Fairness in Machine Learning

Beyond Observational Measures

Michael Yang Prof. Anand Sarwate

July 12, 2018

DIMACS REU 2018

Funding Received from: National Science Foundation CCF-1559855

Overview

Overview

Overview

- Preliminaries
  - Notation
  - The Philosophy of Fairness
- Common Observational Notions of Fairness
- Problems with Observational Notions of Fairness
- Beyond Observational Measures
  - Aggregation ⇒ Finer-Grained Fairness
  - Obliviousness ⇒ Causality
  - Short-Sightedness ⇒ Modeling Long-Term Fairness

Overview

Preliminaries

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

X: covariates; features; independent variables

Ŷ: what the ML program or decision-maker *thinks* Y is; the predicted output

Ŝ: risk scores, which are thresholded into $0-1$ scores $\hat{Y}$

Machine Learning Task:
Learn $f : X \to \hat{Y}$ on a labeled set to minimize error on new observations

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

X: covariates; features; independent variables

$\hat{Y}$: what the ML program or decision-maker *thinks* Y is; the predicted output

$\hat{S}$: risk scores, which are thresholded into $0-1$ scores $\hat{Y}$

Machine Learning Task:
Learn $f : X \rightarrow \hat{Y}$ on a labeled set to minimize error on new observations

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

X: covariates; features; independent variables

$\hat{Y}$: what the ML program or decision-maker *thinks* Y is; the predicted output

$\hat{S}$: risk scores, which are thresholded into $0-1$ scores $\hat{Y}$

Machine Learning Task:
Learn $f : X \to \hat{Y}$ on a labeled set to minimize error on new observations

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

X: covariates; features; independent variables

Ŷ: what the ML program or decision-maker *thinks* Y is; the predicted output

Ŝ: risk scores, which are thresholded into $0 - 1$ scores $\hat{Y}$

Machine Learning Task:
Learn $f : X \rightarrow \hat{Y}$ on a labeled set to minimize error on new observations

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

X: covariates; features; independent variables

$\hat{Y}$: what the ML program or decision-maker *thinks* Y is; the predicted output

$\hat{S}$: risk scores, which are thresholded into $0-1$ scores $\hat{Y}$

Machine Learning Task:
Learn $f : X \to \hat{Y}$ on a labeled set to minimize error on new observations

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

X: covariates; features; independent variables

$\hat{Y}$: what the ML program or decision-maker *thinks* Y is; the predicted output

$\hat{S}$: risk scores, which are thresholded into $0-1$ scores $\hat{Y}$

Machine Learning Task:
Learn $f : X \to \hat{Y}$ on a labeled set to minimize error on new observations

Some ML/stats notation

Y: the target variable; outcome of interest; **the ground truth**

A: group membership in something protected (e.g. race, gender)

X: covariates; features; independent variables

$\hat{Y}$: what the ML program or decision-maker *thinks* Y is; the predicted output

$\hat{S}$: risk scores, which are thresholded into $0-1$ scores $\hat{Y}$

Machine Learning Task:
Learn $f : X \rightarrow \hat{Y}$ on a labeled set to minimize error on new observations

Central Question
What does it mean for an algorithm to be fair?

Philosophically, we might ask:

- Should people be penalized for factors outside their control?
- Should decisions try to ultimate rectify group-level inequalities?
- Are people deterministic or probabilistic?

Central Question
What does it mean for an algorithm to be fair?

Philosophically, we might ask:

- Should people be penalized for factors outside their control?
- Should decisions try to ultimate rectify group-level inequalities?
- Are people deterministic or probabilistic?

Central Question
What does it mean for an algorithm to be fair?

Philosophically, we might ask:

- Should people be penalized for factors outside their control?
- Should decisions try to ultimate rectify group-level inequalities?
- Are people deterministic or probabilistic?

Central Question
What does it mean for an algorithm to be fair?

Philosophically, we might ask:

- Should people be penalized for factors outside their control?
- Should decisions try to ultimate rectify group-level inequalities?
- Are people deterministic or probabilistic?

The Philosophy of Fairness

Central Question
What does it mean for an algorithm to be fair?

Philosophically, we might ask:

- Should people be penalized for factors outside their control?
- Should decisions try to ultimate rectify group-level inequalities?
- Are people deterministic or probabilistic?

Common Observational Notions of Fairness

Observational Notions of Fairness

For assessment, only require black-box access to a predictor (just the set of inputs and outputs)

In theory, equally applicable to human decision-makers as algorithms

For each of the following definitions, one must also consider:

- How to test this fairness definition on an algorithm in a principled way?
- How to learn fairly with respect to this definition (fairness-aware classifiers)?

