

Ethische Fragen bei Entscheidungsalgorithmen

Ideen der Informatik

Kurt Mehlhorn



Januar 2019

mp max planck institut
informatik

SIC Saarland
Informatics Campus

- Qualität und Fairness:
 - Qualität: entspricht die Entscheidung der Wahrheit?
 - Fairness: Gleichbehandlung verschiedener Gruppen.
 - Werden diese Frage diskutieren an Hand von COMPAS, eine Software, die abschätzt, ob Straftäter rückfällig werden.
 - Literatur:
 - Angwin, Julia; Larson, Jeff (2016-05-23). "Machine Bias". ProPublica.
 - Alexandra Chouldechova: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments.
- Wahl zwischen zwei Übeln.
 - Das Trolley-Dilemma
 - Literatur:

<https://autonoblog.de/2019/03/22/ethik-autonomes-fahren-ii-trolley-probleme/>

Excerpt from Wikipedia: **COMPAS** = Decision Support Tool to assess the likelihood of a defendant becoming a recidivist.

The Violent Recidivism score is meant to predict violent offenses following release. The risk score is computed from the following data: age, age-at-first-arrest, history-violence, vocation education, history of non-compliance. The function is learned from past data.

Critique with respect to fairness and accuracy: ProPublica: “blacks are almost twice as likely as whites to be labeled a higher risk but not actually re-offend,” whereas COMPAS “makes the opposite mistake among whites: They are much more likely than blacks to be labeled lower-risk but go on to commit other crimes.”

They also found that only 20 percent of people predicted to commit violent crimes actually went on to do so.

Kreditvergabe: Schufa benutzt automatische Verfahren zur Vorhersage von Kreditwürdigkeit.

Ein Beispiel für Compas

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	1449	362	1811
	Y = 1	1141	744	1885
		2690	1006	3996

Compas hatte die Daten von 3996 Straftätern. Davon wurden 1885 rückfällig ($Y = 1$), 1811 wurden nicht rückfällig (Zeile $Y = 0$).

Vorhersagen der Software: $H = 1$ bedeutet "Vorhersage = wird rückfällig".

- Bei den nicht Rückfälligen (Zeile $Y = 0$): 1449 mal wurde korrekt vorhergesagt ($H = 0$), 362 mal inkorrekt ($H = 1$).
- Bei den Rückfälligen (Zeile $Y = 1$): 1141 wurde nicht korrekt vorhergesagt und 744 wurde korrekt vorhergesagt.
- Die Vorhersage $H = 1$ ist in 744 von 1006 Fällen richtig, das sind etwa 74%.
- Die Rate der falsch Positiven ist 362 von 1811, das sind etwa 20%.
- Die Rate der falsch Negativen ist 1141 von 1885, das sind etwa 60%.



Ein Beispiel für Compas

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	1449	362	1811
	Y = 1	1141	744	1885
		2690	1006	3996

Compas hatte die Daten von 3996 schwarzen Straftätern. Davon wurden 1885 rückfällig ($Y = 1$), 1811 wurden nicht rückfällig (Zeile $Y = 0$). Prozentsatz der Rückfälligen etwa 51%.

- Die Vorhersage $H = 1$ ist in 744 von 1006 Fällen richtig, das sind etwa 74%.
- Die Rate der falsch Positiven ist 362 von 1811, das sind etwa 20%.
- Die Rate der falsch Negativen ist 1141 von 1885, das sind etwa 60%.

- **Aussage der Firma: Unsere Software diskriminiert nicht:** Die positive Vorhersagekorrektheit ist bei Weißen und Schwarzen gleich 74%.
- **ProPublica: Die Software diskriminiert:** Die Rate der falsch Positiven ist bei Schwarzen 20% und bei Weißen nur 10%. Die Rate der falsch Negativen ist dagegen bei den Weißen deutlich höher.

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	1346	150	1496
	Y = 1	691	266	957
		2037	416	2453

Compas hatte die Daten von 2453 weißen Straftätern. Davon wurden 957 rückfällig ($Y = 1$), 1496 wurden nicht rückfällig (Zeile $Y = 0$). Prozentsatz der Rückfälligen = 39%.

- Die Vorhersage $H = 1$ ist in 266 von 416 Fällen richtig, das sind etwa 74%.
- Die Rate der falsch Positiven ist 150 von 1496, das sind etwa 10%.
- Die Rate der falsch Negativen ist 691 von 957, das sind etwa 72%.

Wer hat Recht?



