Fairness and machine learning

Limitations and Opportunities*

Solon Barocas, Moritz Hardt, Arvind Narayanan

Wrocław, 10.11.2020

*LINK NA KOŃCU PREZENTACJI

Fairness criteria

Spośród wielu fairness criteria, dzisiaj wybieramy dwa(Separation i Sufficiency), które są podstawą w swojej dziedzinie, posłużą nam do zaprezentowania ogólnych sposobów osiągania sprawiedliwości przez klasyfikatory.

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

- Oznaczenia:

- A -> sensitive attribute

- Y -> target variable

- R -> classifier lub score

Independence - przypomnienie

- Krótkie przypomnienie czym jest independence, ponieważ większość z tego kryterium pojawiło się na poprzedniej prezentacji dotyczącej publikacji "Fairness and machine learning"* a nam pozostało dopowiedzieć jak nieniejsze kryterium osiągnąć.
- Independence dąży do niezależności wyniku R od wrażliwego atrybutu A.
- Przykład, chcemy aby firma zatrudniała pracowników, nie biorąc pod uwagę ich koloru skóry.

Jak osiągnąć Independence

Są różne sposoby aby osiągnąć kryterium independence budując klasyfikator, my podzielimy je na moment w którym zostają one zaaplikowane do klasyfikatora:

- Pre-processing
- At training time
- Post-processing

Pre-processing

- Dostosowanie danych by były nieskorelowane z wrażliwym atrybutem.
- Zalety
  1. Nie trzeba modyfikować klasyfikatora
  2. Jeżeli zadbamy by kolejne kroki były deterministyczne, nie stracimy niezależności
- Wady:
  1. Osiąga gorsze rezultaty względem dokładności i bycia fair względem kolejnych metod

At training time

- Stosuje ograniczenia w procesie optymalizacji, który konstruuje klasyfikator na podstawie danych uczących.
- Zalety
    1. Może prowadzić do najwyższej jakości rozwiązania
- Wady:
    1. Potrzebuje dostępu do surowych danych
    2. Potrzebuje dostępu do modelu treningowego
    3. Zazwyczaj skupiamy się na jednej klasie by osiągnąć w niej wybitne wyniki, tracimy przez to na uniwersalności

Post-processing

- Dostosowuje wyuczony klasyfikator tak, aby nie był skorelowany z wrażliwym atrybutem. Proces opiera się na korzystaniu z wrażliwego atrybutu oraz dodaniu pewnej randomowości.
- Zalety
  1. Zadziała dla każdego klasyfikatora, niezależnie od jego działania
  2. Nie ma potrzeby ponownego trenowania
  3. Często jest jedyną opcją, bo często mamy jedynie wytrenowany model
- Wady:
  1. Wszystkie powyższe zalety, są wadami metod post-processingu, z którymi wiąże się spadek efektywności

Separation

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

- Separacja polega na przyznaniu, że target variable(Y) jest zależne od sensitive attribute(A) i pozwala na korelację między score(R) a sensitive attribute(A) tak długo jak jest to uzasadnione przez target variable(Y).
- Przykładem mogą być choroby, zależne od koloru skóry, zawody gdzie wymagana jest specyficzna budowa ciała(strażak, chirurg, fizjoterapeuta), oprocentowanie w bankach

Separation

- Definicja: Zmienne losowe (R, A, Y) spełniają warunki separacji jeśli

$$R \perp A \mid Y$$

- W przypadku, gdy R jest klasyfikatorem binarnym, separacja jest równoważna wymaganiu dla wszystkich grup a, b dwóch ograniczeń

$$P\{R = 1 \mid Y = 1, A = a\} = P\{R = 1 \mid Y = 1, A = b\}$$

$$P\{R = 1 \mid Y = 0, A = a\} = P\{R = 1 \mid Y = 0, A = b\}$$

Separation

$$P\{R = 1 \mid Y = 1, A = a\} = P\{R = 1 \mid Y = 1, A = b\}$$

$$P\{R = 1 \mid Y = 0, A = a\} = P\{R = 1 \mid Y = 0, A = b\}$$

- $P\{R = 1 \mid Y = 1\}$ to true positive rate klasyfikatora. Tzn. rate z jakim klasyfikator prawidłowo rozpoznaje pozytywne wystąpienia.
- $P\{R = 1 \mid Y = 0\}$ to false positive klasyfikatora. Tzn. rate z jakim klasyfikator błędnie przypisuje pozytywne wystąpienie do negatywnych instancji.
- Zatem separacja wymaga, aby wszystkie grupy doświadczały tego samego odsetka wyników false negative i tego samego odsetka wyników false positive.