Observational Notions of Fairness

For assessment, only require black-box access to a predictor (just the set of inputs and outputs)

In theory, equally applicable to human decision-makers as algorithms

For each of the following definitions, one must also consider:

- How to test this fairness definition on an algorithm in a principled way?
- How to learn fairly with respect to this definition (fairness-aware classifiers)?

Observational Notions of Fairness

For assessment, only require black-box access to a predictor (just the set of inputs and outputs)

In theory, equally applicable to human decision-makers as algorithms

For each of the following definitions, one must also consider:

- How to test this fairness definition on an algorithm in a principled way?
- How to learn fairly with respect to this definition (fairness-aware classifiers)?

Observational Notions of Fairness

For assessment, only require black-box access to a predictor (just the set of inputs and outputs)

In theory, equally applicable to human decision-makers as algorithms

For each of the following definitions, one must also consider:

- How to test this fairness definition on an algorithm in a principled way?
- How to learn fairly with respect to this definition (fairness-aware classifiers)?

Observational Notions of Fairness

For assessment, only require black-box access to a predictor (just the set of inputs and outputs)

In theory, equally applicable to human decision-makers as algorithms

For each of the following definitions, one must also consider:

- How to test this fairness definition on an algorithm in a principled way?
- How to learn fairly with respect to this definition (fairness-aware classifiers)?

Statistical Parity

Definition

$$\hat{Y} \perp A$$

Statistical Parity

Definition

$$\hat{Y} \perp A$$

The percentage of acceptances and rejections should be equivalent across protected groups:

$$P(\hat{Y} = 1 \mid A = a) \stackrel{?}{=} P(\hat{Y} = 1 \mid A = a')$$

Statistical Parity

Definition

$$\hat{Y} \perp A$$

The percentage of acceptances and rejections should be equivalent across protected groups:

$$P(\hat{Y} = 1 \mid A = a) \stackrel{?}{=} P(\hat{Y} = 1 \mid A = a')$$

In the ideal world, protected attributes are not predictive of outcomes

Statistical Parity

Definition

$$\hat{Y} \perp A$$

The percentage of acceptances and rejections should be equivalent across protected groups:

$$P(\hat{Y} = 1 \mid A = a) \stackrel{?}{=} P(\hat{Y} = 1 \mid A = a')$$

In the ideal world, protected attributes are not predictive of outcomes

Related to p-% rules: The ratio of outcomes $\frac{p(Y|a')}{p(Y|a)}$ should not be less than less than p

Definition

$$\hat{Y} \perp A$$

Problems:

- Not compatible with the ideal predictor $\hat{Y} = Y$
- Too strong: Y may correlate with A for "benign" reasons
- *Also* too weak: Possible for $\hat{Y} \not\perp A \mid V$ for some V

  What if $V = Y$? $\Rightarrow$ "Scapegoating" protected individuals

Statistical Parity

Definition

$$\hat{Y} \perp A$$

Problems:

- Not compatible with the ideal predictor $\hat{Y} = Y$
- Too strong: Y may correlate with A for "benign" reasons
- *Also* too weak: Possible for $\hat{Y} \not\perp A \mid V$ for some V

  What if $V = Y$? $\Rightarrow$ "Scapegoating" protected individuals

Statistical Parity

Definition

$$\hat{Y} \perp A$$

Problems:

- Not compatible with the ideal predictor $\hat{Y} = Y$
- Too strong: Y may correlate with A for "benign" reasons
- *Also* too weak: Possible for $\hat{Y} \not\perp A \mid V$ for some V

  What if $V = Y$? $\Rightarrow$ "Scapegoating" protected individuals

Statistical Parity

Definition

$$\hat{Y} \perp A$$

Problems:

- Not compatible with the ideal predictor $\hat{Y} = Y$
- Too strong: Y may correlate with A for "benign" reasons
- *Also* too weak: Possible for $\hat{Y} \not\perp A \mid V$ for some V

  What if $V = Y$? $\Rightarrow$ "Scapegoating" protected individuals

Statistical Parity

Definition

$$\hat{Y} \perp A$$

Problems:

- Not compatible with the ideal predictor $\hat{Y} = Y$
- Too strong: Y may correlate with A for "benign" reasons
- *Also* too weak: Possible for $\hat{Y} \not\perp A \mid V$ for some V

  What if $V = Y$? $\Rightarrow$ "Scapegoating" protected individuals

Conditional Statistical Parity

Definition

$$\hat{Y} \perp A \mid X$$

Within in each possible bucket of relevant information, the probability of decision is the same across protected groups

With enough X, this definition is equivalent to treating "nearby" individuals similarly (fairness through awareness, Dwork et al., 2012)