Ein Beispiel für Compas

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	1449	362	1811
	Y = 1	1141	744	1885
		2690	1006	3996

Compas hatte die Daten von 3996 schwarzen Straftätern. Davon wurden 1885 rückfällig ($Y = 1$), 1811 wurden nicht rückfällig (Zeile $Y = 0$). Prozentsatz der Rückfälligen etwa 51%.

- Die Vorhersage $H = 1$ ist in 744 von 1006 Fällen richtig, das sind etwa 74%.
- Die Rate der falsch Positiven ist 362 von 1811, das sind etwa 20%.
- Die Rate der falsch Negativen ist 1141 von 1885, das sind etwa 60%.

- **Aussage der Firma: Unsere Software diskriminiert nicht:** Die positive Vorhersagekorrektheit ist bei Weißen und Schwarzen gleich 74%.

- **ProPublica: Die Software diskriminiert:** Die Rate der falsch Positiven ist bei Schwarzen 20% und bei Weißen nur 10%. Die Rate der falsch Negativen ist dagegen bei den Weißen deutlich höher.

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	1346	150	1496
	Y = 1	691	266	957
		2037	416	2453

Compas hatte die Daten von 2453 weißen Straftätern. Davon wurden 957 rückfällig ($Y = 1$), 1496 wurden nicht rückfällig (Zeile $Y = 0$). Prozentsatz der Rückfälligen = 39%.

- Die Vorhersage $H = 1$ ist in 266 von 416 Fällen richtig, das sind etwa 74%.
- Die Rate der falsch Positiven ist 150 von 1496, das sind etwa 10%.
- Die Rate der falsch Negativen ist 691 von 957, das sind etwa 72%.

Wer hat Recht?



Ein Beispiel für Compas

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	1449	362	1811
	Y = 1	1141	744	1885
		2690	1006	3996

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	1346	150	1496
	Y = 1	691	266	957
		2037	416	2453

Compas hatte die Daten von 3996 schwarzen Straftätern. Davon wurden 1885 rückfällig ($Y = 1$), 1811 wurden nicht rückfällig (Zeile $Y = 0$). Prozentsatz der Rückfälligen etwa 51%.

Compas hatte die Daten von 2453 weißen Straftätern. Davon wurden 957 rückfällig ($Y = 1$), 1496 wurden nicht rückfällig (Zeile $Y = 0$). Prozentsatz der Rückfälligen = 39%.

- Die Vorhersage $H = 1$ ist in 744 von 1006 Fällen richtig, das sind etwa 74%.
- Die Rate der falsch Positiven ist 362 von 1811, das sind etwa 20%.
- Die Rate der falsch Negativen ist 1141 von 1885, das sind etwa 60%.
- **Aussage der Firma: Unsere Software diskriminiert nicht:** Die positive Vorhersagekorrektheit ist bei Weißen und Schwarzen gleich 74%.
- **ProPublica: Die Software diskriminiert:** Die Rate der falsch Positiven ist bei Schwarzen 20% und bei Weißen nur 10%. Die Rate der falsch Negativen ist dagegen bei den Weißen deutlich höher.
- Die Vorhersage $H = 1$ ist in 266 von 416 Fällen richtig, das sind etwa 74%.
- Die Rate der falsch Positiven ist 150 von 1496, das sind etwa 10%.
- Die Rate der falsch Negativen ist 691 von 957, das sind etwa 72%.



Ein Beispiel für Compas

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	1449	362	1811
	Y = 1	1141	744	1885
		2690	1006	3996

Compas hatte die Daten von 3996 schwarzen Straftätern. Davon wurden 1885 rückfällig ($Y = 1$), 1811 wurden nicht rückfällig (Zeile $Y = 0$). Prozentsatz der Rückfälligen etwa 51%.

- Die Vorhersage $H = 1$ ist in 744 von 1006 Fällen richtig, das sind etwa 74%.
- Die Rate der falsch Positiven ist 362 von 1811, das sind etwa 20%.
- Die Rate der falsch Negativen ist 1141 von 1885, das sind etwa 60%.
- **Aussage der Firma: Unsere Software diskriminiert nicht:** Die positive Vorhersagekorrektheit ist bei Weißen und Schwarzen gleich 74%.
- **ProPublica: Die Software diskriminiert:** Die Rate der falsch Positiven ist bei Schwarzen 20% und bei Weißen nur 10%. Die Rate der falsch Negativen ist dagegen bei den Weißen deutlich höher.

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	1346	150	1496
	Y = 1	691	266	957
		2037	416	2453

Compas hatte die Daten von 2453 weißen Straftätern. Davon wurden 957 rückfällig ($Y = 1$), 1496 wurden nicht rückfällig (Zeile $Y = 0$). Prozentsatz der Rückfälligen = 39%.

