Wprowadzenie do sprawiedliwego ML

Julia Majkowska | Grzegorz Ciesielski

13 października 2020

Uczenie nadzorowane

Chcemy stworzyć system (klasyfikator), który jest w stanie wykonywać skomplikowane zadania.

Uczenie nadzorowane

Chcemy stworzyć system (klasyfikator), który jest w stanie wykonywać skomplikowane zadania. Mamy:

- zbiór danych X,

Uczenie nadzorowane

Chcemy stworzyć system (klasyfikator), który jest w stanie wykonywać skomplikowane zadania. Mamy:

- zbiór danych X,
- zbiór możliwych etykiet Y,

Uczenie nadzorowane

Chcemy stworzyć system (klasyfikator), który jest w stanie wykonywać skomplikowane zadania. Mamy:

- zbiór danych X,
- zbiór możliwych etykiet Y,
- zbiór treningowy $\{(x_1, y_1), \ldots (x_n, y_n)\}$.

$$\begin{aligned} x_1 &\longrightarrow y_1 \\ x_2 &\longrightarrow y_2 \\ &\vdots \\ x_n &\longrightarrow y_n \end{aligned}$$

Uczenie nadzorowane

Chcemy stworzyć system (klasyfikator), który jest w stanie wykonywać skomplikowane zadania. Mamy:

- zbiór danych X,
- zbiór możliwych etykiet Y,
- zbiór treningowy $\{(x_1, y_1), \ldots (x_n, y_n)\}$.

$$
\begin{aligned}
x_1 &\longrightarrow y_1 \\
x_2 &\longrightarrow y_2 \\
&\vdots \\
x_n &\longrightarrow y_n \\
\\
x_{n+1} &\longrightarrow ?
\end{aligned}
$$

Uczenie nadzorowane

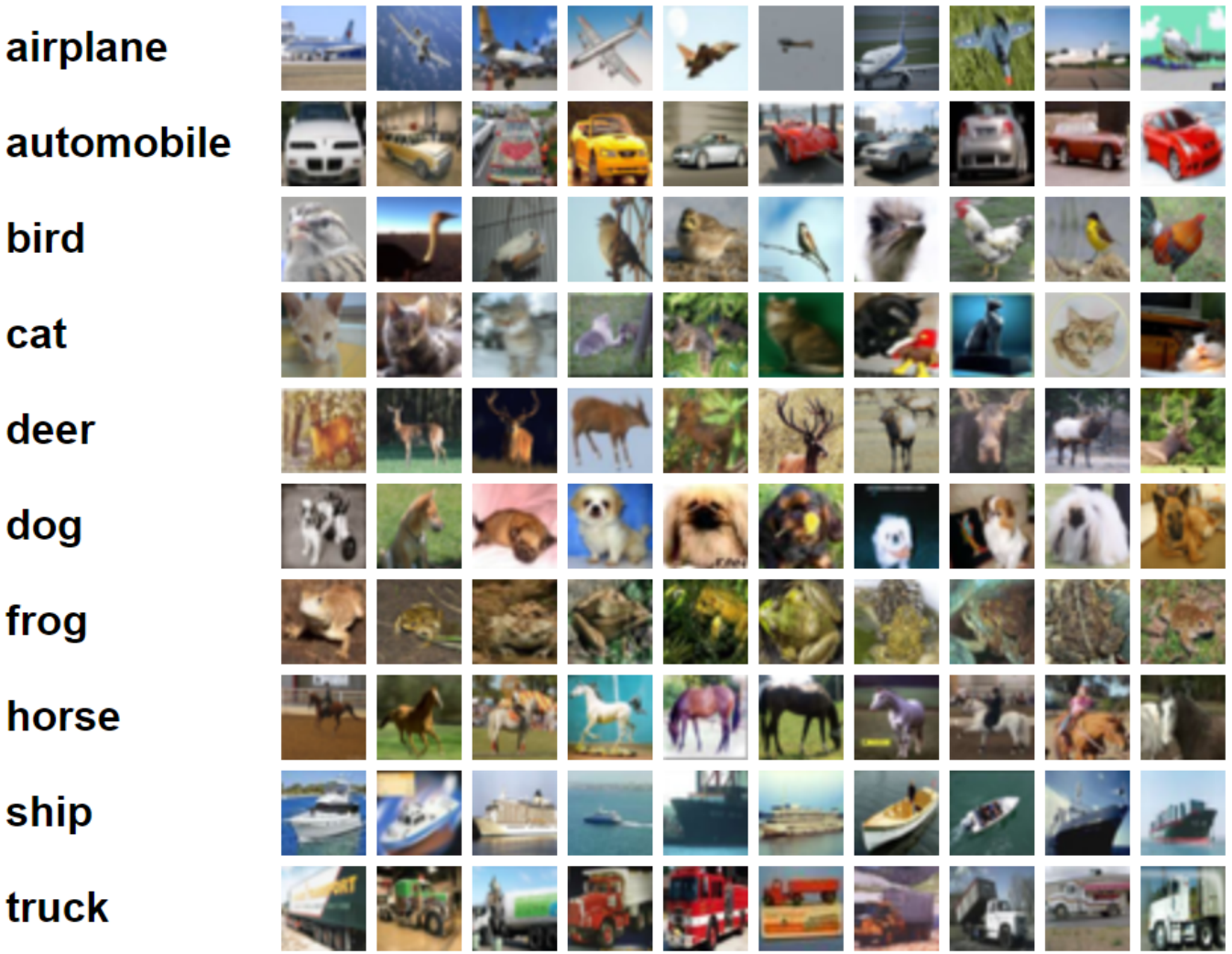
Chcemy stworzyć system (klasyfikator), który jest w stanie wykonywać skomplikowane zadania. Mamy:

- zbiór danych X,
- zbiór możliwych etykiet Y,
- zbiór treningowy $\{(x_1, y_1), \ldots (x_n, y_n)\}$.

$$
\begin{array}{ccc}
x_1 & \longrightarrow & y_1 \\
x_2 & \longrightarrow & y_2 \\
 & \vdots & \\
x_n & \longrightarrow & y_n \\
 & & \\
x_{n+1} & \longrightarrow & ?
\end{array}
$$

Klasyfikator: $f : X \to Y$.

Przykład: cifar-10



Rozkład

Zazwyczaj stosuje się jednak konwencję, że dane pochodzą z pewnego rozkładu prawdopodobieństwa. Wtedy X, Y są (zależnymi) zmiennymi losowymi.

Rozkład

Zazwyczaj stosuje się jednak konwencję, że dane pochodzą z pewnego rozkładu prawdopodobieństwa. Wtedy X, Y są (zależnymi) zmiennymi losowymi.

Definicja

1. $\hat{y} = f(x)$

Rozkład

Zazwyczaj stosuje się jednak konwencję, że dane pochodzą z pewnego rozkładu prawdopodobieństwa. Wtedy X, Y są (zależnymi) zmiennymi losowymi.

Definicja

1. $\hat{y} = f(x)$
2. $\hat{Y} = f(X)$

Chcielibyśmy, żeby $\hat{Y} = Y$.

Przykład: klasyfikator chorób serca

Chcemy móc przewidywać, czy dany człowiek będzie miał choroby serca czy nie.

- X: ludzie
- Y: czy będzie miał choroby serca?