Separation

$$P\{R = 1 \mid Y = 1, A = a\} = P\{R = 1 \mid Y = 1, A = b\}$$
$$P\{R = 1 \mid Y = 0, A = a\} = P\{R = 1 \mid Y = 0, A = b\}$$

- Co ja rozumiem przez napisy wyżej? Otóż, mamy dwie grupy, np. kobiety i mężczyzn. Oraz prestiżową firmę, która wymaga napisania testu z programowania. Jak sprawdziłem w ankiecie, na początku 2020, kobiet stanowiły około 8 procent populacji programistów, mężczyźni około 91. Załóżmy, że ze 100 kobiet 10 zdało test, a ze 100 mężczyzn zdało test 90.*

Separation

$$P\{R = 1 \mid Y = 1, A = a\} = P\{R = 1 \mid Y = 1, A = b\}$$

$$P\{R = 1 \mid Y = 0, A = a\} = P\{R = 1 \mid Y = 0, A = b\}$$

Kobiety	Zdany test	Niezdany test
Przyjęte	5	18
Odrzucone	5	72
Suma	10	90

Mężczyźni	Zdany test	Niezdany test
Przyjęci	45	2
Odrzuceni	45	8
Suma	90	10

Zdany test i przyjęta: 5/10 = 50%
Niezdany, nieprzyjęta: 72/90 = 80%
Procent przyjętych: (5+18)/100 = 23%

Zdany test i przyjęci: 45/90 = 50%
Niezdany, nieprzyjęty: 8/10 = 80%
Procent przyjętych: (45+2)/100 = 47%

Separacja jest spełniona ponieważ kobiety i mężczyźni którzy zdali test mają 50% szans na przyjęcie, a kobiety i mężczyźni którzy nie zdali mają 80% szans na bycie odrzuconym.

Separation

- Zamiast nazwy separation możemy używać nazwy Equalized Odds*

Pytanie, czy poprzedni przykład z problemem zatrudnienia, ma jakieś wady?

Separation

Kobiety	Zdany test	Niezdany test
Przyjęte	5	18
Odrzucone	5	72
Suma	10	90

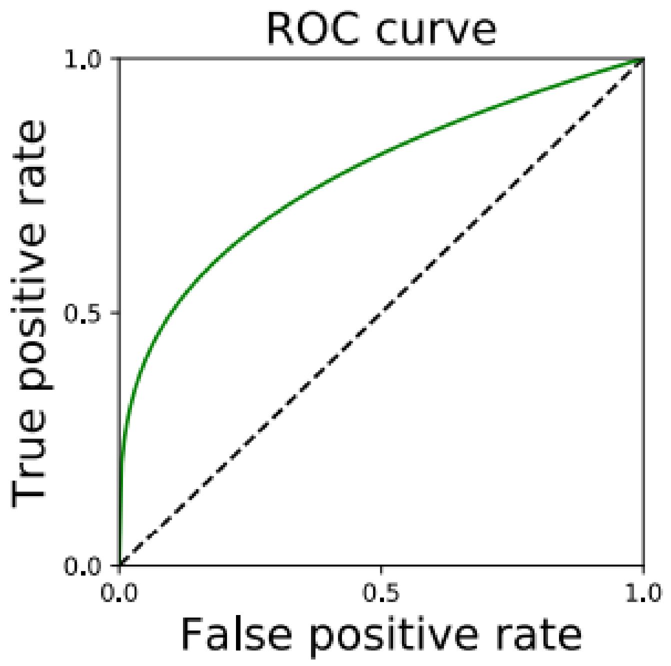
Mężczyźni	Zdany test	Niezdany test
Przyjęci	45	2
Odrzuceni	45	8
Suma	90	10

Separation dąży do zmniejszenia różnicy w liczebności osób które zdały test i zostały przyjęte. Jednak, czy to działa? Mamy 47 mężczyzn z dobrą pracą, którzy swoich synów poślą do dobrych szkół. Mamy 23 kobiet z dobrą pracą które poślą swoje córki do dobrych szkół. Uprzywilejowana grupa pozostaje uprzywilejowaną, a za czasem różnica w liczebności osób które zdały test i zostały przyjęte może wzrosnąć.

Osiąganie separation

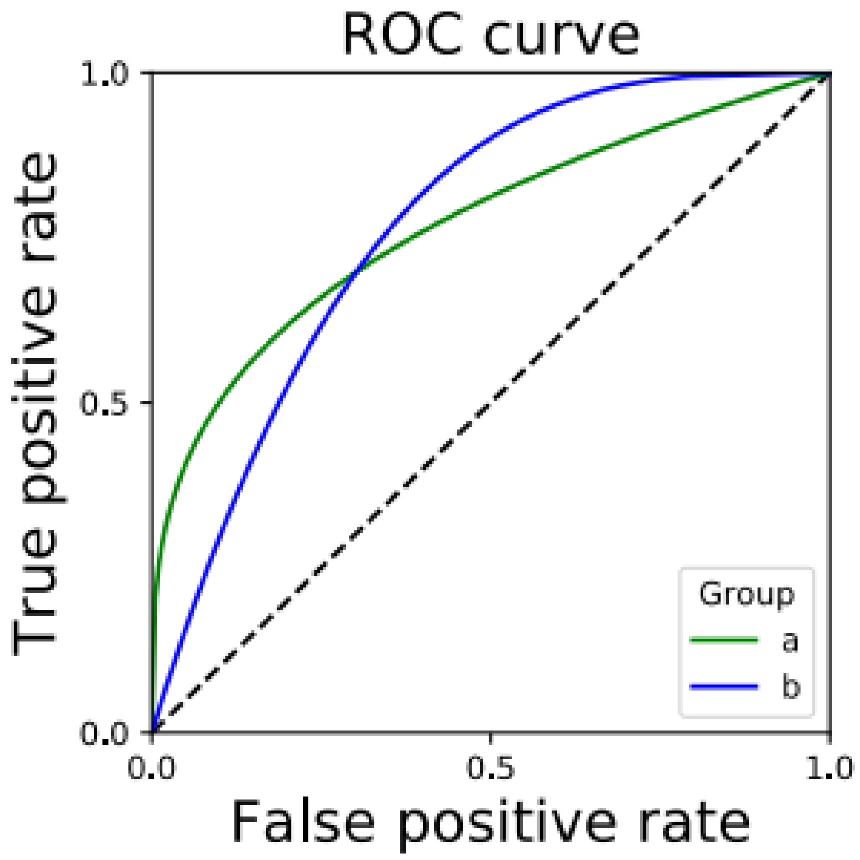
- Tak jak było w przypadku z niezależnością, możemy osiągnąć separację poprzez post-processing na funkcji score, nie trzeba dotrenować klasyfikatora.