Conditional Statistical Parity

Definition

$$\hat{Y} \perp A \mid X$$

Within in each possible bucket of relevant information, the probability of decision is the same across protected groups

With enough X, this definition is equivalent to treating "nearby" individuals similarly (fairness through awareness, Dwork et al., 2012)

Conditional Statistical Parity

Definition

$$\hat{Y} \perp A \mid X$$

Within in each possible bucket of relevant information, the probability of decision is the same across protected groups

With enough X, this definition is equivalent to treating “nearby” individuals similarly (fairness through awareness, Dwork et al., 2012)

Equalized Odds

Definition

$$\hat{Y} \perp A \mid Y$$

Every equally qualified individual has an equal chance of receiving positive classification

Equal rates of false positives (FP) and false negatives (FN)

Compatible with perfect prediction

Related to conditional statistical parity... except Y is not available for new observations

Retrospectively getting the prediction right (how right **were** the predictions by group)

Equalized Odds

Definition

$$\hat{Y} \perp A \mid Y$$

Every equally qualified individual has an equal chance of receiving positive classification

Equal rates of false positives (FP) and false negatives (FN)

Compatible with perfect prediction

Related to conditional statistical parity... except Y is not available for new observations

Retrospectively getting the prediction right (how right **were** the predictions by group)

Equalized Odds

Definition

$$\hat{Y} \perp A \mid Y$$

Every equally qualified individual has an equal chance of receiving positive classification

Equal rates of false positives (FP) and false negatives (FN)

Compatible with perfect prediction

Related to conditional statistical parity... except Y is not available for new observations

Retrospectively getting the prediction right (how right **were** the predictions by group)

Equalized Odds

Definition

$$\hat{Y} \perp A \mid Y$$

Every equally qualified individual has an equal chance of receiving positive classification

Equal rates of false positives (FP) and false negatives (FN)

Compatible with perfect prediction

Related to conditional statistical parity... except Y is not available for new observations

Retrospectively getting the prediction right (how right **were** the predictions by group)

Definition

$$\hat{Y} \perp A \mid Y$$

Every equally qualified individual has an equal chance of receiving positive classification

Equal rates of false positives (FP) and false negatives (FN)

Compatible with perfect prediction

Related to conditional statistical parity... except Y is not available for new observations

Retrospectively getting the prediction right (how right **were** the predictions by group)

Equalized Odds

Definition

$$\hat{Y} \perp A \mid Y$$

Every equally qualified individual has an equal chance of receiving positive classification

Equal rates of false positives (FP) and false negatives (FN)

Compatible with perfect prediction

Related to conditional statistical parity... except Y is not available for new observations

Retrospectively getting the prediction right (how right **were** the predictions by group)

Calibration

Definition

$$Y \perp A \mid \hat{Y}$$

Predicted scores should reflect equal probability of the true status Y across protected groups

The sole criterion for non-discrimination in many social sciences (e.g. psychology, educational tools)

Prospectively getting the prediction right (how right **will** the predictions be by group)

Calibration

Definition

$$Y \perp A \mid \hat{Y}$$

Predicted scores should reflect equal probability of the true status Y across protected groups

The sole criterion for non-discrimination in many social sciences (e.g. psychology, educational tools)

Prospectively getting the prediction right (how right **will** the predictions be by group)

Definition

$$Y \perp A \mid \hat{Y}$$

Predicted scores should reflect equal probability of the true status Y across protected groups

The sole criterion for non-discrimination in many social sciences (e.g. psychology, educational tools)

Prospectively getting the prediction right (how right **will** the predictions be by group)

Calibration

Definition

$$Y \perp A \mid \hat{Y}$$

Predicted scores should reflect equal probability of the true status Y across protected groups

The sole criterion for non-discrimination in many social sciences (e.g. psychology, educational tools)

Prospectively getting the prediction right (how right **will** the predictions be by group)

Problems with Observational Notions of Fairness

Theorem

Calibration and **equalized odds** cannot both hold for a set of predictions, under two conditions:

- Perfection prediction is not achieved
- Background rates of Y are unequal across groups

Theorem resulted from the debate about COMPAS scores

Theorem

Calibration and **equalized odds** cannot both hold for a set of predictions, under two conditions:

- Perfection prediction is not achieved
- Background rates of Y are unequal across groups

Theorem resulted from the debate about COMPAS scores

Theorem

Calibration and **equalized odds** cannot both hold for a set of predictions, under two conditions:

- Perfection prediction is not achieved
- Background rates of Y are unequal across groups

Theorem resulted from the debate about COMPAS scores

Theorem

Calibration and **equalized odds** cannot both hold for a set of predictions, under two conditions:

