Nieodłączne kompromisy w uczciwym określaniu ryzyka

Seminarium: Informatyka politycznie poprawna

Bartosz Jaśkiewicz, Andrzej Tkaczyk

20/27 października 2020

Słowem wstępu

Inherent Trade-Offs in the Fair Determination of Risk Scores

Jon Kleinberg * Sendhil Mullainathan † Manish Raghavan ‡

Abstract

Recent discussion in the public sphere about algorithmic classification has involved tension between competing notions of what it means for a probabilistic classification to be fair to different groups. We formalize three fairness conditions that lie at the heart of these debates, and we prove that except in highly constrained special cases, there is no method that can satisfy these three conditions simultaneously. Moreover, even satisfying all three conditions approximately requires that the data lie in an approximate version of one of the constrained special cases identified by our theorem. These results suggest some of the ways in which key notions of fairness are incompatible with each other, and hence provide a framework for thinking about the trade-offs between them.

1 Introduction

There are many settings in which a sequence of people comes before a decision-maker, who must make a judgment about each based on some observable set of features. Across a range of applications, these judgments are being carried out by an increasingly wide spectrum of approaches ranging from human expertise to algorithmic and statistical frameworks, as well as various combinations of these approaches.

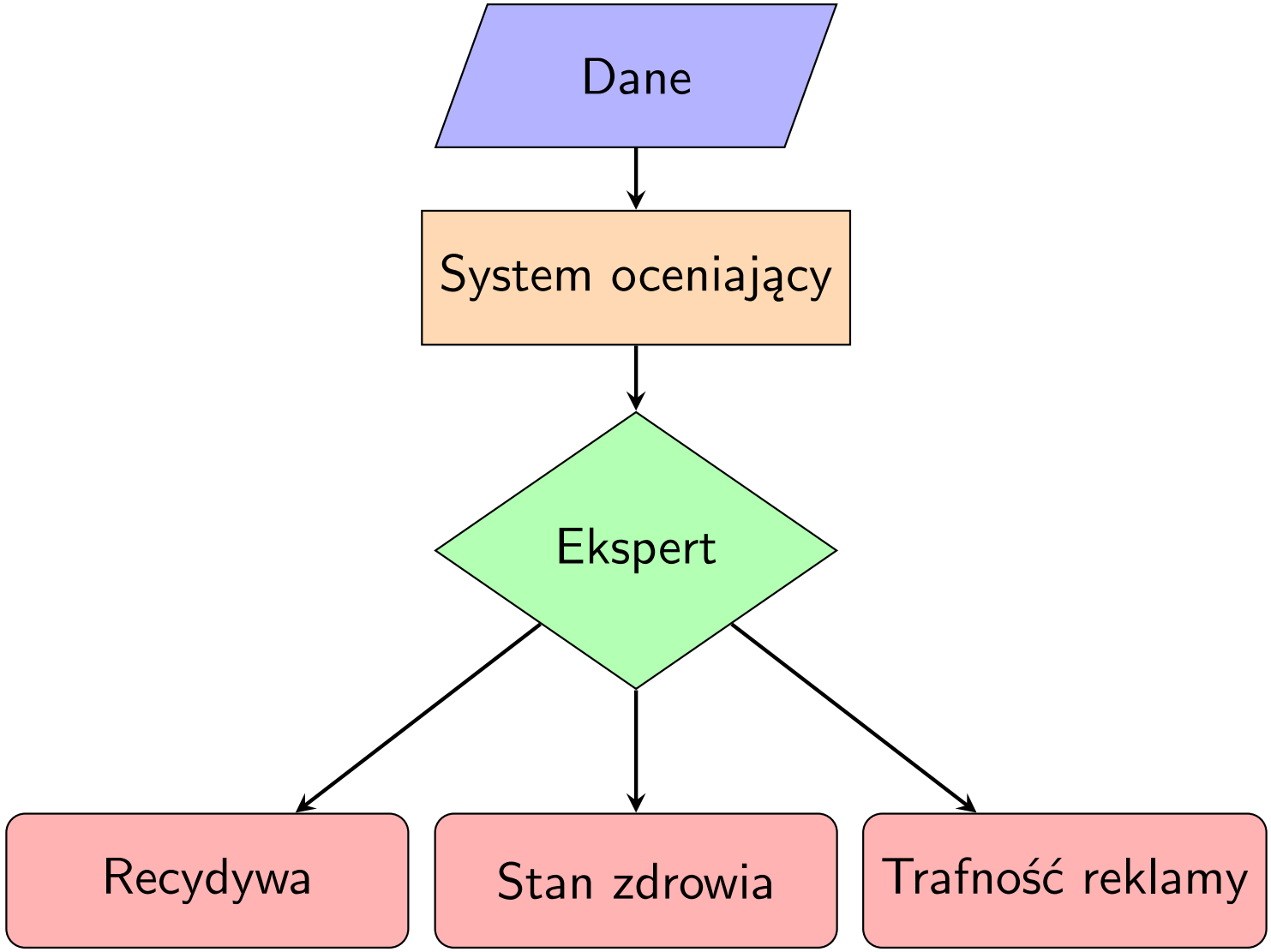
Along with these developments, a growing line of work has asked how we should reason about issues of bias and discrimination in settings where these algorithmic and statistical techniques, trained on large datasets of past instances, play a significant role in the outcome. Let us consider three examples where such issues arise, both to illustrate the range of relevant contexts, and to surface some of the challenges.