Przypomnienie
Krzywa ROC jest wykresem graficznym, który ilustruje zdolność diagnostyczną binarnego systemu klasyfikatora, gdy jego próg dyskryminacji jest zmienny.



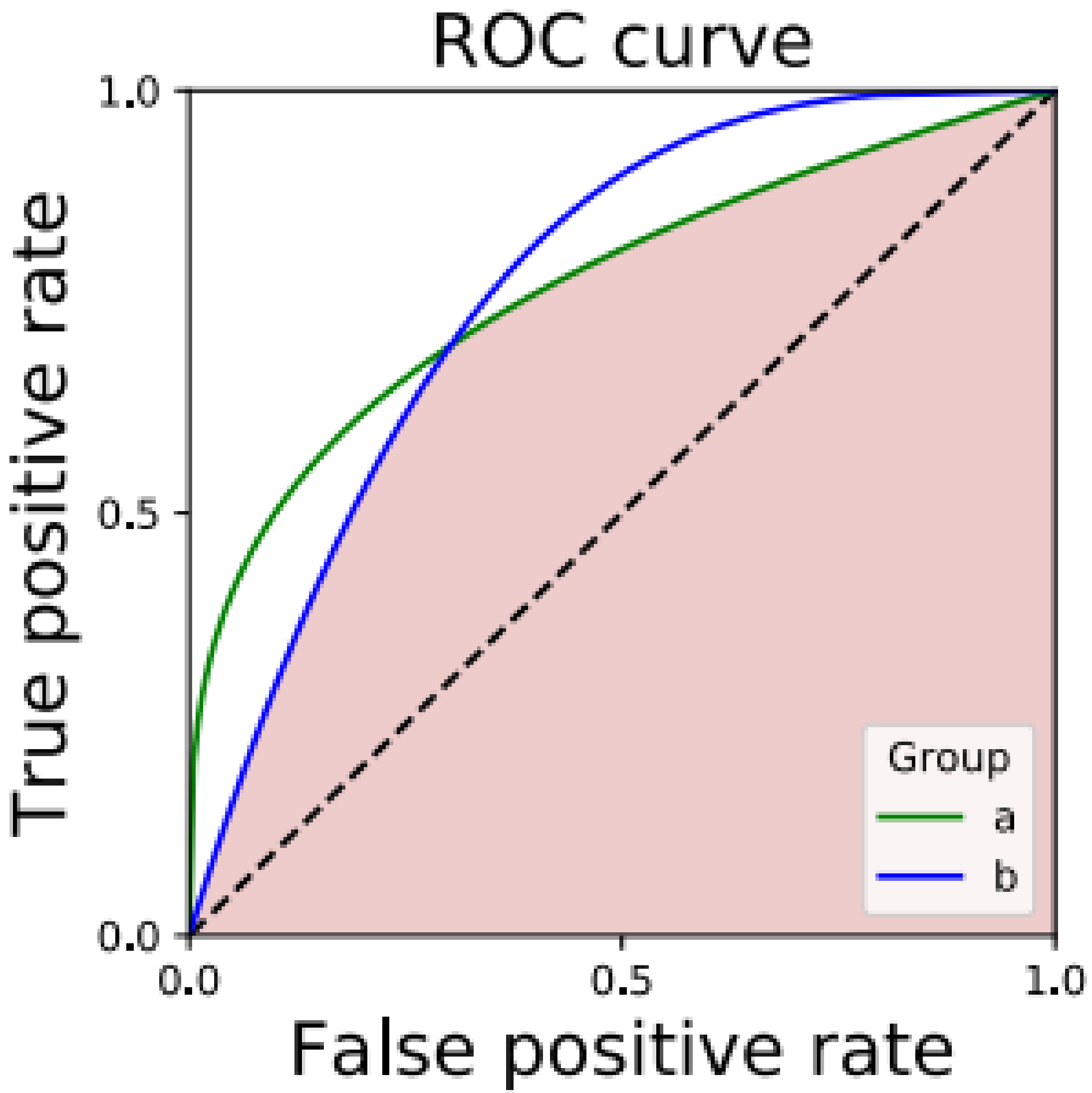
Osiąganie separation

- Klasyfikator binarny który spełnia warunki separacji, musi osiągnąć te same true positive rates i false positive rates we wszystkich grupach. Ten warunek odpowiada wzięciu przecięcia wszystkich krzywych ROC poszczególnych grup. Oraz regionu ograniczonym przez krzywe, gdzie możemy następnie wybrać klasyfikator, który minimalizuje podany koszt.

