DLI Accelerated Data Science Teaching Kit

Lecture 4.2 - Tools for Discovering and Interpreting Bias in Models

Discovering and mitigating bias

Outline

- Avoiding Bias
  - Diversity in team composition
  - Understand task, stakeholders, and potential for errors
  - Verify data representation
  - Understand model decision
- Discovering Bias
  - Targeted test
  - Quick test
  - Comprehensive test
  - Ecologically valid test
  - Adversarial test
- Mitigating Bias
  - Adversarial training
  - Correlation loss
  - Constrained optimization
- Tools

Bias Avoidance

Best practice

1. **Diversity** in team composition
2. Understand task, stakeholders, and potential for errors
3. Verify data representation
4. Understand model decision

Bias Avoidance

Best practice

1. Diversity in team composition
2. Understand **task, stakeholders** and potential for **errors**
3. Verify data representation
4. Understand model decision

Bias Avoidance

Best practice

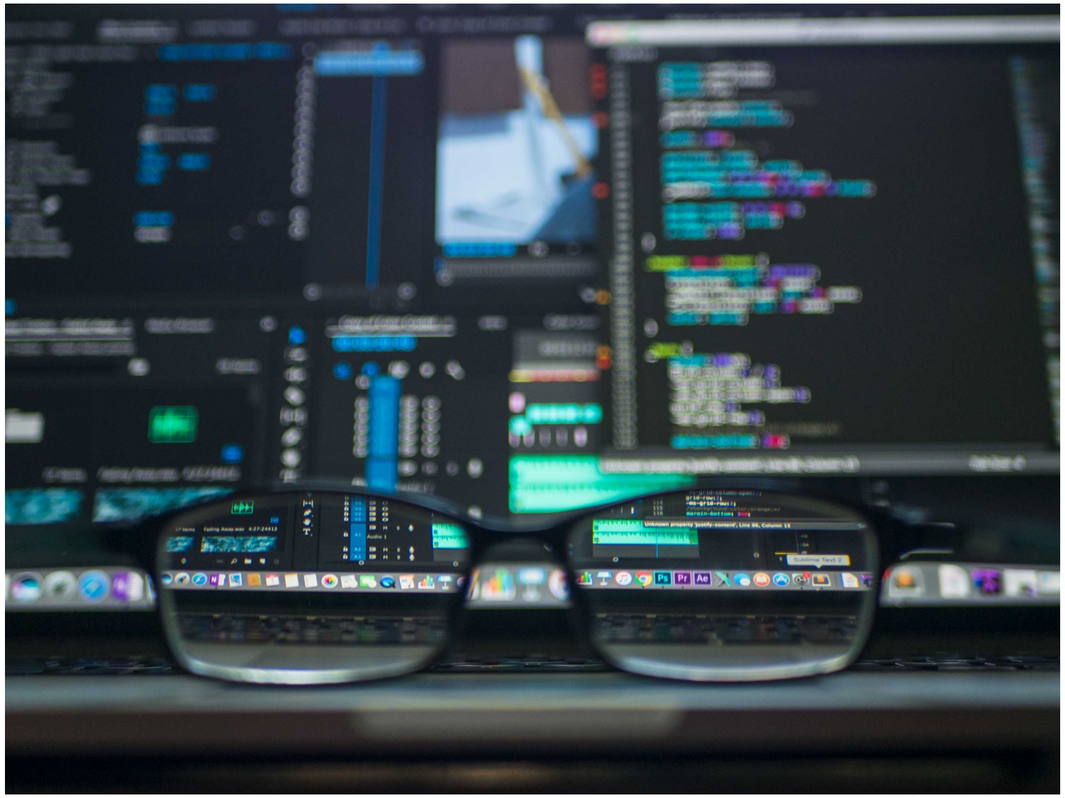
1. Diversity in team composition
2. Understand task, stakeholders, and potential for errors
3. Verify data **representation**
4. Understand model decision

Through experiments, surveys and other methods

Bias Avoidance

Best practice

1. Diversity in team composition
2. Understand task, stakeholders, and potential for errors
3. Verify data representation
4. **Understand** model decision



Use ML explainability methods to understand the model decision making process

Discovering Bias

Targeted test

1. Test models on instances that are likely to be affected by bias based on prior experience/knowledge
2. Test images with different skin colors on facial recognition models
3. Test gender stereotypes on natural language processing models

Discovering Bias

Quick test

1. Check extreme cases
2. Low coverage but quick to spot major issues
3. Consider marginalized groups

Discovering Bias

Comprehensive test

1. Include enough test data for each subgroup
2. Include all relevant combinations of attributes
3. Critical if the model will be used in system that has high social impacts

Discovering Bias

Ecologically valid test

1. Test models on real-world settings
2. Test how model generalize to newer data
3. Useful when historical data is available or there are ways to estimate real data distribution

Discovering Bias

Adversarial test

1. Test for rare but extreme harm
2. For example, test high-risk options in a medical treatment recommender system
3. Requires domain knowledge

Mitigating Bias

Adversarial learning

1. Train a predictor (predict label Y) and an adversary (predict attribute Z) simultaneously
2. Maximize predictor's power to predict Y and minimize adversary's ability to predict Z