A set of example domains. First, at various points in the criminal justice system, including decisions about bail, sentencing, or parole, an officer of the court may use quantitative *risk tools* to assess a defendant's probability of recidivism — future arrest — based on their past history and other attributes. Several recent analyses have asked whether such tools are mitigating or exacerbating the sources of bias in the criminal

Kilka przykładów

- The COMPAS risk tool
  - Czarnoskórzy i białoskórzy skazani
- Reklamy internetowe
  - Reklamy sportów typowo męskich
  - Reklamy mniej płatnych miejsc pracy
- Badania medyczne i diagnostyka
  - Różnice w rozpoznawaniu i przewidywaniu przebiegu chorób

Proces podejmowania decyzji



Podstawowe założenia dla sprawiedliwości

- *Well-calibrated*
- *Balance for the positive class*
- *Balance for the negative class*
- *Statistical parity?*

Chcemy spełnić wszystkie założenia.
Czy tak się da w ogóle zrobić?

Opis modelu

Pozytywne/negatywne instancje

Ludzie posiadający porządaną cechę / nieposiadający porządanej cechy. Pozytywne instancje należą do **klasy pozytywnej**, a negatywne do **klasy negatywnej**.

Wektor cech

Każda osoba ma wektor σ.

p_σ - prawdopodobieństwo, że osoba z wektorem σ jest pozytywną instancją.

Grupa

Każda osoba należy do grupy 1. lub 2.

$a_{\sigma t}$ - prawdopodobieństwo, że osoba z grupy t ma wektor σ.

Przypisywanie ryzyka

1. Tworzymy kubełki z etykietą *score* v_b, gdzie b to nazwa kubełka, a v_b to prawdopodobieństwo, że osoba w kubełku b jest pozytywną instancją.

2. Wkładamy ludzi do poszczególnych kubełków, według jakiejś zasady opartej o ich wektor σ. $X_{\sigma b}$ to prawdopodobieństwo że osoba posiadająca wektor σ trafi do kubełka b.

Warunki sprawiedliwego przypisywania ryzyka

(A) *Calibration within groups* - dla każdej grupy t i dla każdego kubełka b z przypisanym *score* v_b, oczekiwana liczba ludzi z grupy t w kubełku b, należących do klasy pozytywnej powinna być v_b częścią oczekiwanej liczby wszystkich ludzi z grupy t w kubełku b.

(B) *Balance for the negative* - średni *score* dla osób z grupy 1., którzy należą do klasy negatywnej, powinien być taki sam dla osób z grupy 2., którzy należą do klasy negatywnej.

(C) *Balance for the positive class* - średni *score* dla osób z grupy 1., którzy należą do klasy pozytywnej, powinien być taki sam dla osób z grupy 2., którzy należą do klasy pozytywnej.

Specjalne przypadki

- *Perfect prediction* - załóżmy, że dla każdego wektora cech σ, $p_\sigma = 0$, albo $p_\sigma = 1$. Wtedy możemy wrzucić wszystkie osoby z $p_\sigma = 1$ do kubełka b z $v_b = 1$, a osoby z $p_\sigma = 0$ do kubełka b' z $v_{b'} = 0$.