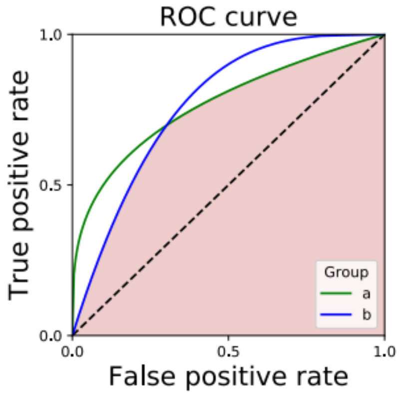


Osiąganie separation

- Nie wszystkie kompromisy między true positive rates i false positive rates są osiągalne.
- Punkty które nie leżą dokładnie na krzywych ale w zaznaczonym regionie, wymagają *randomizacji*.



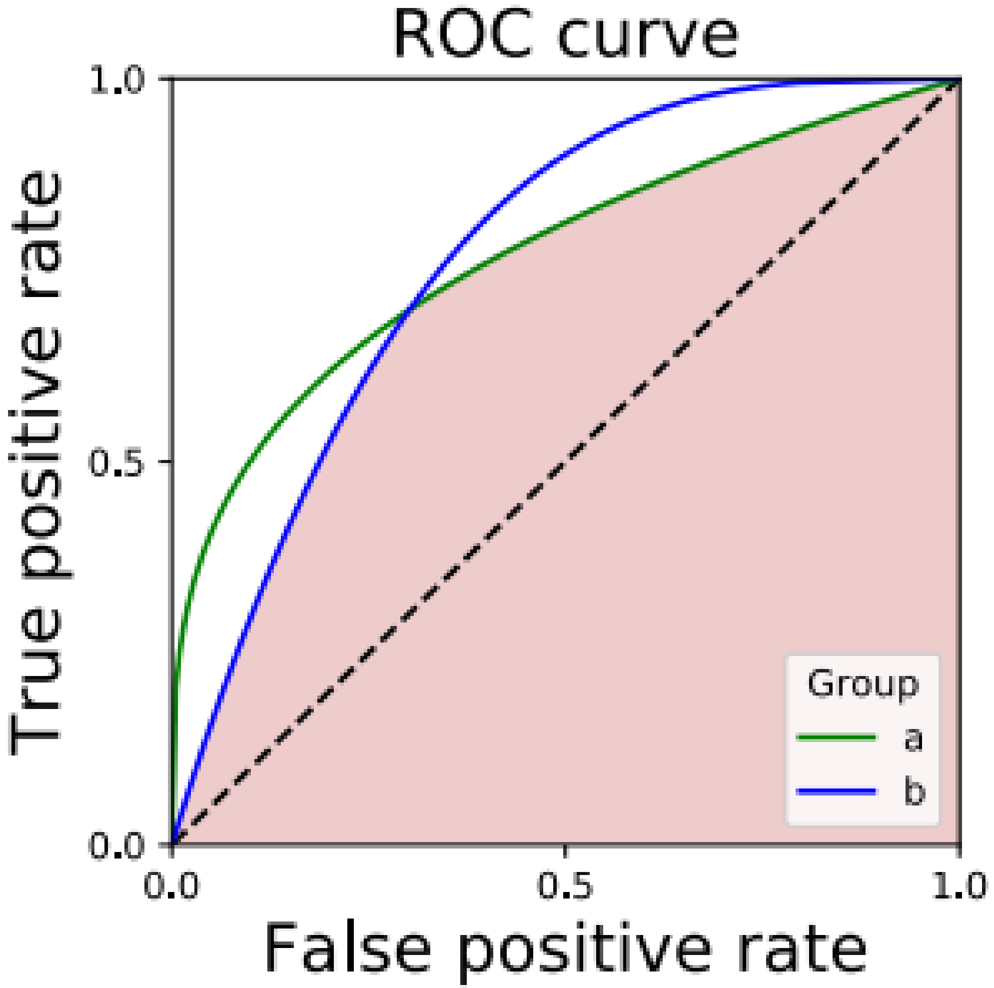
Osiąganie separation - *randomizacja*



- Aby zrozumieć randomizajcę, rozważ klasyfikator który akceptuje wszystkich, i przypisuje wartość true i false positive 1, czyli prawy górny róg wykresu. Rozważmy też klasyfikator który nie akceptuje nikogo, czyli true i false positive 0, lewy dolny róg wykresu. Rozważmy w końcu trzeci klasyfikator, który z prawdopodobieństwem1-p wybiera pierwszy kla. i z prawdopodobieństwem p drugi. Ten klasyfikator osiąga wartość true i false positive p, co daje nam jeden punkt na przerywanej linii na wykresie.