- Perfection prediction is not achieved
- Background rates of Y are unequal across groups

Theorem resulted from the debate about COMPAS scores

Theorem

Calibration and **equalized odds** cannot both hold for a set of predictions, under two conditions:

- Perfection prediction is not achieved
- Background rates of Y are unequal across groups

Theorem resulted from the debate about COMPAS scores

COMPAS Refresher

What COMPAS got right:

- Scores were *well-calibrated*:

$$E[Y = 1 \mid y = 0, A = \text{black}] = E[Y = 1 \mid y = 0, A = \text{white}]$$

  Translation: Black people with a score of 7 were as likely to recidivate as white people with a score of 7

What COMPAS got wrong:

- Unequal false negative rates:

$$E[y = 0 \mid Y = 1, A = \text{black}] \neq E[y = 0 \mid Y = 1, A = \text{white}]$$

  Translation: White people who would actually recidivate almost twice as likely to be scored "low risk"

- Unequal false positive rates:

$$E[y = 1 \mid Y = 0, A = \text{black}] \neq E[y = 1 \mid Y = 0, A = \text{white}]$$

  Translation: Black people who would not actually recidivate almost twice as likely to be scored "higher risk"

COMPAS Refresher

What COMPAS got right:

- Scores were *well-calibrated*:

$$E[Y = 1 \mid y = 0, A = \text{black}] = E[Y = 1 \mid y = 0, A = \text{white}]$$

  Translation: Black people with a score of 7 were as likely to recidivate as white people with a score of 7

What COMPAS got wrong:

- Unequal false negative rates:

$$E[y = 0 \mid Y = 1, A = \text{black}] \neq E[y = 0 \mid Y = 1, A = \text{white}]$$

  Translation: White people who would actually recidivate almost twice as likely to be scored “low risk”

- Unequal false positive rates:

$$E[y = 1 \mid Y = 0, A = \text{black}] \neq E[y = 1 \mid Y = 0, A = \text{white}]$$

  Translation: Black people who would not actually recidivate almost twice as likely to be scored “higher risk”

Fairness v. Fairness Trade-Off

Impossibility matters: Not only is there a fairness v. accuracy trade-off

...There's also a *fairness v. fairness* trade-off

What matters more: Getting answers prospectively v. retrospectively correct?

But wait, there's more!

Fairness v. Fairness Trade-Off

Impossibility matters: Not only is there a fairness v. accuracy trade-off

...There's also a *fairness v. fairness* trade-off

What matters more: Getting answers prospectively v. retrospectively correct?

But wait, there's more!

Fairness v. Fairness Trade-Off

Impossibility matters: Not only is there a fairness v. accuracy trade-off

...There's also a *fairness v. fairness* trade-off

What matters more: Getting answers prospectively v. retrospectively correct?

But wait, there's more!

Impossibility matters: Not only is there a fairness v. accuracy trade-off

...There's also a *fairness v. fairness* trade-off

What matters more: Getting answers prospectively v. retrospectively correct?

But wait, there's more!

Where does one set the threshold for decisions?

First, trying to be “fair” as defined above leads to different thresholds for different groups

But the direction of unfairness can be reversed with different distributions of risk scores! (Simoiu, Corbett-Davies, and Goel, 2017; Corbett-Davies et al., 2017)

Infra-Marginality (aka Fairness v. Fairness Trade-Off, Part 2)

Where does one set the threshold for decisions?

First, trying to be "fair" as defined above leads to different thresholds for different groups

But the direction of unfairness can be reversed with different distributions of risk scores! (Simoiu, Corbett-Davies, and Goel, 2017; Corbett-Davies et al., 2017)

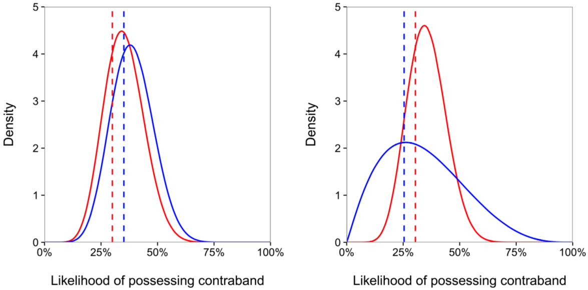
Infra-Marginality (aka Fairness v. Fairness Trade-Off, Part 2)

Where does one set the threshold for decisions?

