Machine Learning (CSE 446): Bias and Fairness (continued)

Noah Smith

University of Washington
nasmith@cs.washington.edu

November 15, 2017

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased.

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased. E.g., reproducing already-biased judges' decisions, or training on data collected only on people charged with crimes.

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased.
- The design/definition of the task might encode bias.

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased.
- The design/definition of the task might encode bias.
  E.g., advertising that assumes binary categories like male/female, Democrat/Republican, etc.

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased.
- The design/definition of the task might encode bias.
- The features might encode bias.

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased.
- The design/definition of the task might encode bias.
- The features might encode bias.
  E.g., language translation systems that ignore context that disambiguates appropriate pronouns.

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased.
- The design/definition of the task might encode bias.
- The features might encode bias.
- The loss function might encode bias.

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased.
- The design/definition of the task might encode bias.
- The features might encode bias.
- The loss function might encode bias.
  E.g., if a minority class is infrequent, it may end up being ignored completely.

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased.
- The design/definition of the task might encode bias.
- The features might encode bias.
- The loss function might encode bias.
- Deployed systems that affect their own future inputs can create feedback loops and exacerbate their own biases.

Bias in Data

- The real-world process that produced the labels, or the sample, might be biased.
- The design/definition of the task might encode bias.
- The features might encode bias.
- The loss function might encode bias.
- Deployed systems that affect their own future inputs can create feedback loops and exacerbate their own biases.
  E.g., spammers adapt to spam-filtering tools, changing the data distribution.

Fairness and Disparate Impact

U.S. labor and housing laws measure discrimination using a rule like this:

$$p(Y = +1 \mid G \neq \textsf{male}) \geq 0.8 \cdot p(Y = +1 \mid G = \textsf{male})$$

and similarly for other protected attributes.

Fairness and Disparate Impact

U.S. labor and housing laws measure discrimination using a rule like this:

$$p(Y = +1 \mid G \neq \text{male}) \geq 0.8 \cdot p(Y = +1 \mid G = \text{male})$$

and similarly for other protected attributes.

- Can we just take "male" (and other protected attributes) out of the feature set?

Fairness and Disparate Impact

U.S. labor and housing laws measure discrimination using a rule like this:

$$p(Y = +1 \mid G \neq \text{male}) \geq 0.8 \cdot p(Y = +1 \mid G = \text{male})$$

and similarly for other protected attributes.

- Can we just take "male" (and other protected attributes) out of the feature set? (No.)

Fairness and Disparate Impact

U.S. labor and housing laws measure discrimination using a rule like this:

$$p(Y = +1 \mid G \neq \text{male}) \geq 0.8 \cdot p(Y = +1 \mid G = \text{male})$$

and similarly for other protected attributes.

- Can we just take "male" (and other protected attributes) out of the feature set? (No.)
- Can we satisfy this rule and still obtain high accuracy?

Fairness and Disparate Impact

U.S. labor and housing laws measure discrimination using a rule like this:

$$p(Y = +1 \mid G \neq \text{male}) \geq 0.8 \cdot p(Y = +1 \mid G = \text{male})$$

and similarly for other protected attributes.

- Can we just take "male" (and other protected attributes) out of the feature set? (No.)
- Can we satisfy this rule and still obtain high accuracy?
- Are there other (better?) measurements of fairness?