Algoritmisk rettferdighet

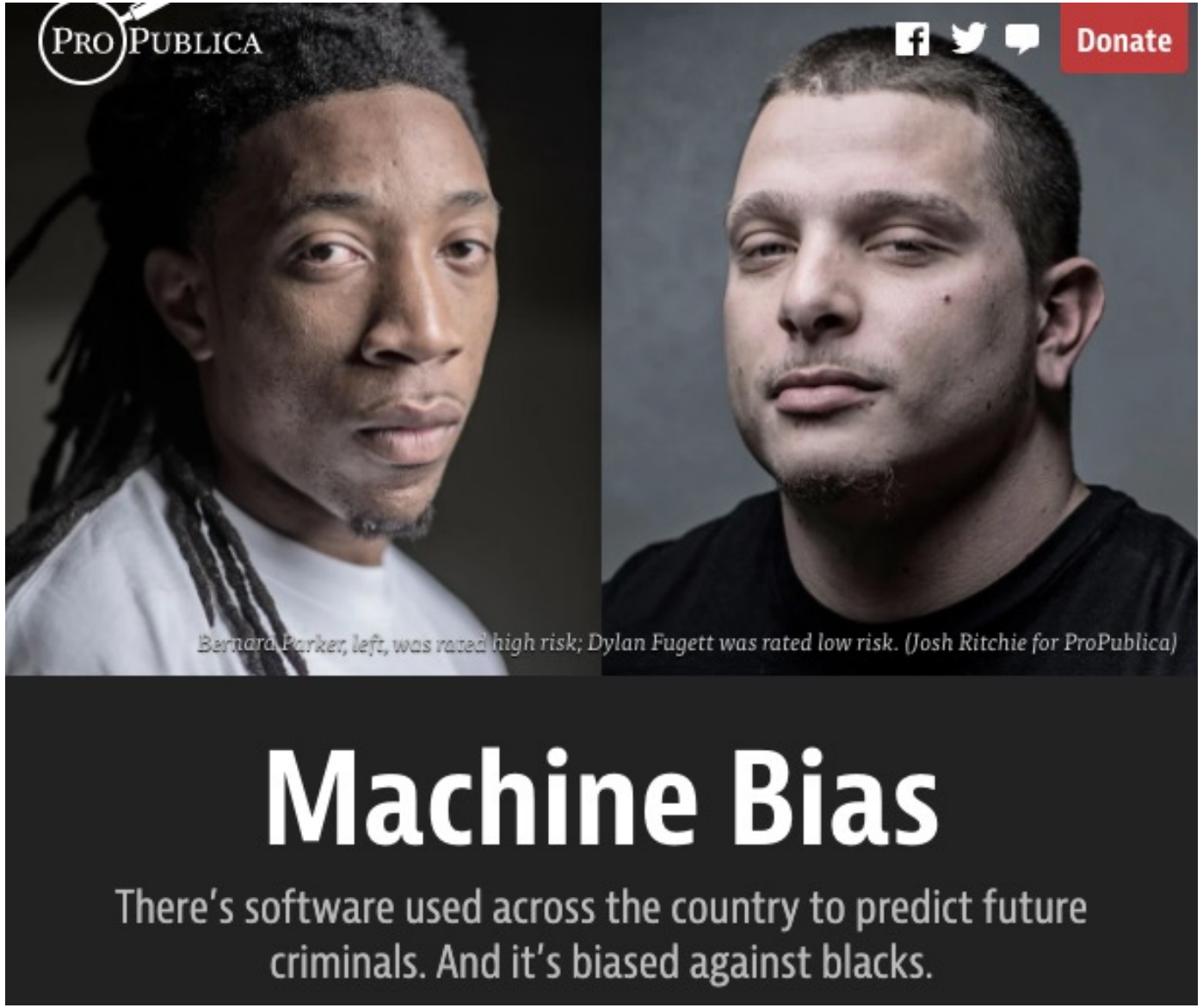
Aksel Braanen Sterri

akselbst@gmail.com

14. Februar 2023

Problemet

- Maskinlæring tas i bruk på flere områder. Mange er bekymret.
- Overordnede spørsmål:
  - Bør vi være bekymret?
  - Hva bør vi være bekymret for?
  - Hva bør vi gjøre?
- Mål:
  - Sette debatten om «algoritmisk rettferdighet» i en større kontekst.
  - Vise relevansen av etisk tenkning.