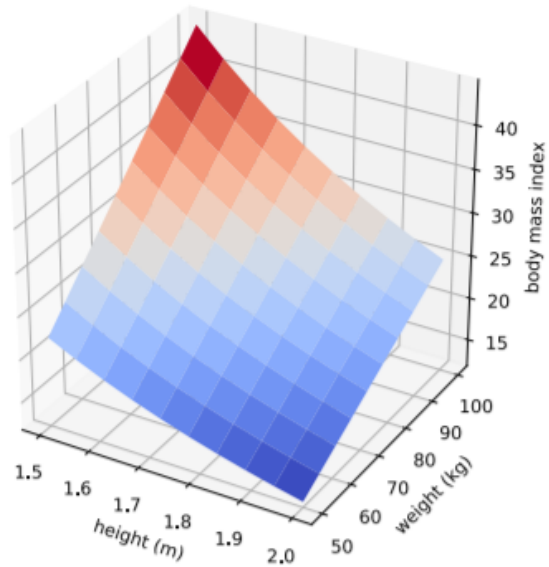
Przykład: klasyfikator chorób serca

Chcemy móc przewidywać, czy dany człowiek będzie miał choroby serca czy nie.

- X: ludzie
- Y: czy będzie miał choroby serca?

Możemy stworzyć funckję $f(X) = W/H^2$, gdzie $W = \text{waga}(X)$, a $H = \text{wzrost}(X)$.

Taka funkcja nazywa się funkcją oceny (ang. *score function*).



Przykład: klasyfikator chorób serca

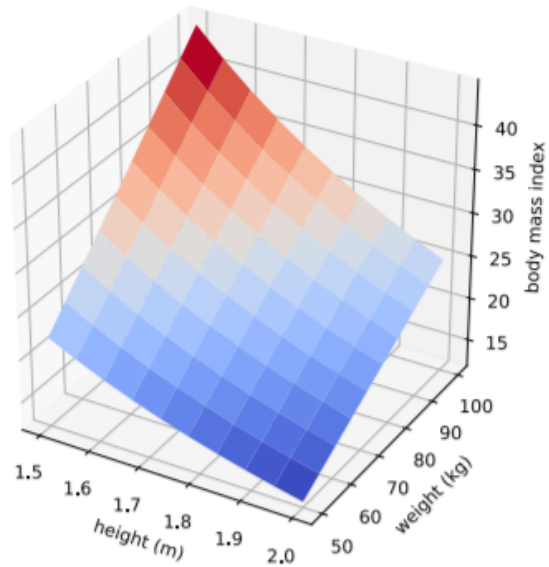
Chcemy móc przewidywać, czy dany człowiek będzie miał choroby serca czy nie.

- X: ludzie
- Y: czy będzie miał choroby serca?

Możemy stworzyć funckję $f(X) = W/H^2$, gdzie $W = \text{waga}(X)$, a $H = \text{wzrost}(X)$.

Taka funkcja nazywa się funkcją oceny (ang. *score function*).

Słaba!



Przykład: klasyfikator chorób serca

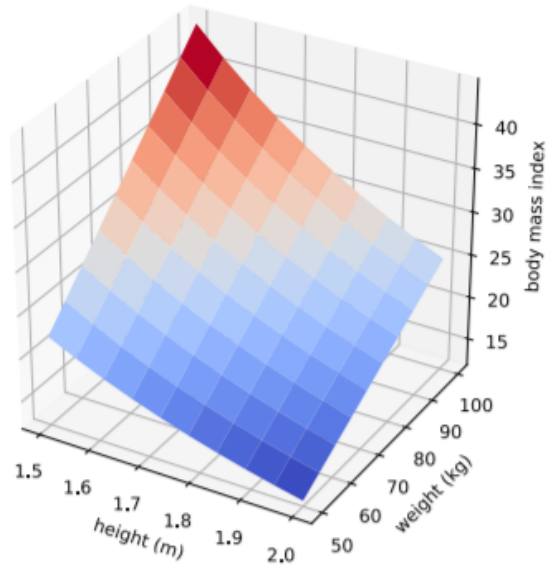
Chcemy móc przewidywać, czy dany człowiek będzie miał choroby serca czy nie.

- X: ludzie
- Y: czy będzie miał choroby serca?

Możemy stworzyć
funckję $f(X) = W/H^2$, gdzie
$W = \text{waga}(X)$, a $H = \text{wzrost}(X)$.

Taka funkcja nazywa się
funkcją oceny (ang. *score function*).

Słaba! I dyskryminuje ludzi wysokich!