First, trying to be “fair” as defined above leads to different thresholds for different groups

But the direction of unfairness can be reversed with different distributions of risk scores! (Simoiu, Corbett-Davies, and Goel, 2017; Corbett-Davies et al., 2017)

Infra-Marginality



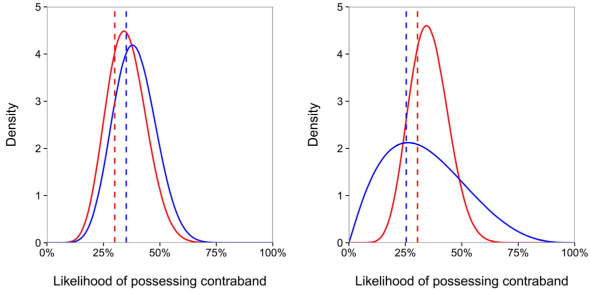
(a)
(b)

Both graphs are observationally equivalent:

- Red individuals searched more often (area under the curve right of threshold)
- Searches of red individuals are less successful

But in (b), blue individuals have the lower threshold!

Infra-Marginality



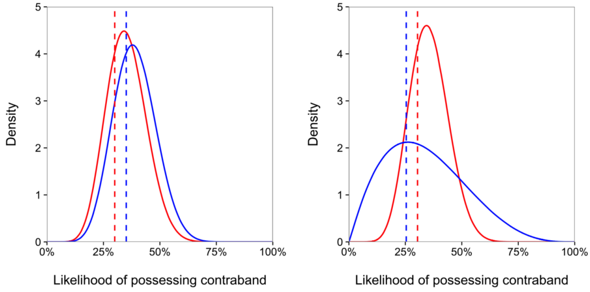
(a)
(b)

Both graphs are observationally equivalent:

- Red individuals searched more often (area under the curve right of threshold)
- Searches of red individuals are less successful

But in (b), blue individuals have the lower threshold!

Infra-Marginality



(a)

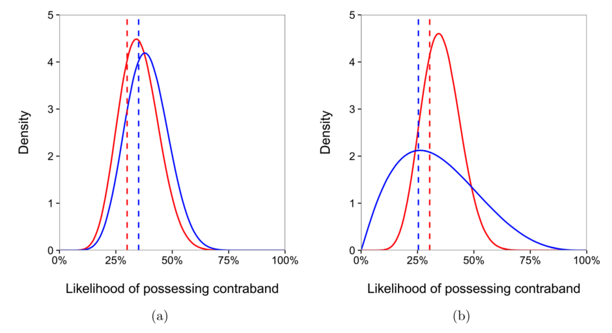
(b)

Both graphs are observationally equivalent:

- Red individuals searched more often (area under the curve right of threshold)
- Searches of red individuals are less successful

But in (b), blue individuals have the lower threshold!

Infra-Marginality



Both graphs are observationally equivalent:

- Red individuals searched more often (area under the curve right of threshold)
- Searches of red individuals are less successful

But in (b), blue individuals have the lower threshold!

Beyond Observational Measures

Aggregation

We have to fix from the onset what a protected group is

Consider the following binary prediction task:

- Protected attributes are race (red or blue) and gender (male or female)
- An algorithm only predicts $\hat{Y} = 1$ for red males and blue females

"Fair" with respect to either race or gender *considered alone*

Aggregation

We have to fix from the onset what a protected group is

Consider the following binary prediction task:

- Protected attributes are race (red or blue) and gender (male or female)
- An algorithm only predicts $\hat{Y} = 1$ for red males and blue females

"Fair" with respect to either race or gender *considered alone*

Aggregation

We have to fix from the onset what a protected group is

Consider the following binary prediction task:

- Protected attributes are race (red or blue) and gender (male or female)
- An algorithm only predicts $\hat{Y} = 1$ for red males and blue females

"Fair" with respect to either race or gender *considered alone*

Aggregation

We have to fix from the onset what a protected group is

Consider the following binary prediction task:

- Protected attributes are race (red or blue) and gender (male or female)
- An algorithm only predicts $\hat{Y} = 1$ for red males and blue females

"Fair" with respect to either race or gender *considered alone*

Aggregation

We have to fix from the onset what a protected group is

Consider the following binary prediction task:

- Protected attributes are race (red or blue) and gender (male or female)
- An algorithm only predicts $\hat{Y} = 1$ for red males and blue females

“Fair” with respect to either race or gender *considered alone*

Finer-Grained Fairness

Methods for *assessing fairness* on arbitrary subgroups:

- Learn a classification tree (Chouldechova, 2017)
- Use (carefully-constructed) statistical tests to find differential predictions across exponentially many groups (in linear time) (Zhang and Neill, 2016)

Methods for *learning fairly* on arbitrary subgroups:

- Kearns et al., 2017 provide a method for learning a fixed notion of fairness with respect to arbitrarily many subgroups (uses Learner-Auditor dynamics)
- Hébert-Johnson et al., 2017; Kim, Ghorbani, and Zou, 2018 learn more accurately with respect to arbitrary many subgroups (boosting)

