Fairness in Machine Learning

袁至誠
Chih-cheng Rex Yuan

Institute of Information Science

Tuesday 16th April, 2024

Bias

- Decision making by machine learning may be biased.
- Bias can come from several sources:
  - Biased data. ML is deisgned to replicate this.
  - Missing data. The datasets might not be representative.
  - Biased algorithms. The objective functions might introduce bias.
  - Sensitive attributes: Age, Gender, ..., etc..

Protected Attributes

What are the protected(sensitive) attributes?

Age	Gender	Occupation	Income	Education
28	M	Engineering	$80,000	Master
28	F	Engineering	$65,000	Master
45	M	Medicine	$100,000	Doctorate
40	F	Legal	$150,000	Law Degree
32	M	Education	$55,000	Bachelor

Table: Example Dataset

Fairness Through Unawareness

The most straightforward solution to fairness seems to be that just simply dropping all the protected columns.

- This is called fairness through unawareness.
- Formally it's

$$X_i = X_j \rightarrow \hat{Y}_i = \hat{Y}_j$$

  where i, j are individuals; X is the set of attributes except protected attributes; and $\hat{Y}$ is the prediction.
- Also known as fairness through blindness and anti-classification.

Fairness Through Unawareness

The downside of this is there could still be "proxy" attributes that correlate with protected attributes: like Occupation still correlates with Income.

~~Age~~	~~Gender~~	Occupation	~~Income~~	Education
~~28~~	~~M~~	Engineering	~~$80,000~~	Master
~~28~~	~~F~~	Engineering	~~$65,000~~	Master
~~45~~	~~M~~	Medicine	~~$100,000~~	Doctorate
~~40~~	~~F~~	Legal	~~$150,000~~	Law Degree
~~32~~	~~M~~	Education	~~$55,000~~	Bachelor

Table: Example Dataset

Disparate Impact

- In 1971, the US spreme court ruled it is illegal for hiring decisions to have "disparate impact" by race. It is taken as unintentional discrimination (as opposed to "disparate treatment", intentional discrimination).
- Legal issues involving disparate impact usually refer to the "80% Rule", advocated by the US Equal Employment Opportunity Commission, where the selection rate of a minority group is to be no less than 80% of that of a majority group.
- Formally, it requires that

$$\frac{P[\hat{Y} = 1|S \neq 1]}{P[\hat{Y} = 1|S = 1]} \geq 1 - \epsilon$$

where $\hat{Y} = 1$ represents acceptance(positive); $S = 1$ represents privileged group; $S \neq 1$ represents unprivileged group where S is some protected attributes.

Disparate Impact

For example, if for some job opening there are 10 female applicants and 100 male applicants, and there are 2 accepted females and 90 accepted males. The measure does not hold for $\epsilon = 0.2$ because

$$\frac{2/10}{80/100} = 0.25 \not\geq 0.8$$

while if there were 9 accepted females then it does hold because

$$\frac{9/10}{80/100} = 1.125 \geq 0.8$$

Demographic Parity

- Demographic parity is similar to disparate impact but, instead of ratio, difference is taken.
- Formally, it requires that

$$|P[\hat{Y} = 1|S = 1] - P[\hat{Y} = 1|S \neq 1]| \leq \epsilon$$

where $\hat{Y} = 1$ represents acceptance(positive); $S = 1$ represents privileged group; $S \neq 1$ represents unprivileged group where S is some protected attributes.

- This is usually known as affirmative action.

Disparate Impact and Demographic Parity

If for some job opening there are 10 female applicants and 100 male applicants, and there are 8 accepted females and 90 accepted males:

$$\frac{8/10}{90/100} = 0.\overline{8} \qquad |8/10 - 90/100| = 0.1$$

while if there were 1 accepted females and 20 accepted males:

$$\frac{1/10}{20/100} = 0.5 \qquad |1/10 - 20/100| = 0.1$$

Disparate Impact and Demographic Parity

Disadvantage of these two measures:

- "A fully accurate classifier may be considered unfair, when the base rates (i.e., the proportion of actual positive outcomes) of the various groups are significantly different."
- "The notion permits that we accept the qualified applicants in one demographic, but random individuals in another, so long as the percentages of acceptance match."
- It "often cripples the utility that we might hope to achieve, especially in the common scenario in which an outcome to be predicated, e.g. whether the loan be will defaulted, is correlated with the protected attribute"

Equalized Odds

- Equalized odds is designed to address the downsides of the previous two by taking into accounts the actual "ground truths" and consider the difference between false-positive rates and true-positive rates of the groups.
- Formally, it requires that

$$|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]| \leq \epsilon$$
$$|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S \neq 1, Y = 1]| \leq \epsilon$$

  where Y represents ground truths.
- A fully accurate classifier will necessarily satisfy the two equalized odds constraints.