Avgrensning: Seleksjon

- Seleksjon
- Ikke ChatGPT3.
- Observasjoner → Prediksjoner → Seleksjon → Ulikhet.
- Fordele et **gode** eller **onde** basert på **forventninger** om individers atferd.
- Lån, straff, plass på universitetet, jobben, besøk av barnevernet etter fødsel, et nytt organ.
- Men ikke for dårlige formål.

Algoritme er et beslutningshjelpemiddel

Holistisk vurdering

Jobbintervju, straff, besøk av barnevernet

Statistisk vurdering

Opptak til universitet, siling av jobbsøknader basert på kvantitative kriterier

Maskinlæringsvurdering

Kan det kvantifiseres, kan det maskinlæres.

Etiske idealer

Immanuel Kant. Av Sahroe/Shutterstock. Begrenset gjenbruk

Platon (til venstre) og Aristoteles (høyre).

Utsnitt av fresken Skolen i Athen
Av Rafael, 1509.
Lisens: Falt i det fri (Public domain)

Pro Publica

- COMPAS - Anklagen: Algoritmene er «biased mot svarte».
- **Falsk positiv-falsk negativ ratio:** Høyere andel falske positive for svarte enn for hvite og høyere andel falske negative for hvite enn svarte.
- Men ifølge Northpointe (COMPAS): Algoritmen er ikke “biased” fordi prediksjonene var like **treffsikre** for begge grupper (Dietrich et al. 2016)
- Anthony Flores et al. (2016): Algoritmen er like **kalibrert** for begge grupper: For hver risikoskår var det omtrent en like stor andel svarte som hvite som begikk en ny kriminell handling.
- Dette er bare tre av (minst) 11 definisjoner av statistisk algoritmisk rettferdighet.
- Umulighetsteoremer.
- Hvordan forholder vi oss til uenigheten?

Test-case: Kron og mynt

Mynter med ulike “vekter” (noen har >0.5 for kron, andre har >0.5, mellom 0 og 1, ingen mynter har akkurat 0.5)

Randomiserer folk til ulike mynter.

Randomiser deretter folk og mynt til to rom A og B (gruppetilhørighet)

Vi skal predikere for hver person om deres mynt blir mynt eller kron.

Vi vet hvem som har hvilke mynter og alle mynter er korrekt merket med myntens «vekt».

Hvordan vil en rettferdig algoritme se ut?

Et forslag

- Binær: For hver person, ta deres mynt og les av vekten. Hvis det står «x», gi en risiko-skår lik x til personen med denne mynten. Hvis $x > 0.5$ prediker at personen kommer til å få kron. Hvis $x < 0.5$, prediker at personen kommer til å få mynt. (Rom er ikke relevant.)

Fordelingen

Rom A

- 12 stk med mynter som har 0.75 i vekt
- 8 stk med 0.125 i vekt

Rom B

- 10 stk med 0.6
- 10 stk med 0.4

Prediksjon

Rom A:

- Blant de 12 med 0.75 vil det være 9 med kron og 3 med mynt
- Blant de 8 med 0.125 vil det være 1 kron og 7 mynt

Rom B:

- Blant de 10 med 0.6 vil det være 6 kron og 4 mynt
- Blant de 10 med 0.4 vil det være 4 kron og 6 mynt

Resultat

Rom A:

- Blant de 12 med 0.75 vil det være 9 med kron og 3 med mynt
- Blant de 8 med 0.125 vil det være 1 kron og 7 mynt

Rom B:

- Blant de 10 med 0.6 vil det være 6 kron og 4 mynt
- Blant de 10 med 0.4 vil det være 4 kron og 6 mynt

Test mot kriteriene for rettferdighet

- Samtlige standard-definisjoner av statistisk rettferdighet med unntak av **kalibrering** impliserer (i dette eksemplet) at algoritmen er urettferdig. Men dette stemmer ikke.
- Ett eksempel:
  - Den falske-positive ratio i rom A er 3/10
  - Den falske-positive ratio i rom B er 4/10
- Lærdommer: «Biasen» eller «urettferdigheten» skyldes at algoritmen gir mer presise prediksjoner jo mer mynten er vektet. Jo mindre den er vektet, jo mer upresis blir algoritmen. Jo mer feil gjør den.