Finer-Grained Fairness

Methods for *assessing fairness* on arbitrary subgroups:

- Learn a classification tree (Chouldechova, 2017)
- Use (carefully-constructed) statistical tests to find differential predictions across exponentially many groups (in linear time) (Zhang and Neill, 2016)

Methods for *learning fairly* on arbitrary subgroups:

- Kearns et al., 2017 provide a method for learning a fixed notion of fairness with respect to arbitrarily many subgroups (uses Learner-Auditor dynamics)
- Hébert-Johnson et al., 2017; Kim, Ghorbani, and Zou, 2018 learn more accurately with respect to arbitrary many subgroups (boosting)

Finer-Grained Fairness

Methods for *assessing fairness* on arbitrary subgroups:

- Learn a classification tree (Chouldechova, 2017)
- Use (carefully-constructed) statistical tests to find differential predictions across exponentially many groups (in linear time) (Zhang and Neill, 2016)

Methods for *learning fairly* on arbitrary subgroups:

- Kearns et al., 2017 provide a method for learning a fixed notion of fairness with respect to arbitrarily many subgroups (uses Learner-Auditor dynamics)
- Hébert-Johnson et al., 2017; Kim, Ghorbani, and Zou, 2018 learn more accurately with respect to arbitrary many subgroups (boosting)

Finer-Grained Fairness

Methods for *assessing fairness* on arbitrary subgroups:

- Learn a classification tree (Chouldechova, 2017)
- Use (carefully-constructed) statistical tests to find differential predictions across exponentially many groups (in linear time) (Zhang and Neill, 2016)

Methods for *learning fairly* on arbitrary subgroups:

- Kearns et al., 2017 provide a method for learning a fixed notion of fairness with respect to arbitrarily many subgroups (uses Learner-Auditor dynamics)
- Hébert-Johnson et al., 2017; Kim, Ghorbani, and Zou, 2018 learn more accurately with respect to arbitrary many subgroups (boosting)

Finer-Grained Fairness

Methods for *assessing fairness* on arbitrary subgroups:

- Learn a classification tree (Chouldechova, 2017)
- Use (carefully-constructed) statistical tests to find differential predictions across exponentially many groups (in linear time) (Zhang and Neill, 2016)

Methods for *learning fairly* on arbitrary subgroups:

- Kearns et al., 2017 provide a method for learning a fixed notion of fairness with respect to arbitrarily many subgroups (uses Learner-Auditor dynamics)
- Hébert-Johnson et al., 2017; Kim, Ghorbani, and Zou, 2018 learn more accurately with respect to arbitrary many subgroups (boosting)

Finer-Grained Fairness

Methods for *assessing fairness* on arbitrary subgroups:

- Learn a classification tree (Chouldechova, 2017)
- Use (carefully-constructed) statistical tests to find differential predictions across exponentially many groups (in linear time) (Zhang and Neill, 2016)

Methods for *learning fairly* on arbitrary subgroups:

- Kearns et al., 2017 provide a method for learning a fixed notion of fairness with respect to arbitrarily many subgroups (uses Learner-Auditor dynamics)
- Hébert-Johnson et al., 2017; Kim, Ghorbani, and Zou, 2018 learn more accurately with respect to arbitrary many subgroups (boosting)

Obliviousness

Causal Limitation of Observational Measures

There are two scenarios with identical joint distributions, but completely different interpretations for fairness (Hardt, Price, and Srebro, 2016).

Causal models help us make this distinction.

We can now answer:

- What unobserved variables are in our scenario, and what are their values? (What's the inherent risk?)
- What are the functional or *causal* relationship between variables in our scenario? (Kilbertus et al., 2017)
- **Importantly,** what would predictions have been if A had been different? (Kusner et al., 2017)

Causal Limitation of Observational Measures

There are two scenarios with identical joint distributions, but completely different interpretations for fairness (Hardt, Price, and Srebro, 2016).

Causal models help us make this distinction.

We can now answer:

- What unobserved variables are in our scenario, and what are their values? (What's the inherent risk?)
- What are the functional or *causal* relationship between variables in our scenario? (Kilbertus et al., 2017)
- **Importantly,** what would predictions have been if A had been different? (Kusner et al., 2017)

Obliviousness

Causal Limitation of Observational Measures

There are two scenarios with identical joint distributions, but completely different interpretations for fairness (Hardt, Price, and Srebro, 2016).

Causal models help us make this distinction.