Equal Opportunity

- Equal opportunity is similar to equalized odds but focuses on true-positive rates of the groups only.
- Formally, it requires that

$$|P[\hat{Y} = 1|S = 1, Y = 1] - P[\hat{Y} = 1|S \neq 1, Y = 1]| \leq \epsilon$$

  where Y represents ground truths.
- It is a relaxation of equalized odds that focuses on the typically considered "advantaged" group $Y = 1$. For example, it "requires that people who pay back their loan, have an equal opportunity of getting the loan in the first place (without specifying any requirement for those that will ultimately default)."

COMPAS

- COMPAS is an algorithm used by U.S. courts for predicting recidivism(reoffending) based on a questionaire.
- Supreme Court ruled that it can be considered by judges during sentencing, but there must be warnings about the tool's "limitations and cautions."
- Risk score: $s = a(-w_1) + a_{1st}(-w_2) + h_{vio}w_3 + v_{edu}w_4 + h_{nc}w_5$
- In 2016, ProPublica found that the algorithm is biased.

  Black defendants were often predicted to be at a higher risk of recidivism than they actually were. White defendants were often predicted to be less risky than they were.

- The false-positive rates vary significantly across black people and white people, violating equalized odds.

[1](Link) ProPublica - How We Analyzed the COMPAS Recidivism Algorithm

[2](Link) Vsauce2 - The Dangerous Math Used To Predict Criminals

Individual Fairness

- All of the above fairness measures consider fairness across groups, while individual fairness requires that similar individuals will be treated similarly.
- Formally, it requires that

$$|P[\hat{Y}^{(i)} = y|X^{(i)}, S^{(i)}] - P[\hat{Y}^{(j)} = y|X^{(j)}, S^{(j)}]| \leq \epsilon \text{ if } d(i,j) \approx 0$$

  where i, j denotes two individuals; $S^{(\cdot)}$ refers to sensitive attributes; $X^{(\cdot)}$ refers to associated features; and $d(i,j)$ is a distance metric between individuals.
- "This measure considers other individual attributes for defining fairness rather than just the sensitive attributes."

Other Measures

- Overall accuracy equality:

$$|P[Y = \hat{Y}|S = 1] - P[Y = \hat{Y}|S \neq 1]| \leq \epsilon$$

  where $Y = \hat{Y}$ means that the prediction was correct.

- Predictive parity:

$$|P[Y = 1|S = 1, \hat{Y} = 1] - P[Y = 1|S \neq 1, \hat{Y} = 1]| \leq \epsilon$$

  This requires that the "positive predictive values" are similar across groups, meaning the probability of an individual with a positive prediction actually experiencing a positive outcome.

- Equal calibration:

$$|P[Y = 1|S = 1, V = v] - P[Y = 1|S \neq 1, V = v]| \leq \epsilon$$

  where V is the predicted probability value. When V is binary, this is equivalent to predictive parity.

Other Measures

- Conditional statistical parity:

$$|P[\hat{Y} = 1|S = 1, L = l] - P[\hat{Y} = 1|S \neq 1, L = l]| \leq \epsilon$$

  where L is a set of additional "legitimate" factors; for example, black and white defendants who have the same number of prior convictions.
- Predictive equality:

$$|P[\hat{Y} = 1|S = 1, Y = 0] - P[\hat{Y} = 1|S \neq 1, Y = 0]| \leq \epsilon$$

  This is similar to equal opportunity, but instead of focusing only on true-positive, this focuses only on false-positive rates.
- Conditional use accuracy equality:

$$|P[Y = 1|S = 1, \hat{Y} = 1] - P[Y = 1|S \neq 1, \hat{Y} = 1]| \leq \epsilon$$
$$|P[Y = 0|S = 1, \hat{Y} = 0] - P[Y = 0|S \neq 1, \hat{Y} = 0]| \leq \epsilon$$

  This is similar to predictive parity but it additionally requires negative predictive values to be similar across groups.

Other Measures

- Treatment equality:

$$|\frac{FN_{S=1}}{FP_{S=1}} - \frac{FN_{S\neq 1}}{FP_{S\neq 1}}| \leq \epsilon$$

  This is similar to equalized odds but difference of ratio is taken instead.

- Balance for the positive class:

$$|E[V|Y=1,S=1] - E[V|Y=1,S\neq 1]| \leq \epsilon$$

  where V is the predicted probability value.

- Balance for the negative class:

$$|E[V|Y=0,S=1] - E[V|Y=0,S\neq 1]| \leq \epsilon$$

  where V is the predicted probability value.

Other Measures

- Mutual information:

$$\sum P(\hat{y}, s) \log(\frac{P(\hat{y}, s)}{P(\hat{y})P(s)}) \leq \epsilon$$

  This measures the mutual dependence between the sensitive feature and the predicted outcome.

- Mean difference:

$$|E[\hat{Y}|S = 1] - E[\hat{Y}|S\neg 1]| \leq \epsilon$$

  This measures the difference between the means of the predictions across groups. This is similar to demographic parity when the target is binary.