Implikasjoner for COMPAS?

- Betyr det at COMPAS ikke er urettferdig?

Begrep og forståelse

- Skille mellom *begrep (concept)* og *forståelse* (conception)?
- Begrep: Hva er det vi alle kan enes om at rettferdighet betyr? (Aristoteles)
- Forståelse: Den beste forståelsen av rettferdighet (refleksiv likevekt)
- Hvilken forståelse av rettferdighet er best?

Flere former for rettferdighet

- Prosedural og substantiv rettferdighet (Robert Nozick og Rawls)
- Ulike former for rettferdighet (Aristoteles og fløyta)
- Omfordelingsrettferdighet og skyld-rettferdighet (velferd versus straff)
- Global eller lokal rettferdighet (Rawls og Walzer)
- *Snever*: *Ikke* legge vekt på «irrelevant» info, legger vekt på *relevante* forskjeller, og behandler like tilfeller likt.
- *Vid*: Ikke gjør samfunnet mer urettferdig eller gjør samfunnet mer rettferdig.

Case: Universitet

Iril Flatebø deltar på Honours-programmet til Universitetet i Oslo, og visste at debatten om elitestudenter ville komme. Foto: SIr Øverland Eriksen Foto: Siri Øverland Eriksen

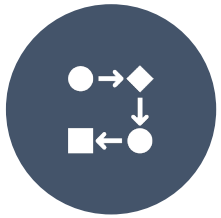
Første dag på honours-studiet: «Jeg ser ikke på meg selv som elitestudent»

Maskin-seleksjon

STEG 1: SAMLE OBSERVASJONER OM TIDLIGERE STUDENTER

STEG 2: HVA ER SUKSESS-KRITERIENE?

STEG 3: MASKINEN FINNER SAMMENHENGER MELLOM STUDENTERS EGENSKAPER OG MÅLOPPNÅELSE



STEG 4: TESTE MASKINEN PÅ ANDRE OBSERVASJONER (ET ANNET DATASET)

STEG 5: SELEKSJONSALGORITMEN ER DEN SOM ER BEST TIL Å FORUTSI FREMTIDIG SUKSESS

STEG 6: SAMLE OBSERVASJONER OM NYE SØKERE TIL UNIVERSITETET

STEG 7: SELEKSJONSALGORITMEN RANGERER KANDIDATENE ETTER FORVENTET SUKSEESS

STEG 8: DE KANDIDATENE SOM ER RANGERT HØYEST FÅR TILBUD FØRST.

Umiddelbare tanker

HVA ER UKLART?

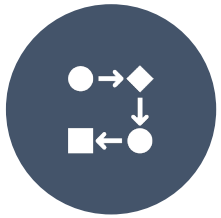
FORDELER OG ULEMPER MED Å BRUKE MASKINLÆRINGSALGORITMER TIL Å SELEKTERE STUDENTER VED OPPTAK TIL UNIVERSITET?

Maskin-seleksjon

STEG 1: SAMLE OBSERVASJONER OM TIDLIGERE STUDENTER

STEG 2: HVA ER SUKSESS, IFØLGE UNIVERSITETS-ADMINISTRASJONEN?

STEG 3: MASKINEN FINNER SAMMENHENGER MELLOM STUDENTERS EGENSKAPER OG MÅLOPPNÅELSE



STEG 4: TESTE MASKINEN PÅ ANDRE OBSERVASJONER (ET ANNET DATASET)

STEG 5: SELEKSJONSALGORITMEN ER DEN SOM ER BEST TIL Å FORUTSI FREMTIDIG SUKSESS

STEG 6: SAMLE OBSERVASJONER OM NYE SØKERE TIL UNIVERSITETET

STEG 7: SELEKSJONSALGORITMEN RANGERER KANDIDATENE ETTER FORVENTET SUKSEESS

STEG 8: DE KANDIDATENE SOM ER RANGERT HØYEST FÅR TILBUD FØRST.

