DLI Accelerated Data Science Teaching Kit

Lecture 4.1 - Sources of Bias and Fairness Measures

The Accelerated Data Science Teaching Kit is licensed by NVIDIA, Georgia Institute of Technology, and Prairie View A&M University under the Creative Commons Attribution-NonCommercial 4.0 International License.

Bias Sources and Fairness Measures

Outline

Bias Source

- Types of Bias
- Bias from the training data
- Bias from unequal model performance

Fairness Measures

- False/true positive/negative rate
- Demographic parity
- Equalized odds
- Disaggregated values

NVIDIA DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

Why do we care about bias and fairness?

Example 1: Skin Color Bias in Facial Recognition

1. Some software could not detect dark-skinned faces
2. Facial attribute classifiers have lower accuracy on black women than white men

Image credit: https://www.nytimes.com/2020/11/11/movies/coded-bias-review.html

Why do we care about bias and fairness?

Example 2: Gender Bias in Word Embeddings

1. Word embedding transforms words into vectors
2. The distance in the word embedding vector space has semantic meaning
3. Word embedding's arithmetic properties reveal gender bias

"Woman" + "King" – "Man" = "Queen"

"Woman" + "Doctor" – "Man" = "Nurse"

Example: Bolukbasi, T., Chang, K. W., Zou, J., Saligrama, V., & Kalai, A. (2016). Man is to computer programmer as woman is to homemaker? debiasing word embeddings. arXiv preprint arXiv:1607.06520.

NVIDIA DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

Where are the bias from?

Key source: training data

1. Data is a social mirror
2. Bias exist in real life and data collection
3. ML models learn from data patterns, including bias

NVIDIA DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

Types of Bias

1. **Reporting bias**: what people share does not align with real-world frequencies
2. **Selection bias**: data selection does not reflect random sample
3. **Temporal bias**: difference in populations and behaviors over time
4. **In-group bias**: preference of an individual for characteristics you share
5. **Out-group homogeneity bias**: tendency to stereotype individuals from other group
6. **Confirmation bias**: processing data in a way that affirm pre-existing hypothesis and beliefs
7. And 100+ other bias types listed on Wikipedia