- *Equal base rates* - załóżmy, że średnie p_σ dla obu grup jest takie same. Tworzymy zatem jeden kubełek b z etykietą v_b równą p_σ i umieszczamy w nim wszystkich.

Dla obu powyższych przypadków zasady (A), (B) i (C) są zachowane.

I TYLKO DLA NICH!

Twierdzenie 1

Jeśli instancja problemu przypisania ryzyka spełnia jednocześnie warunki (A), (B) i (C), to instancja ta musi być przypadkiem specjalnym *Perfect prediction* lub *Equal base rates*.

Twierdzenie 2

Istnieje funkcja ciągła $f(x)$ dążąca do 0 przy x dążącym do 0, taka że dla każdego $\epsilon > 0$ i dla każdej instancji problemu przypisania ryzyka, jeśli ta instancja spełnia ϵ-dokładną wersję warunków (A), (B) i (C), to musi być $f(\epsilon)$-dokładną wersją *prerfect prediction* lub $f(\epsilon)$-dokładną wersją *equal base rates*.

Szkic dowodu twierdzenia 1

N_t - liczba wszystkich osób w grupie t.
μ_t - liczba pozytywnych instancji w grupie t

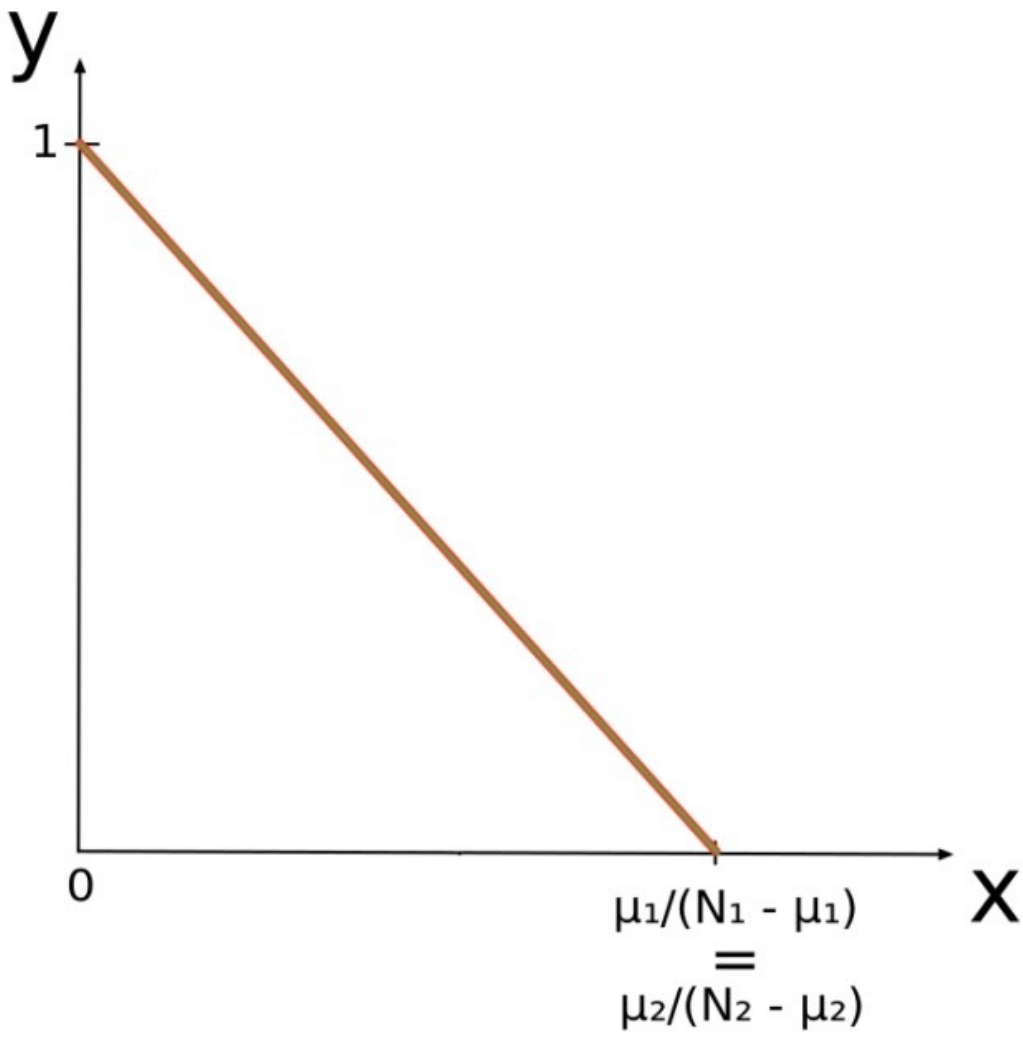
$$\sum_b (a_{t\sigma} \cdot N_t \cdot X_{\sigma b} \cdot v_b) = \mu_t$$

x - średni *score* przypisany osobie z klasy negatywnej.
y - średni *score* przypisany osobie z klasy pozytywnej.

