Discrimination in Machine Learning[*,†]

© Navdeep Gill

February 28, 2021

[†] This presentation is not, and should not be construed as, legal advice or requirements for regulatory compliance.

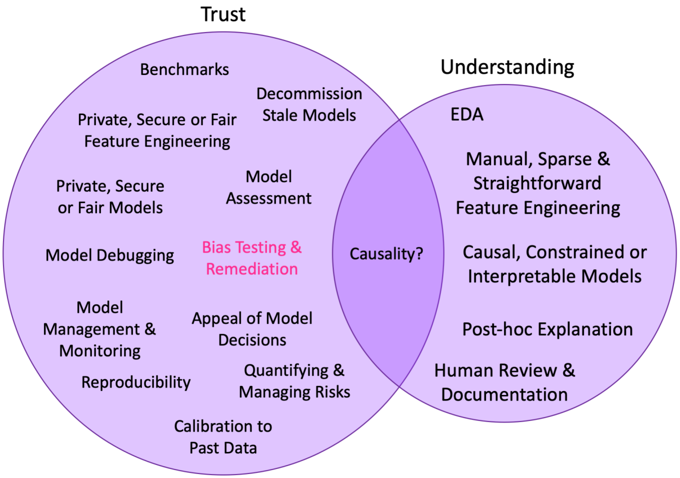
Contents

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."‡
    - Consumers will not differentiate between intentional or unintentional unethical behavior in a ML system ...
    - If trust is lost in a product/service by a consumer, then it will be difficult to obtain again.
- **Responsible practice of ML**:
  - Clear understanding and trust of a ML system's behavior.
    - Analysts, engineers, physicians, researchers, scientists, and humans in general have the need to understand and trust models and modeling results that affect our work and our lives.
  - Rigorous vetting of a ML system, especially when it comes to high stake applications e.g.:
    - Fair lending
    - Credit scoring
    - Facial recognition
    - Healthcare diagnosis
    - Recidivism
- **Non-compliance fines and litigation costs**.

‡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.
    - Data used for models are a sample and hence are an estimation of reality, which will carry some bias on its own.
    - Models have their own sense of reality based on data given to the model and the algorithm itself.
- Sometimes, bias is helpful, e.g. shrunken, robust β_j coefficients in penalized linear models.
    - Bias-Variance trade off
- Other types of bias might be unwanted, unhelpful, or illegal discrimination, e.g. penalizing observations based on race, gender, socioeconomic status, etc.

What is Discrimination in ML?

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

- High stake ML applications

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]

Several forms of discrimination may manifest in ML, including:

- Intentional discrimination, i.e. *disparate treatment*.
  - Using a demographic group variable as an input to a ML model, e.g. race, gender, etc.
- Unintentional discrimination, i.e. *disparate impact* (DI).
  - Using a variable that is a reasonable predictor of the outcome but introduces some demographic group bias, e.g. a variable correlated to a demographic group variable.
  - Most ML models can be checked for this post hoc, e.g., disparate impact analysis (DIA).
- Discrimination in ML may or may not be illegal, depending on how it arises and applicable discrimination laws.[2]
  - Regulated industries might investigate potential discrimination in ML applications more so than unregulated industries. However, this is not always the case.

[§] 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?

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.

What is Discrimination in ML?

Many kinds of group disparities can be measured[‖], e.g.:

- 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),

$\%\text{ accepted}_{\text{group}} = 100 \cdot \frac{\text{tn}_{\text{group}} + \text{fn}_{\text{group}}}{N_{\text{group}}}$, and $\text{accuracy}_{\text{group}} = \frac{\text{tp}_{\text{group}} + \text{tn}_{\text{group}}}{N_{\text{group}}}$. ¶

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

[‖] Python notebook showing different ways to measure group disparities: https://github.com/navdeep-G/interpretable-ml/blob/master/python/jupyter-notebooks/credit/binomial/dia.ipynb

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]

Fix the model:

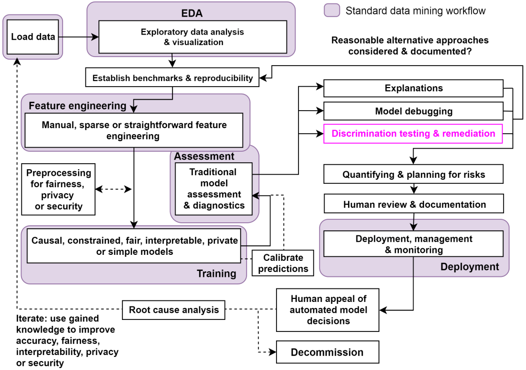
- Consider fairness metrics when selecting hyperparameters and cutoff thresholds.
- Train fair models directly:
  - Learning fair representations (LFR) and adversarial de-biasing.[6], [7]
  - Use dual objective functions that consider both accuracy and fairness metrics.
- Edit model mechanisms to ensure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



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.

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.

References

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.