Bias types: https://en.wikipedia.org/wiki/List_of_cognitive_biases

Bias in Data

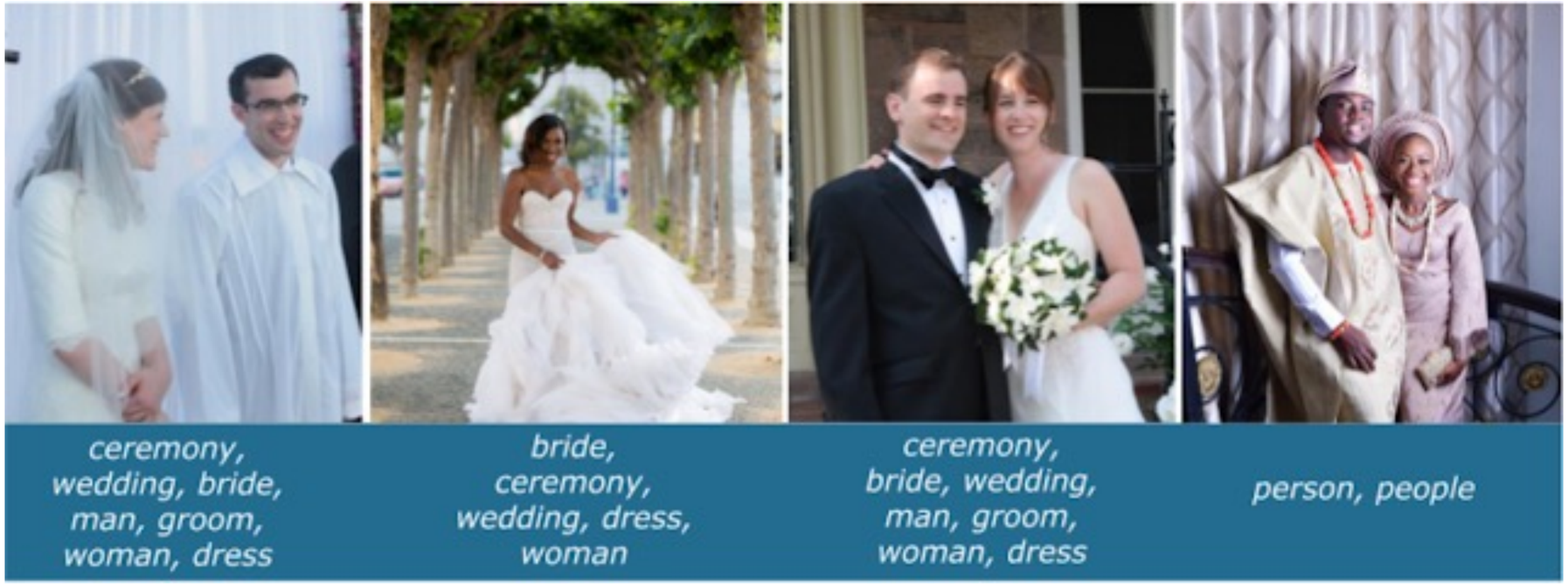
Biased data representation

Even when the data include an appropriate amount of samples for each group, some group can be represented less positively than others.

NVIDIA DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

Bias in Data

Biased labels



Data label reflects annotation’s pre-existing beliefs.

Image credit: https://ai.googleblog.com/2018/09/introducing-inclusive-images-competition.html

NVIDIA DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

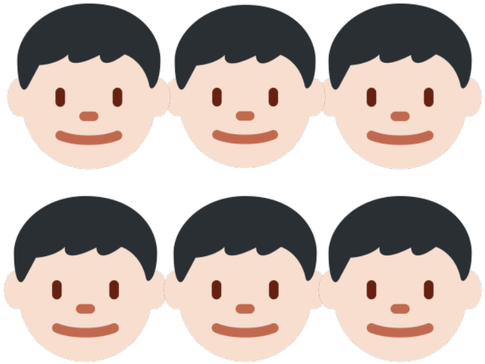
Bias in Interpretation

Bias from unequal model performance

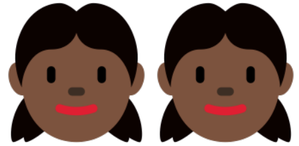
1. Bias from overgeneralization
2. Majority group enjoys higher accuracy in decision making