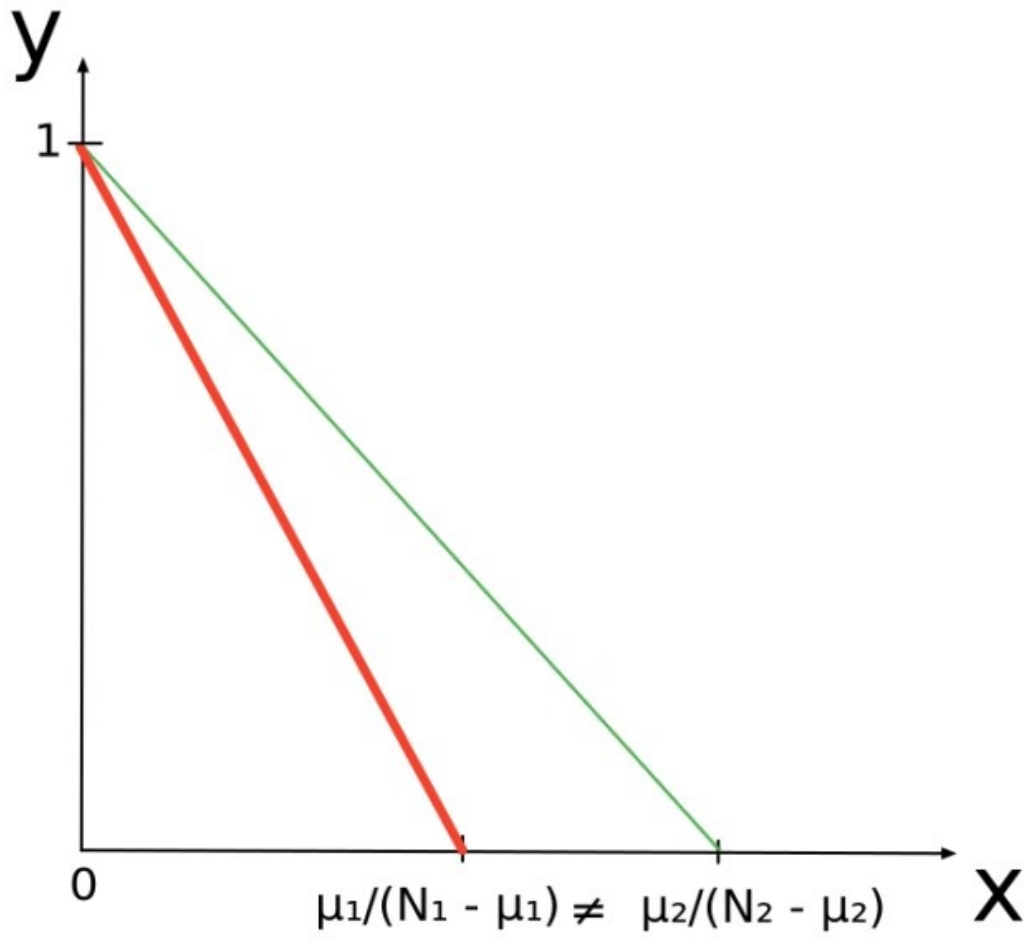
$$(N_t - \mu_t)x + \mu_t y = \mu_t$$

$$y = 1 - \frac{N_t - \mu_t}{\mu_t} x$$

Szkic dowodu twierdzenia 1



Szkic dowodu twierdzenia 1



Minimalizowanie "straty" w *Equal Base Rates*

Cel: jak największy *score* dla instancji pozytywnych i jak najmniejszy dla negatywnych.