Osiąganie separation - *randomizacja*

- Wniosek, każdy punkt w zaznaczonym obszarze możemy osiągnąć poprzez randomizację pomiędzy pewnymi klasyfikatorami.



Osiąganie separation - podsumowanie

- Aby osiągnąć separację, tworzymy wykres ROC, patrzymy jakie wyniki są możliwe, a potem tworzymy klasyfikatory i randomizujemy pomiędzy nimi, aby otrzymać to co chcemy, czyli ilość kontra jakość.

Sufficiency

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

- Kryterium formalizuje, że rezultat już obejmuje wrażliwą cechę w celu przewidzenia wyniku.
- Często mówi się, że R satysfakcjonuje sufficiency kiedy A i Y są jasne dzięki kontekstowi.
- Dla binarnego przypadku:

$$P\{Y = 1 \mid R = r, A = a\} = P\{Y = 1 \mid R = r, A = b\}$$

Sufficiency

$$\mathbf{P\{Y = 1 \mid R = r, A = a\} = P\{Y = 1 \mid R = r, A = b\}}$$

- Klasyfikator spełnia tę definicję, jeśli zarówno grupy chronione, jak i niechronione mają równe prawdopodobieństwo, że podmiot o przewidywanej pozytywnej wartości naprawdę należy do klasy pozytywnej.
- Również to jest wtedy prawdą:

$$P\{Y = 0 \mid R = r, A = a\} = P\{Y = 0 \mid R = r, A = b\}$$

Sufficiency

Kobiety	Zdany test	Niezdany test	Mężczyźni	Zdany test	Niezdany test
Przyjęte	8	2	Przyjęci	72	0
Odrzucone	2	88	Odrzuceni	18	10
Suma	10	90	Suma	90	10

- W powyższym przykładzie przyjmujemy 80 procent obydwu płci z pozytywnym wynikiem testu, oraz 2 procent osób z negatywnym wynikiem. Od razu widać że kryterium sufficiency w tym przypadku jest bardziej korzystne dla pracodawcy od kryterium separation.
- Problem z tym kryterium może być taki, że również może pogłębiać różnicę między grupami.

Calibration and sufficiency

- Dotychczas nasze R było równe 0 lub 1, teraz zajmiemy się kalibracją tzn. że nasze R będzie przyjmować wartości z przedziału [0, 1]. Dlaczego? Ponieważ sufficiency po kalibracji nie musi zwracać ostatecznego wyniku, a być jedynie jednym z kroków na których podstawie budujemy nasz klasyfikator.
- Formalnie, R jest skalibrowany jeśli dla wszystkich wyników r z R mamy:

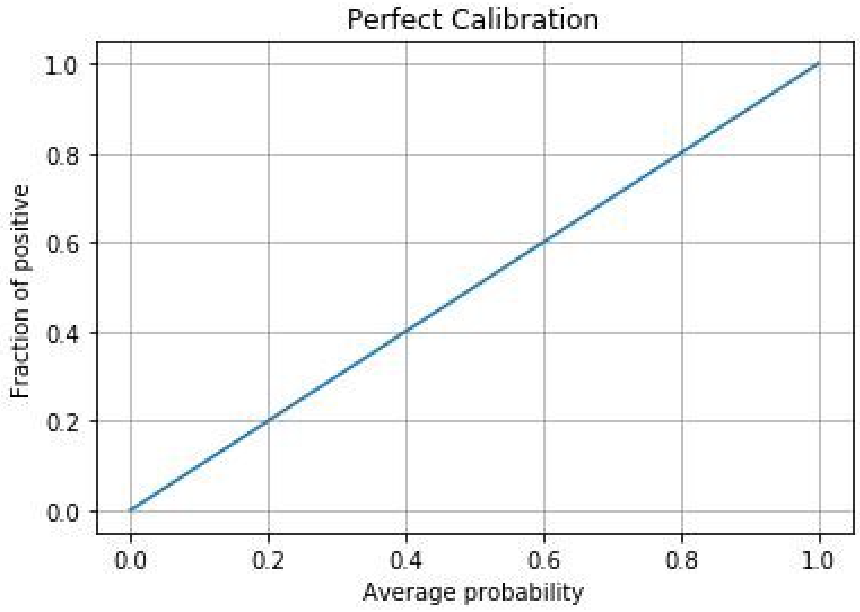
$$P\{Y = 1 \mid R = r\} = r$$

Calibration

P{Y = 1 | R = r} = r

- To oznacza, że wśród instancji którym przypisany jest score r, r-ta część instancji ma wynik pozytywny. To nie oznacza, że instancja ze score r, ma prawdopodobieństwo r na bycie pozytywną.

- Układamy osoby rosnąco względem r, dzielimy na równe grupy.