We can now answer:

- What unobserved variables are in our scenario, and what are their values? (What's the inherent risk?)
- What are the functional or *causal* relationship between variables in our scenario? (Kilbertus et al., 2017)
- **Importantly,** what would predictions have been if A had been different? (Kusner et al., 2017)

Causal Limitation of Observational Measures

There are two scenarios with identical joint distributions, but completely different interpretations for fairness (Hardt, Price, and Srebro, 2016).

Causal models help us make this distinction.

We can now answer:

- What unobserved variables are in our scenario, and what are their values? (What's the inherent risk?)
- What are the functional or *causal* relationship between variables in our scenario? (Kilbertus et al., 2017)
- **Importantly,** what would predictions have been if A had been different? (Kusner et al., 2017)

Causal Limitation of Observational Measures

There are two scenarios with identical joint distributions, but completely different interpretations for fairness (Hardt, Price, and Srebro, 2016).

Causal models help us make this distinction.

We can now answer:

- What unobserved variables are in our scenario, and what are their values? (What's the inherent risk?)
- What are the functional or *causal* relationship between variables in our scenario? (Kilbertus et al., 2017)
- **Importantly,** what would predictions have been if A had been different? (Kusner et al., 2017)

Causal Limitation of Observational Measures

There are two scenarios with identical joint distributions, but completely different interpretations for fairness (Hardt, Price, and Srebro, 2016).

Causal models help us make this distinction.

We can now answer:

- What unobserved variables are in our scenario, and what are their values? (What's the inherent risk?)
- What are the functional or *causal* relationship between variables in our scenario? (Kilbertus et al., 2017)
- **Importantly,** what would predictions have been if A had been different? (Kusner et al., 2017)

Causal Limitation of Observational Measures

There are two scenarios with identical joint distributions, but completely different interpretations for fairness (Hardt, Price, and Srebro, 2016).

Causal models help us make this distinction.

We can now answer:

- What unobserved variables are in our scenario, and what are their values? (What's the inherent risk?)
- What are the functional or *causal* relationship between variables in our scenario? (Kilbertus et al., 2017)
- **Importantly,** what would predictions have been if A had been different? (Kusner et al., 2017)

(The Problem with) Causality

The problem is finding the model $\mathcal{M}$:

"Counterfactuals assumptions such as structural equations are *in general unfalsifiable* even if interventional data for all variables is available… Having passed testable implications, the remaining components of a counterfactual model should be understood as conjectures formulated according to the best of our knowledge." (Kusner et al., 2017)

One workaround, "Multi-world fairness" (Russell et al., 2017), allows for optimizing counterfactual fairness with respect to multiple different models

But this doesn't address the issue of discovering path-specific discrimination.

The problem is finding the model $\mathcal{M}$:

"Counterfactuals assumptions such as structural equations are *in general unfalsifiable* even if interventional data for all variables is available… Having passed testable implications, the remaining components of a counterfactual model should be understood as conjectures formulated according to the best of our knowledge." (Kusner et al., 2017)

One workaround, "Multi-world fairness" (Russell et al., 2017), allows for optimizing counterfactual fairness with respect to multiple different models

But this doesn't address the issue of discovering path-specific discrimination.

The problem is finding the model $\mathcal{M}$:

"Counterfactuals assumptions such as structural equations are *in general unfalsifiable* even if interventional data for all variables is available... Having passed testable implications, the remaining components of a counterfactual model should be understood as conjectures formulated according to the best of our knowledge." (Kusner et al., 2017)

One workaround, "Multi-world fairness" (Russell et al., 2017), allows for optimizing counterfactual fairness with respect to multiple different models

But this doesn't address the issue of discovering path-specific discrimination.

(The Problem with) Causality

The problem is finding the model $\mathcal{M}$:

"Counterfactuals assumptions such as structural equations are *in general unfalsifiable* even if interventional data for all variables is available... Having passed testable implications, the remaining components of a counterfactual model should be understood as conjectures formulated according to the best of our knowledge." (Kusner et al., 2017)

One workaround, "Multi-world fairness" (Russell et al., 2017), allows for optimizing counterfactual fairness with respect to multiple different models

But this doesn't address the issue of discovering path-specific discrimination.

"Delayed Impact of Fair Machine Learning" (Liu et al., 2018)

- Fairness of a single decision (e.g. loan approval) does not consider the effect on the overall group
- So, what's the effect of different thresholds on populations on the *change* in *group credit score?*
- We can figure this out by calculating the effect of false positives and false negatives on credit score...
- "Fair" algorithms (demographic parity and equal opportunity) do worse than the unconstrained decision threshold!