Indywidualna strata

Jeśli osoba ma przypisany *score* równy v to jej *indywidualna strata* wynosi v, gdy należy do klasy negatywnej i $1 - v$, gdy należy do klasy pozytywnej.

Strata w grupie

Suma *indywidualnych strat* osób należących do grupy t.

Strata ogólna

Średnia ważona ze strat w każdej grupie.

Minimalizowanie "straty" w *Equal Base Rates*

$$\mathcal{L}_t(X) = n_t^T(I - P)Xv + (\mu_t - n_t^T PXv)$$

$(I - P)$ - macierz prawdopodobieństwa, że osoba z wektorem σ jest negatywną instancją

$n_t^T(I - P)$ - wektor liczby osób z grupy t, które należą do klasy negatywnej i mają dany wektor σ

$n_t^T(I - P)Xv$ - strata z negatywnych instancji w grupie t

$n_t^T PXv$ - suma *score* uzyskana przez pozytywne instancje z grupy t

$(\mu_t - n_t^T PXv)$ - strata z pozytywnych instancji w grupie t

$$n_t^T(I - P)Xv = n_t^T Xv - n_t^T PXv = \mu_t - n_t^T PXv$$

$$\mathcal{L}_t(X) = 2(\mu_t - n_t^T PXv)$$

Minimalizowanie "straty" w *Equal Base Rates*

Weźmy dwie grupy 1 i 2, obie po 100 osób.
Niech 50 osób z grupy 1 ma $p_\sigma = 0.6$, i 50 osób ma $p_\sigma = 0.4$.
Niech 50 osób z grupy 2 ma $p_\sigma = 0.7$, i 50 osób ma $p_\sigma = 0.3$.

γ_t - średni *score* pozytywnych instancji z grupy t.

$$\gamma_1 = \frac{(50 * 0.6) * 0.6 + (50 * 0.4) * 0.4}{50} = 0.52$$

$$\gamma_2 = \frac{(50 * 0.7) * 0.7 + (50 * 0.3) * 0.3}{50} = 0.58$$

Zatem nie został zachowany warunek (C).

Charakteryzacja dobrze skalibrowanych rozwiązań

γ_t - średni *score* pozytywnych instancji z grupy t.

fairness difference $= \gamma_1 - \gamma_2$.

Jeśli *fairness difference* > 0, to model **lekko faworyzuje** grupę 1,
a jeśli *fairness difference* < 0, to model **lekko faworyzuje** grupę 2.

Charakteryzacja dobrze skalibrowanych rozwiązań

Lemat 1

Jeśli grupy 1 i 2 posiadają *Equal base rates*, to dla każdych dwóch nietrywialnych, dobrze skalibrowanych przyporządkowań ryzyka o *fairness difference* d_1 i d_2 i dla każdego $d_3 \in [d1, d2]$ istnieje nietrywialne, dobrze skalibrowane przypisanie ryzyka z *fairness difference* równym d_3.

Wniosek 1

Nietrywialne, uczciwe przyporządkowanie ryzyka istnieje wtedy i tylko wtedy, gdy istnieją dwa nietrywialne, dobrze skalibrowane przyporządkowania ryzyka, z których jedno lekko faworyzuje grupę 1, a drugie lekko faworyzuje grupę 2.

Charakteryzacja dobrze skalibrowanych rozwiązań

Niech $X^{(1)}$ i $X^{(2)}$ to macierze X przyporządkowań ryzyka z *fairness difference* d_1 i d_2 (bez straty ogólności $d_1 < d_2$).

Wybierzmy λ, t. że $\lambda d_1 + (1 - \lambda) d_2 = d_3$.
Widzimy, że $= \frac{d_2 - d_3}{d_2 - d_1}$.

Wtedy $X^{(3)} = [\lambda X^{(1)} \quad (1 - \lambda) X^{(2)}]$.

NP-zupełność nietrywialnego integralnego przyporządkowania ryzyka

Integralne przyporządkowanie ryzyka

Takie przyporządkowanie, że dwie osoby o takim samym wektorze cech σ muszą trafić do tego samego kubełka.

Problem sumy podzbioru

Dane są liczby $w_1, w_2, ..., w_n$ oraz liczba T.
Czy istnieje podzbiór liczb $w_1, w_2, ..., w_n$ o sumie T.

Konkluzja

Źródło

Publikacja autorstwa:
Jon Kleinberg, Sendhil Mullainathan, Manish Raghavan