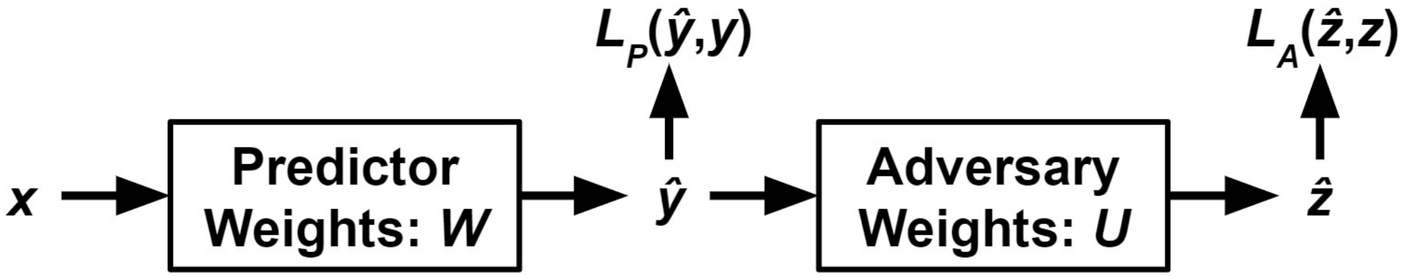
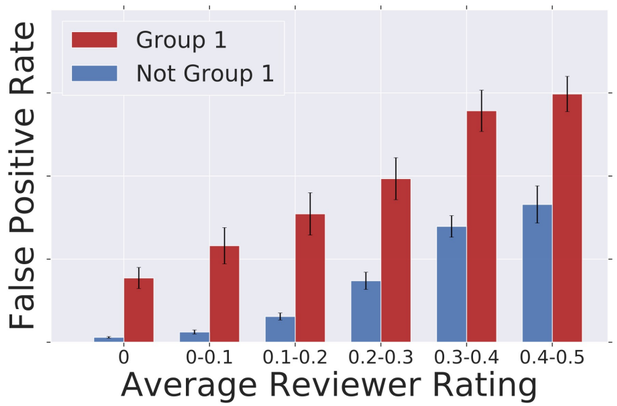


Figure credit: Zhang, B. H., Lemoine, B., & Mitchell, M. (2018, December). Mitigating unwanted biases with adversarial learning. In Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society (pp. 335-340).

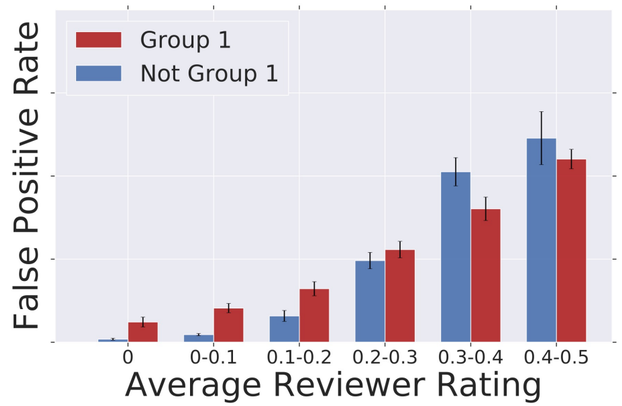
Mitigating Bias

Correlation loss

1. Include fairness metrics in the loss function (of machine learning algorithms)
2. Maximizing model's prediction power while minimizing the predicted label distribution across different protected groups



Before



After

Figure credit: Beutel, A., Chen, J., Doshi, T., Qian, H., Woodruff, A., Luu, C., ... & Chi, E. H. (2019, January). Putting fairness principles into practice: Challenges, metrics, and improvements. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society* (pp. 453-459).

Mitigating Bias

Constrained optimization

1. Explicitly encode fairness goals as optimization constraints
2. Trained model minimizes the loss function subject to the fairness constraints

$$\min \quad f(x)$$

$$\text{s.t.g} \quad FPR_{G1} \leq 1.1 \times FPR_{G2}$$

Figure credit: Cotter, A., Gupta, M., Jiang, H., Srebro, N., Sridharan, K., Wang, S., ... & You, S. (2019, May). Training well-generalizing classifiers for fairness metrics and other data-dependent constraints. In International Conference on Machine Learning (pp. 1397-1405). PMLR.

Practical Tools and Libraries

Discovering, understanding, and mitigating bias

1. Facets — Visualizations for ML datasets
2. FairVis — Discovering Bias in Machine Learning Using Visual Analytics
3. Microsoft InterpretML —A toolkit to help understand ML models and enable responsible AI
4. Microsoft Azure Interpretability Toolkit — SDK packages for ML explainability
5. IBM Open Scale — A system to monitor model fairness
6. IBM Fairness 360 — A toolkit for examine, report, and mitigate bias
7. Microsoft Fairlearn — A toolkit for assessing and improving fairness in AI
8. Datasheets for Datasets — Fairness information for shared datasets
9. Model cards — Fairness information for shared models
10. Fact sheets — Fairness information for AI service
11. Open source libraries: Aequitas, Audit-AI, FairML, Fairness Comparison, Fairness Measures, FairTest, Themis™, Themis-ML.

DEEP
LEARNING
INSTITUTE

PRAIRIE VIEW
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You