- Die Vorhersage $H = 1$ ist in 266 von 416 Fällen richtig, das sind etwa 74%.
- Die Rate der falsch Positiven ist 150 von 1496, das sind etwa 10%.
- Die Rate der falsch Negativen ist 691 von 957, das sind etwa 72%.

Wer hat Recht?



Eine Erklärung

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	TN	FP	$\#(Y = 0)$
	Y = 1	FN	TP	$\#(Y = 1)$
		$\#(H = 0)$	$\#(H = 1)$	S

S = Größe der Grundmenge, $\#(Y = 0)$ und $\#(Y = 1)$ sind gegeben.

- Die Vorhersage H = 1 ist richtig mit Bruchteil $PPV = \frac{TP}{\#(H=1)}$.
- Die Rate der falsch Positiven ist $FPR = \frac{FP}{\#(Y=0)}$.
- Die Rate der falsch Negativen ist $FNR = \frac{FN}{\#(Y=1)}$.

Die Größen sind nicht unabhängig voneinander. Je zwei bestimmen die Dritte. Genauer: Sei $p = \#(Y = 1)/S$. Dann

$$FPR = \frac{p}{1-p} \frac{1-PPV}{PPV} (1 - FNR).$$

Merke: Wenn die p -Werte für die beiden Populationen (Schwarze und Weiße) unterschiedlich sind, dann können die drei anderen Werte nicht sämtlich gleich sein.

Eine Erklärung

		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	TN	FP	$\#(Y = 0)$
	Y = 1	FN	TP	$\#(Y = 1)$
		$\#(H = 0)$	$\#(H = 1)$	S

S = Größe der Grundmenge, $\#(Y = 0)$ und $\#(Y = 1)$ sind gegeben.

- Die Vorhersage H = 1 ist richtig mit Bruchteil $PPV = \frac{TP}{\#(H=1)}$.
- Die Rate der falsch Positiven ist $FPR = \frac{FP}{\#(Y=0)}$.
- Die Rate der falsch Negativen ist $FNR = \frac{FN}{\#(Y=1)}$.

Die Größen sind nicht unabhängig voneinander. Je zwei bestimmen die Dritte. Genauer: Sei $p = \#(Y = 1)/S$. Dann

$$FPR = \frac{p}{1-p} \frac{1-PPV}{PPV} (1 - FNR).$$

Merke: Wenn die p -Werte für die beiden Populationen (Schwarze und Weiße) unterschiedlich sind, dann können die drei anderen Werte nicht sämtlich gleich sein.

Die Rechnung

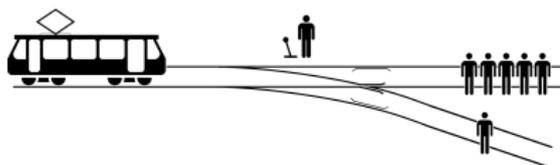
		Vorhersage		
		H = 0	H = 1	
Wahrheit	Y = 0	TN	FP	$\#(Y = 0)$
	Y = 1	FN	TP	$\#(Y = 1)$
		$\#(H = 0)$	$\#(H = 1)$	S

S = Größe der Grundmenge, $p = \#(Y = 1)/S$, $PPV = \frac{TP}{\#(H=1)}$,
 $FPR = \frac{FP}{\#(Y=0)}$, $FNR = \frac{FN}{\#(Y=1)}$. Dann ist

$$FPR = \frac{p}{1-p} \frac{1-PPV}{PPV} (1-FNR).$$

$$\begin{aligned} \frac{p}{1-p} \frac{1-PPV}{PPV} (1-FNR) &= \frac{\#(Y=1)/S}{\#(Y=0)/S} \cdot \frac{FP/\#(H=1)}{TP/\#(H=1)} \cdot \left(1 - \frac{FN}{\#(Y=1)}\right) \\ &= \frac{\#(Y=1)}{\#(Y=0)} \cdot \frac{FP}{TP} \cdot \frac{TP}{\#(Y=1)} \\ &= \frac{FP}{\#(Y=0)} \\ &= FPR. \end{aligned}$$

Entscheidung zwischen zwei Übeln



Schematische Darstellung eines Trolley-Problems (copyright: Zapyon, CC BY-SA 4.0)

- Prototyp der Wahl zwischen zwei Übeln. Viele Varianten (Greise und Kinder, Fat Man auf der Brücke, ...).
- War ein Problem für Philosophen. Wird es durch autonomes Fahren ein Problem der Realität?
- Antworten hängen stark vom Kulturkreis und der Variante ab.
- Eine ausführliche Diskussion finden Sie unter <https://autonoblog.de/2019/03/22/ethik-autonomes-fahren-ii-trolley-probleme/>