Calibration

Regresja logistyczna – w statystyce wykorzystywana do określania prawdop. przynależności do danej klasy.

- Jest wiele sposobów na osiągnięcie kalibracji, Platt scalling* jest jedną z popularniejszych.
- W tej technice używamy zmodyfikowanej funkcji sigmoidalnej, aby dopasować rozkład przewidywanych prawdopodobieństw do rozkładu prawdopodobieństwa obserwowanego w danych uczących. W rzeczywistości wykonujemy regresję logistyczną na wyniku modelu w odniesieniu do faktycznej etykiety.

Calibration

- Bardziej formalnie, Platt scaling wyznacza parametry a i b takie, że

$$S = \frac{1}{1 + \exp(aR + b)}$$

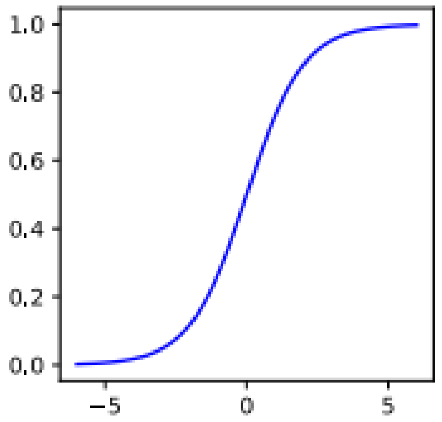
Funkcja sigmoidalna

Minimalizuje wartości funkcji utraty logarytmu względem target variable

$$-E[Y \log S + (1 - Y)\log(1 - S)].$$

Funkcja utraty logarytmu jest używana do oceny wydajności algorytmu używanego do rozwiązania zadania

[8] A plot of the sigmoid function $1/(1 + \exp(-x))$.



Calibration by group

- Z definicji widać, że sufficiency jest blisko związana z ideą kalibracji. Aby sformalizować to połączenie, mówimy, że wynik R spełnia kryteria kalibracji dla grupy, jeśli spełnia

$$P\{Y = 1 \mid R = r, A = a\} = r$$

dla wszystkich wartości score r i grup a.

- Kalibracja ma ten sam wymóg na poziomie populacji, bez warunkowania na A.

Fakt 1. Calibration by group implikuje sufficiency

- I odwrotnie, sufficiency jest tylko nieznacznie słabsze niż kalibracja według grup.

Twierdzenie 1.

- Jeśli score R spełnia sufficiency, to istnieje funkcja

$$f: [0, 1] \to [0, 1],$$

taka, że f(R) spełnia kalibrację według grup.

Dowód tw. 1

- Jeśli score R spełnia sufficiency, to istnieje funkcja

$$f: [0, 1] \to [0, 1],$$

taka, że f(R) spełnia kalibrację według grup.

- Ustalmy grupę a, mamy $f(r) = P\{Y=1 \mid R=r, A=a\}$. Ponieważ R spełnia sufficiency, prawdop. to jest takie samo dla dowolnej grupy a, zatem stosując f do innych grup niż a, otrzymamy to samo prawdop. Teraz, rozważmy dwie grupy, a i b. Mamy

$$r = P\{Y = 1 \mid f(R) = r, A = a\}$$
$$= P\{Y = 1 \mid R \in f^{-1}(r), A = a\}$$
$$= P\{Y = 1 \mid R \in f^{-1}(r), A = b\}$$
$$= P\{Y = 1 \mid f(R) = r, A = b\},$$

Zatem f(R) jest skalibrowane według grup.

Podsumowanie kalibracji i sufficiency

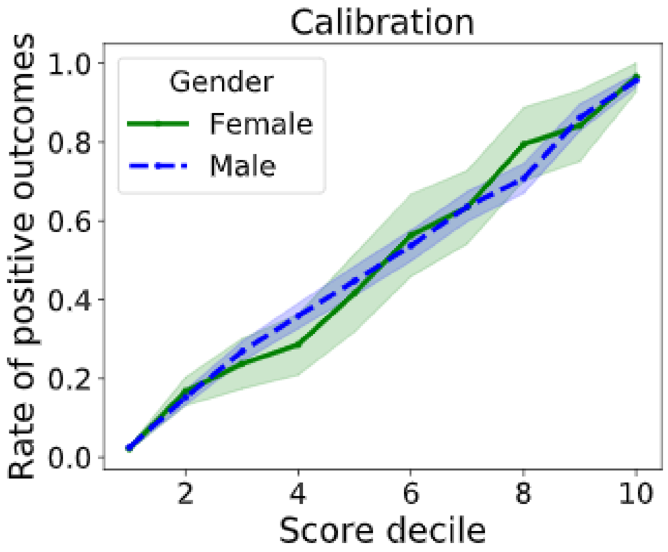
- Sufficiency i kalibracja według grup są zasadniczo równoważnymi pojęciami. W szczególności daje nam to duży repertuar metod osiągania sufficiency. Moglibyśmy na przykład zastosować skalowanie Platt'a dla każdej z grup zdefiniowanych przez atrybut wrażliwy A.