College Admission Model
91% accuracy

95% accuracy



32% accuracy

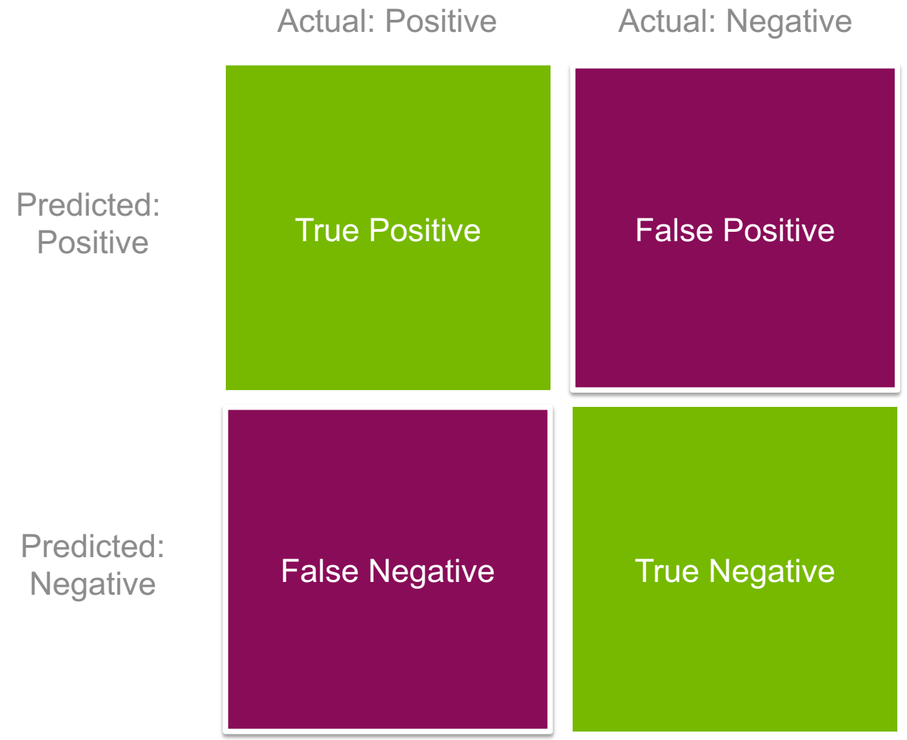


NVIDIA DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

How do we measure fairness?

Confusion matrix

1. Analyze confusion matrix in addition to test accuracy
2. Choose evaluation metrics based on the goal of your model
3. False positive might be better than false negative (COVID testing)



NVIDIA DEEP LEARNING INSTITUTE
Georgia Tech
PRAIRIE VIEW A&M UNIVERSITY

How do we measure fairness?

Equality of opportunity

1. Equality of opportunity: **sensitivity** (true positive rate) is equal across subgroups
2. It measures whether the people with positive labels are equally likely to be classified as positive regardless of their group membership

Male applicant

True Positive (TP) = 20	False Positive (FP) = 5
False Negative (FN) = 1	True Negative (TN) = 100

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{20}{21} = 0.9523$$

Female applicant

True Positive (TP) = 5	False Positive (FP) = 4
False Negative (FN) = 3	True Negative (TN) = 30

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{5}{8} = 0.625$$

Fairness metrics: https://developers.google.com/machine-learning/glossary/fairness

NVIDIA DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

How do we measure fairness?

Equalized odds

1. Equalized odds: **sensitivity** (true positive rate) and **specificity** (true negative rate) are both equal across subgroups
2. It measures for any given label and attribute, whether a model predicts that label equally well for all values of that attribute

Male applicant

True Positive (TP) = 20	False Positive (FP) = 5
False Negative (FN) = 1	True Negative (TN) = 100

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{20}{21} = 0.9523$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{100}{105} = 0.9523$$

Female applicant

True Positive (TP) = 5	False Positive (FP) = 4
False Negative (FN) = 3	True Negative (TN) = 30

$$\text{Sensitivity} = \frac{TP}{TP + FN} = \frac{5}{8} = 0.625$$

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{30}{34} = 0.8824$$

Fairness metrics: https://developers.google.com/machine-learning/glossary/fairness

How do we measure fairness?

Demographic parity

1. Demographic parity: model **positive rate** is equal across subgroups
2. It measures whether a model's positive prediction rate is not dependent on a given attribute

Male applicant

True Positive (TP) = 20	False Positive (FP) = 5
False Negative (FN) = 1	True Negative (TN) = 100

$$\text{Positive Rate} = \frac{TP + FP}{TP + TN + FP + FN} = \frac{25}{126} = 0.1984$$

Female applicant

True Positive (TP) = 5	False Positive (FP) = 4
False Negative (FN) = 3	True Negative (TN) = 30

$$\text{Positive Rate} = \frac{TP + TN}{TP + TN + FP + FN} = \frac{9}{42} = 0.2143$$

Fairness metrics: https://developers.google.com/machine-learning/glossary/fairness

NVIDIA DEEP LEARNING INSTITUTE | Georgia Tech | PRAIRIE VIEW A&M UNIVERSITY

DEEP
LEARNING
INSTITUTE

PRAIRIE VIEW
A&M UNIVERSITY

DLI Accelerated Data Science Teaching Kit

Thank You