Counterfactual Prediction and Fairness in Risk Assessment Tools

Alan Mishler (amishler@stat.cmu.edu), Edward H. Kennedy (edward@stat.cmu.edu)

Department of Statistics & Data Science, Carnegie Mellon University

Introduction

Risk Assessment Tools (RATs) predict **observed outcomes**. Evaluations of the fairness of these instruments also rely on observed outcomes.

Problem: Observed outcomes confound *risk* and the *effect of interventions*.

- Interventions reflect risk *and* affect outcomes.

$\Longrightarrow$ RATs have limited use to decision-makers.

$\Longrightarrow$ Hard to evaluate performance or fairness.

Solution: Use **potential outcomes** under various intervention(s) instead.

$\Longrightarrow$ More useful information for decision-makers.

$\Longrightarrow$ More sensible definitions of fairness.

Example: COMPAS Scores

The COMPAS recidivism prediction tool (Northpointe, inc.) predicts whether criminal defendants will be arrested for a future crime within 2 years.

Notation:

Observed variables

$S \in \{0,1\}$ = COMPAS score, 1 = "high risk"

$A \in \{0,1\}$ = Incarceration indicator

$Y \in \{0,1\}$ = Outcome: recidivism indicator

$\mathbf{X}$ = Covariates: gender, prior convictions, etc.

R = race (b = black, w = white)

Counterfactual variables

$S^{R=r} \in \{0,1\}$ = Potential score under race r

$Y^{A=a} \in \{0,1\}$ = Outcome under treatment a

$\mu_a(X) = P(Y^{A=a} = 1|X)$

$\Longrightarrow$ $\mu_0(X)$ = P(recidivism without intervention)

COMPAS aims to predict $P(Y = 1|X)$. Why is this problematic?

Plausible scenario:

- Defendants with high $\mu_0(X)$ get incarcerated.
- Defendants with low $\mu_0(X)$ don't.
- Then low $\mu_0(X)$ implies higher recidivism rates.

Result:

- Accuracy maximized by assigning $S = 1$ for low $\mu_0(X)$!
- Not clear how decision-makers can use this.
- Doesn't predict what happens if the prediction itself is used for decision-making.
- Unclear how to evaluate fairness.

COMPAS: Counterfactual Reanalysis

Data (ProPublica, 2016)

- 5,278 arrest cases from Broward County, FL
- 3,175 black; 2,103 white
- Jail durations, recidivism outcomes, covariates
- 2 COMPAS scores:

  General recidivism risk (G)

  Violent recidivism risk (V)

Previous Analyses

- ProPublica (2016): Found different error rates and predicted score ratios based on race.
- Northpointe (2016): Found scores show predictive parity for whites and blacks.

Analysis 1: Do scores differ by race?

ProPublica (logistic regression):

$$\frac{\hat{P}(S = 1|X = x, R = b)}{\hat{P}(S = 1|X = x, R = w)}$$

for an arbitrarily chosen x.

Reanalysis (doubly robust estimator):

$$\frac{\hat{P}(S^{R=b} = 1)}{\hat{P}(S^{R=w} = 1)}$$

Assumes $R \perp\!\!\!\perp S^{R=r}|X$

Results:

	(G)	(V)
ProPublica	1.45	1.77
Reanalysis	1.21	1.29

Counterfactual scores show much less bias.

Analysis 2: False Positive Rates

ProPublica:

$$\hat{P}(S = 1|Y = 0, R = r)$$

Reanalysis (doubly robust estimator):

$$\hat{P}(S = 1|Y^{A=0} = 0, R = r)$$

Assumes $A \perp\!\!\!\perp Y^{A=a}|S = 1, R = r, X$

Results:

	(G)		(V)	
	White	Black	White	Black
ProPublica	0.23	0.45	0.18	0.38
Reanalysis	0.24	0.43	0.17	0.30

Counterfactual scores show similar bias.

Analysis 3: False Negative Rates

ProPublica:

$$\hat{P}(S = 0|Y = 1, R = r)$$

Reanalysis (doubly robust estimator):

$$\hat{P}(S = 0|Y^{A=0} = 1, R = r)$$

Assumes $A \perp\!\!\!\perp Y^{A=a}|S = 0, R = r, X$.

Results:

	(G)		(V)	
	White	Black	White	Black
ProPublica	0.48	0.28	0.63	0.38
Reanalysis	0.51	0.29	0.71	0.45

Counterfactual scores show similar bias.

Analysis 4: Positive Predictive Values

Northpointe:

$$\hat{P}(Y = 1|S = 1, R = r)$$

Reanalysis (doubly robust estimator):

$$\hat{P}(Y^{A=0} = 1|S = 1, R = r)$$

Assumes $A \perp\!\!\!\perp Y^{A=a}|S = 1, R = r, X$.

Results:

	(G)		(V)	
	White	Black	White	Black
Northpointe	0.59	0.63	0.17	0.21
Reanalysis	0.65	0.69	0.14	0.18

COMPAS Reanalysis Conclusions

- Bias in score ratios, but much less than in ProPublica results.
- Error rates (FPR, FNR) similar to ProPublica results.
- Approximate predictive parity, similar to Northpointe results.
- Slightly higher PPVs for blacks than whites.
- General recidivism: Slightly higher PPVs than from observed outcomes.

General Conclusions

Risk prediction should start with well-defined notion of risk.

Current standard: $E(Y|X)$

- Merely predicts outcomes under current system.
- Doesn't consider effects of interventions.
- Not clear how decision-maker should use this.
- Not clear how to evaluate fairness.

Our proposal: $E(Y^{A=a}|X)$ under various interventions a, including no intervention ($a = 0$).

- Predicts outcomes under different interventions.
- Allows decision maker to weigh interventions.
- Yields sensible notions of fairness.

Measure	Current	Proposed
FPR	$P(S = 1\|Y = 0)$	$P(S = 1\|Y^{A=0} = 0)$
FNR	$P(S = 0\|Y = 1)$	$P(S = 0\|Y^{A=0} = 1)$
PPV	$P(Y = 1\|S = 1)$	$P(Y^{A=0} = 1\|S = 1)$

References

Julia Angwin, Jeff Larson, Surya Mattu, & Lauren Kirchner. How we analyzed the COMPAS recidivism algorithm. 2016. (ProPublica)

William Dieterich, Christina Mendoza, & Tim Brennan. COMPAS risk scales: Demonstrating accuracy equity and predictive parity. 2016. (Northpointe)