Etisk relevante steg

SAMLE OBSERVASJONER OM TIDLIGERE STUDENTER

HVA ER SUKSESS?

ALGORITMEN: HVA ENN PREDIKERER SUKSESS

SAMLE OBSERVASJONER OM NYE SØKERE TIL UNIVERSITETET

RANGERE KANDIDATENE ETTER FORVENTET SUKSEESS

DE BESTE KANDIDATENE FÅR TILBUD FØRST.

- Privatliv
- Hvor pålitelig er historien?
- Strukturell diskriminering påvirker hvem som lykkes
- Gjennomsiktighet i kriteriene (I dag er det gjennomsiktig i kriteriene, men ikke I målet, motsatt med ML)
- Er individuell “suksess” det eneste som betyr noe?
- Reduksjonisme?
- Er alle egenskaper relevante prediktorer?
- Ulikhet mellom grupper?

Urettferdige utslag

- Hva hvis karakterer har urettferdige utslag? Kvinner lykkes mer enn menn? «Etnisk norske» mer enn nordmenn med innvandrerbakgrunn?

- Tilsynelatende nøytrale kriterier kan ha urettferdige implikasjoner. Opptak til politi basert på fysiske kriterier.

- Hva er rettferdig/urettferdig når det kommer til opptak til høyere utdanning?

Ulike former for rettferdighet

1. Individuell rettferdighet: Like tilfeller behandles likt (men hva innebærer det?)
2. Utfalls-likhet: Grupper skal være likt representert (eller tilnærmet likt).
3. Kriteriet skal være like treffsikkert for alle relevante grupper (men hvilke typer feil?).
4. Vi skal gjøre opp for historisk urettferdighet (ligger ofte bak kvotering)

Hvem har krav på plass på universitetet?

Bakoverskuende:

1. De som *fortjener* det (f.eks. har de jobbet hardt)
2. Gjøre opp for historisk urettferdighet ved å favorisere historisk undertrykte grupper

Framoverskuende:

1. De som *forventes* å gjøre en best jobb (på universitetet, i arbeidslivet, i samfunnet?)
2. Forhindre nepotisme
3. Øke representasjon av under-representerte grupper
4. Bidra til en mer rettferdig fordeling av ferdigheter

- Er karakterer en **treffende indikator** på de målene vi forsøker å oppnå?

Fordeler ved maskinlæring

- Tar «interseksjonalitet på alvor». Er mer «holistisk» enn enn «reduksjonistisk» statistisk analyse.
- Er mer gjennomsiktig enn en holistisk avgjørelse. Kan testes for diskriminering og urettferdighet (gitt at treningsdata og algoritmen er tilgjengelig for andre) f.eks. ved hjelp av «kontrafaktisk» analyse.
- En kan få maskiner til å gjøre arbeid vi ikke trenger å gjøre.
- Presisjon er bra:
  - mer rettferdig (mindre vilkårlig og mindre systematiske skjevheter)
  - Mer presise algoritmer brukt for å bedømme varetekt kan redusere antallet som sitter i fengsel.
- Kvantifisere avveining mellom hensyn som kvotering og organisasjoners primære måloppnåelse.

Kilder til urettferdighet

- Økologisk feilslutning
- Upresise
- Kan reflektere og forsterke eksisterende skjevheter (hvis grunnen til at noen ikke lykkes på jobben og i studier er at de blir diskriminert mot, kan algoritmer opprettholde det).
- Enkelte egenskaper vi ikke skal vurderes etter (kjønn, rase, hva med IQ?)
    - Men: Hvis minoriteter har møtt større barrierer som gjør at deres vgs-snitt undervurderer hvor bra de kommer til å gjøre det på universitetet, burde ikke deres karakterer vektes opp?
- Kan ha urettferdige utslag (ikke alltid ønskelig at forsikringsselskapet har så presis informasjon som mulig?)
- Kan brukes for gale formål