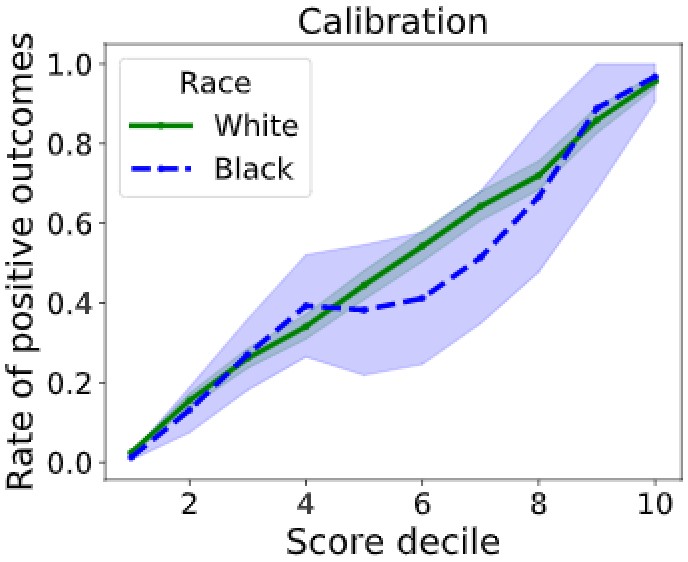
Calibration by group jako konsekwencja nauczania nieo*graniczonego*

- Sufficiency jest osiągana często bez konieczności interweniowania w proces nauczania. Zazwyczaj przecież oczekujemy, że score spełni sufficiency gdzie sensitive attribute może być przewidziany dzięki innym atrybutom.
- Żeby to zilustrować, twórcy pracy aplikują regresję logistyczną, korzystają ze standardowego UCI data setu dorosłych* i biblioteki skylearn w Pythonie. Twórcy nie dostrajają i nie kalibrują wyniku.

Grupa testowa

Czarni	Biali	Kobiety	Mężczyźni
1561	13946	5421	10860





- Widzimy, że model jest nieźle skalibrowany dla płci.
- Gorzej wygląda to dla rasy. Jest tak ponieważ mamy mało danych z atrybutem *czarny*.
- Zatem, jeszcze raz, sufficiency jest osiągana często bez konieczności interweniowania w proces nauczania, ponieważ standardowe techniki już o to dbają.

Relacje między kryteriami

- Jak podpowiada intuicja, stosowanie więcej niż jednego kryterium na raz, doprowadzi do miernego klasyfikatora.
- Jeśli jednak poluzujemy wymagania kryteriów, możemy liczyć na sensowny klasyfikator, który jest kompromisem pomiędzy kryteriami.

Independence kontra Sufficiency

- Jeśli sensitive attribute A i target variable Y są zależne, to znaczy, że jedna grupa ma wyższy odsetek pozytywnych wyników.
- Twierdzenie 2. Załóżmy, że A i Y są zależne. Wtedy sufficiency i independence wykluczają się.

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Twierdzenie 2. Załóżmy, że A i Y są zależne. Wtedy sufficiency i independence wykluczają się.

- Proof, by the contraction rule for conditional independence,

$$A \perp R \text{ and } A \perp Y \mid R \Rightarrow A \perp (Y, R) \Rightarrow A \perp Y$$

- $A \perp (Y, R)$ oznacza że A jest niezależne od pary (Y, R). Pominięcie R nie da nam zależności między A i Y.
- Stąd w kontrapozycji,
    - $A \, !\!\perp Y \Rightarrow A \, !\!\perp R$ lub $A \, !\!\perp R \mid Y$

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

c.k.d

Independence kontra Separation

- Tutaj podobnie, kryteria wykluczają się wzajemnie. Aby to udowodnić wymagane jest założenie, że target variable Y jest binarne. Oraz score jest zależny od target variable. To ostatnie uzasadniamy tym, że chcemy aby klasyfikator korelował z target variable.

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Twierdzenie 3. Załóżmy, że Y jest binarne, A jest zależne od Y, R jest zależne od Y.

- Załóżmy że, $Y \in \{0, 1\}$. Musimy udowodnić

$$A \perp R \text{ and } A \perp R \mid Y \Rightarrow A \perp Y \text{ or } R \perp Y$$

Dzięki twierdzeniu o prawdop. całkowitym

$$P\{R = r \mid A = a\} = \sum_y P\{R = r \mid A = a, Y = y\} P\{Y = y \mid A = a\}$$

Tw. o prawdop. całkowitym - twierdzenie pozwalające na obliczanie prawdopodobieństw zdarzeń, które mogą zajść w konsekwencji zajścia innych zdarzeń, takich jak doświadczenia wieloetapowe.

Twierdzenie 3. Załóżmy, że Y jest binarne, A jest zależne od Y, R jest zależne od Y.