Modeling Long-Term Fairness

"Delayed Impact of Fair Machine Learning" (Liu et al., 2018)

- Fairness of a single decision (e.g. loan approval) does not consider the effect on the overall group
- So, what's the effect of different thresholds on populations on the *change* in *group credit score*?
- We can figure this out by calculating the effect of false positives and false negatives on credit score...
- "Fair" algorithms (demographic parity and equal opportunity) do worse than the unconstrained decision threshold!

Modeling Long-Term Fairness

"Delayed Impact of Fair Machine Learning" (Liu et al., 2018)

- Fairness of a single decision (e.g. loan approval) does not consider the effect on the overall group
- So, what's the effect of different thresholds on populations on the *change* in *group credit score?*
- We can figure this out by calculating the effect of false positives and false negatives on credit score...
- "Fair" algorithms (demographic parity and equal opportunity) do worse than the unconstrained decision threshold!

"Delayed Impact of Fair Machine Learning" (Liu et al., 2018)

- Fairness of a single decision (e.g. loan approval) does not consider the effect on the overall group
- So, what's the effect of different thresholds on populations on the *change* in *group credit score?*
- We can figure this out by calculating the effect of false positives and false negatives on credit score...
- "Fair" algorithms (demographic parity and equal opportunity) do worse than the unconstrained decision threshold!

Modeling Long-Term Fairness

"Delayed Impact of Fair Machine Learning" (Liu et al., 2018)

- Fairness of a single decision (e.g. loan approval) does not consider the effect on the overall group
- So, what's the effect of different thresholds on populations on the *change* in *group credit score?*
- We can figure this out by calculating the effect of false positives and false negatives on credit score...
- "Fair" algorithms (demographic parity and equal opportunity) do worse than the unconstrained decision threshold!

References

Chouldechova, Alexandra (June 2017). “Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments”. en. In: *Big Data* 5.2, pp. 153–163. ISSN: 2167-6461, 2167-647X. DOI: `10.1089/big.2016.0047`.

Corbett-Davies, Sam et al. (Jan. 2017). “Algorithmic Decision Making and the Cost of Fairness”. In: *arXiv:1701.08230 [cs, stat]*. DOI: `10.1145/3097983.309809`. arXiv: `1701.08230 [cs, stat]`.

Dwork, Cynthia et al. (2012). “Fairness Through Awareness”. In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS ’12. New York, NY, USA: ACM, pp. 214–226. ISBN: 978-1-4503-1115-1. DOI: 10.1145/2090236.2090255.

Hardt, Moritz, Eric Price, and Nathan Srebro (Oct. 2016). “Equality of Opportunity in Supervised Learning”. In: *arXiv:1610.02413 [cs]*. arXiv: 1610.02413 [cs].

Hébert-Johnson, Úrsula et al. (Nov. 2017). “Calibration for the (Computationally-Identifiable) Masses”. In: *arXiv:1711.08513 [cs, stat]*. arXiv: 1711.08513 [cs, stat].

Kearns, Michael et al. (Nov. 2017). “Preventing Fairness Gerrymandering: Auditing and Learning for Subgroup Fairness”. In: *arXiv:1711.05144 [cs]*. arXiv: 1711.05144 [cs].

- Kilbertus, Niki et al. (June 2017). “Avoiding Discrimination through Causal Reasoning”. In: *arXiv:1706.02744 [cs, stat]*. arXiv: `1706.02744 [cs, stat]`.
- Kim, Michael P., Amirata Ghorbani, and James Zou (May 2018). “Multiaccuracy: Black-Box Post-Processing for Fairness in Classification”. In: *arXiv:1805.12317 [cs, stat]*. arXiv: `1805.12317 [cs, stat]`.
- Kusner, Matt J. et al. (Mar. 2017). “Counterfactual Fairness”. In: *arXiv:1703.06856 [cs, stat]*. arXiv: `1703.06856 [cs, stat]`.
- Liu, Lydia T. et al. (Mar. 2018). “Delayed Impact of Fair Machine Learning”. In: *arXiv:1803.04383 [cs, stat]*. arXiv: `1803.04383 [cs, stat]`.

Russell, Chris et al. (2017). “When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness”. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon et al. Curran Associates, Inc., pp. 6414–6423.

Simoiu, Camelia, Sam Corbett-Davies, and Sharad Goel (Sept. 2017). “The Problem of Infra-Marginality in Outcome Tests for Discrimination”. en. In: *The Annals of Applied Statistics* 11.3, pp. 1193–1216. ISSN: 1932-6157. DOI: 10.1214/17-AOAS1058.

Zhang, Zhe and Daniel B. Neill (Nov. 2016). “Identifying Significant Predictive Bias in Classifiers”. In: *arXiv:1611.08292 [cs, stat]*. arXiv: 1611.08292 [cs, stat].