, i.e. *disparate impact* (DI).
- Discrimination in ML may or may not be illegal, depending on how it arises and applicable discrimination laws.[2]

What is Discrimination in ML?

Discrimination originates from training data:

- Incomplete or inaccurate data, particularly under-representation of minorities, e.g. Gender Shades[1].
- Accurate but differing patterns of causation, correlation, or dependency between demographic groups and past outcomes, e.g. traditional FICO credit scores.¶
- Explicit encoding of historical social biases into training data, e.g. criminal records.¶

¶See: Responsible Data Science: Identifying and Fixing Biased AI.

What is Discrimination in ML?

ML models can perpetuate or exacerbate discrimination.

Group disparities, i.e. different or inaccurate treatment of entire demographic groups:

- Learning different correlations between demographic groups and favorable model outcomes, i.e. *adverse impact*.
- Exhibiting different accuracies across demographic groups, i.e. *differential validity*.¶

Locally, i.e. different or inaccurate treatment of similar individuals:

- Local response function or decision boundary form.
- Capacity to form local complex or latent demographic proxies.

Common Metrics of Discrimination in ML

Common metrics of ***group*** disparities:

- Accuracy disparity: $\frac{\text{accuracy}_p}{\text{accuracy}_r}$
- Adverse impact ratio: $\frac{\%\ \text{accepted}_p}{\%\ \text{accepted}_r}$
- Marginal effect: $\%\ \text{accepted}_p - \%\ \text{accepted}_r$
- Standardized mean difference: $\frac{\bar{\hat{y}}_p - \bar{\hat{y}}_r}{\sigma_{\hat{y}}}$

where, $p \equiv$ protected group and $r \equiv$ reference group (often white males),

$\%\ \text{accepted}_{\text{group}} = 100 \cdot \frac{\mathbf{tn_{group}} + \mathbf{fn_{group}}}{N_{\mathbf{group}}}$, and $\text{accuracy}_{\text{group}} = \frac{\mathbf{tp_{group}} + \mathbf{tn_{group}}}{N_{\mathbf{group}}}$. ¶

Local bias is much trickier to measure and often an unmitigated risk for consumer-facing ML systems.

How to Fix Discrimination in ML?

Fix the process: ensure diversity of experience in design, training, and review of ML systems.

Fix the data:

- Collect demographically representative training data.
- Select features judiciously, e.g. using `time_on_file` as an input variable as opposed to `bankruptcy_flag`.¶
- Sample and reweigh training data to minimize discrimination.[3]

How to Fix Discrimination in ML?

Fix the model:

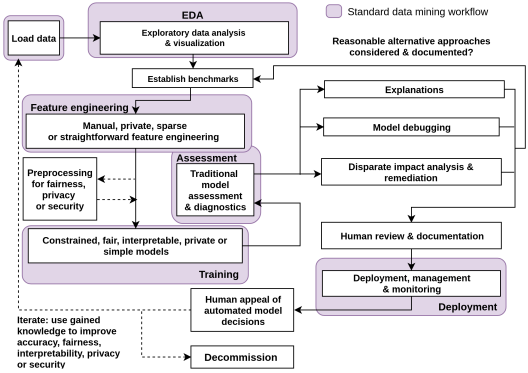
- Consider fairness metrics when selecting hyperparameters and cutoff thresholds.
- Train fair models directly:
  - Learning fair representations (LFR) and adversarial de-biasing.[6], [7]
  - Use dual objective functions that consider both accuracy and fairness metrics.
- Edit model mechanisms to ensure less biased predictions, e.g. with GA2M models.

Fix the predictions:

- Balance model predictions, e.g. reject-option classification.[4]
- Correct or override predictions with model assertions or appeal mechansims.[2], [5]

How to Fix Discrimination in ML?

As part of a responsible ML workflow.



Acknowledgements

This presentation borrows heavily from the expertise of Nicholas Schmidt and Bryce Stephens of BLDS, LLC, a leading fair lending compliance firm.

References

Joy Buolamwini and Timnit Gebru. "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification." In: *Conference on Fairness, Accountability and Transparency*. URL: http://proceedings.mlr.press/v81/buolamwini18a/buolamwini18a.pdf. 2018, pp. 77–91.

Patrick Hall, Navdeep Gill, and Nicholas Schmidt. "Guidelines for the Responsible Use of Explainable Machine Learning." In: *arXiv preprint arXiv:1906.03533* (2019). URL: https://arxiv.org/pdf/1906.03533.pdf.

Faisal Kamiran and Toon Calders. "Data Preprocessing Techniques for Classification Without Discrimination." In: *Knowledge and Information Systems* 33.1 (2012). URL: https://bit.ly/2lH95lQ, pp. 1–33.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. "Decision Theory for Discrimination-aware Classification." In: *2012 IEEE 12th International Conference on Data Mining*. URL: http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.722.3030&rep=rep1&type=pdf. IEEE. 2012, pp. 924–929.

Daniel Kang et al. *Debugging Machine Learning Models via Model Assertions*. URL: https://www-cs.stanford.edu/~matei/papers/2018/mlsys_model_assertions.pdf. 2019.

References

Rich Zemel et al. "Learning Fair Representations." In: *International Conference on Machine Learning*. URL: http://proceedings.mlr.press/v28/zemel13.pdf. 2013, pp. 325–333.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating Unwanted Biases with Adversarial Learning." In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. URL: https://arxiv.org/pdf/1801.07593.pdf. ACM. 2018, pp. 335–340.