- $P\{R = r \mid A = a\} = \sum_y P\{R = r \mid A = a, Y = y\}P\{Y = y \mid A = a\}$

Powyższe po zaaplikowaniu $A \perp R$ i $A \perp R \mid Y$, upraszcza się do

$P\{R = r\} = \sum_y P\{R = r \mid Y = y\}P\{Y = y \mid A = a\}$

Inaczej aplikując dane, otrzymamy również

$P\{R = r\} = \sum_y P\{R = r \mid Y = y\}P\{Y = y\}$

Łącząc dwa powyższe napisy

$\sum_y P\{R = r \mid Y = y\}P\{Y = y\} = \sum_y P\{R = r \mid Y = y\}P\{Y = y \mid A = a\}$

$\sum_y P\{R = r \mid Y = y\}P\{Y = y\} = \sum_y P\{R = r \mid Y = y\}P\{Y = y \mid A = a\}$

- Jeśli y przyjmuje tylko dwie wartości, ta równość może być spełniona tylko jeśli $A \perp Y$ lub $R \perp Y$.
- Inaczej zapisując równość, użyjemy symboli $p = P\{Y = 0\}$,

$p_a = P\{Y = 0 \mid A = a\}$, $r_y = P\{R = r \mid Y = y\}$

$$pr_0 + (1 - p)r_1 = p_a r_0 + (1 - p_a)r_1$$

Odpowiednio, $p(r_0 - r_1) = p_a(r_0 - r_1)$.

To równanie może być spełnione tylko jeśli $r_0 = r_1$ wtedy $R \perp Y$, lub jeśli $p = p_a$ dla wszystkich a, wtedy $Y \perp A$.

c. k. d

Separation kontra Sufficiency

- Na koniec, okazuje się że ponieważ obydwa kryteria korzystają ze zmiennych A, R, Y ich stosowanie w jednym klasyfikatorze prowadzi do zdegenerowanej przestrzeni rozwiązań.

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Separation kontra Sufficiency

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

- Tw. 4. Załóżmy, że wszystkie zdarzenia we wspólnym rozkładzie (A, R, Y) mają pozytywne prawdopodobieństwo, oraz A !⊥ Y. Wtedy, separacja i sufficiency wykluczają się.

Załóżmy, że wszystkie zdarzenia we wspólnym rozkładzie (A, R, Y) mają pozytywne prawdopodobieństwo, oraz A !⊥ Y. Wtedy, separacja i sufficiency wykluczają się.

- Ze standardowego faktu o warunkowej niezależności*

$$A \perp R \mid Y \text{ and } A \perp Y \mid R \Rightarrow A \perp (R,Y)$$

Zatem

$$A \perp (R,Y) \Rightarrow A \perp R \text{ and } A \perp Y$$

c. k. d

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Tw. 5. Załóżmy, że Y jest zależne od A oraz, że Y' jest binarnym klasyfikatorem z niezerowym false positive rate. Wtedy, separcja i sufficiency wykluczają się.

- Ponieważ Y jest zależne od A, muszą istnieć dwie grupy, nazwijmy je 0 i 1 takie, że

$$p_0 = P\{Y = 1 \mid A = 0\} \mathrel{!}= P\{Y = 1 \mid A = 1\} = p_1$$

Załóżmy, że separacja zachodzi. Ponieważ klasyfikator nie jest idealny, to wszystkie grupy mają ten sam dodatni false positive rate FPR>0 i taki sam true positive rate TPR>0. Pokażemy, że sufficiency nie zachodzi.

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Tw. 5. Załóżmy, że Y jest zależne od A oraz, że Y' jest binarnym klasyfikatorem z niezerowym false positive rate. Wtedy, separcja i sufficiency wykluczają się.

- Przypomnijmy, sufficiency implikuje, że wszystkie grupy mają ten sam positive predictive value. PPV w grupie a, satysfakconuje

$$PPV_a = \frac{TPRp_a}{TPRp_a + FPR(1 - p_a)}$$

Z tego wynika, że $PPV_0 = PPV_1$ tylko jeśli TPR = 0 lub FPR = 0. Drugie nie zajdzie z założenia. Zatem TPR = 0, jednak możemy sprawdzić, że negative predictive value NPV_0 rózni się od NVP_1.

Independence	Separation	Sufficiency
$R \perp A$	$R \perp A \mid Y$	$Y \perp A \mid R$

Zatem TPR = 0, jednak możemy sprawdzić,
że negative predictive value NPV_0 rózni się od NVP_1.

$$NPVa = \frac{(1 - FPR)(1 - p_a)}{(1 - TPR)p_a + (1 - FPR)(1 - p_a)}$$

Zatem sufficiency nie zachodzi.

c. n. d.

Bibliografia

- https://fairmlbook.org/
- https://www.statista.com/statistics/1126823/worldwide-developer-gender/#:~:text=According%20to%20a%20global%20software,of%20the%20software%20development%20job.
- https://towardsdatascience.com/a-tutorial-on-fairness-in-machine-learning-3ff8ba1040cb
- Learning Fair Representations. Richard Zemel, Yu (Ledell) Wu, Kevin Swersky, Toniann Pitassi
- https://towardsdatascience.com/calibrating-classifiers-559abc30711a
- https://archive.ics.uci.edu/ml/datasets/adult
- https://medium.com/analytics-vidhya/calibration-in-machine-learning-e7972ac93555
- 17.2 in Wasserman, All of Statistics