Warunkowe oczekiwania

Definicja

$$R(X) = E(Y|X)$$

Age. Curt.	Per-sons.	Age. Curt.	Per-sons	Age. Curt.	Per-sons	Age. Curt.	Per-sons	Age. Curt.	Per-sons	Age. Curt.	Per-sons
1	1000	8	680	15	628	22	585	29	539	36	481
2	855	9	670	16	622	23	579	30	531	37	472
3	798	10	661	17	616	24	573	31	523	38	463
4	760	11	653	18	610	25	567	32	515	39	454
5	732	12	646	19	604	26	560	33	507	40	445
6	710	13	640	20	598	27	553	34	499	41	436
7	692	14	634	21	592	28	546	35	490	42	427
Age. Curt	Per-sons.	Age. Curt.	Per-sons	Age. Curt.	Per-sons	Age Curt	Per-sons	Age. Curt.	Per-sons	Age. Curt.	Per-sons
43	417	50	346	57	272	64	202	71	131	78	58
44	407	51	335	58	262	65	192	72	120	79	49
45	397	52	324	59	252	66	182	73	109	80	41
46	387	53	313	60	242	67	172	74	98	81	34
47	377	54	302	61	232	68	162	75	88	82	28
48	367	55	292	62	222	69	152	76	78	83	23
49	357	56	282	63	212	70	142	77	68	84	20

Age.	Persons.
7	5547
14	4584
21	4270
28	3564
35	3604
42	3178
49	2709
56	2194
63	1694
70	1204
77	692
84	253
100	107
	34000
	Sum Total.

Figure 4: Halley's life table (1693)

Tworzenie klasyfikatora

Mając funkcję oceny, możemy stworzyć binarny klasyfikator następująco: wybrać próg (ang. *threshold*) t i wszystkim kandydatom poniżej przyznać negatywną odpowiedź, a powyżej - pozytywną.

Problemy:

- funkcja oceny wcale niekoniecznie ustawia kandydatów w poprawnej kolejności - jest tylko przybliżeniem faktycznego rozkładu
- jaki próg wybrać?

Kryteria klasyfikacyjne - accuracy

Definicja

Accuracy (Dokładność) - $P(\hat{Y} = Y) = \frac{TP+TN}{TP+FP+TN+FN}$

Event	Condition	Resulting notion ($\mathbb{P}\{\text{event} \mid \text{condition}\}$)
$\hat{Y} = 1$	$Y = 1$	True positive rate, recall
$\hat{Y} = 0$	$Y = 1$	False negative rate
$\hat{Y} = 1$	$Y = 0$	False positive rate
$\hat{Y} = 0$	$Y = 0$	True negative rate

Kryteria klasyfikacyjne - precision

Definicja

Precision - $\frac{TP}{TP+FP}$

Kryteria klasyfikacyjne - precision

Definicja

Precision - $\frac{TP}{TP+FP}$

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

Kryteria klasyfikacyjne - precision

Definicja

Precision - $\frac{TP}{TP+FP}$

$$\text{precision} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{retrieved documents}\}|}$$

W przypadku przewidywania czy dany film na youtubie może być pokazywany dzieciom istotne jest, żeby jak najbardziej minimalizować fałszywe pozytywy i precyzja jest tu dobrą metryką.

Kryteria klasyfikacyjne - recall

Definicja

Recall - $\frac{TP}{TP+FN}$

Kryteria klasyfikacyjne - recall

Definicja

Recall - $\frac{TP}{TP+FN}$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

Kryteria klasyfikacyjne - recall

Definicja

Recall - $\frac{TP}{TP+FN}$

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

W przypadku systemu oceniającego czy więzień powinien zostać zwolniony warunkowo nie branie pod uwagi fałszywych negatywów może być niebezpieczne, i recall jest lepszą metryką niż precision.

Kryteria klasyfikacyjne - F1

Definicja

$$\text{F1 - } 2\frac{precision * recall}{precision + recall} = 2\frac{TP}{TP + \frac{1}{2}(fp + fn)}$$

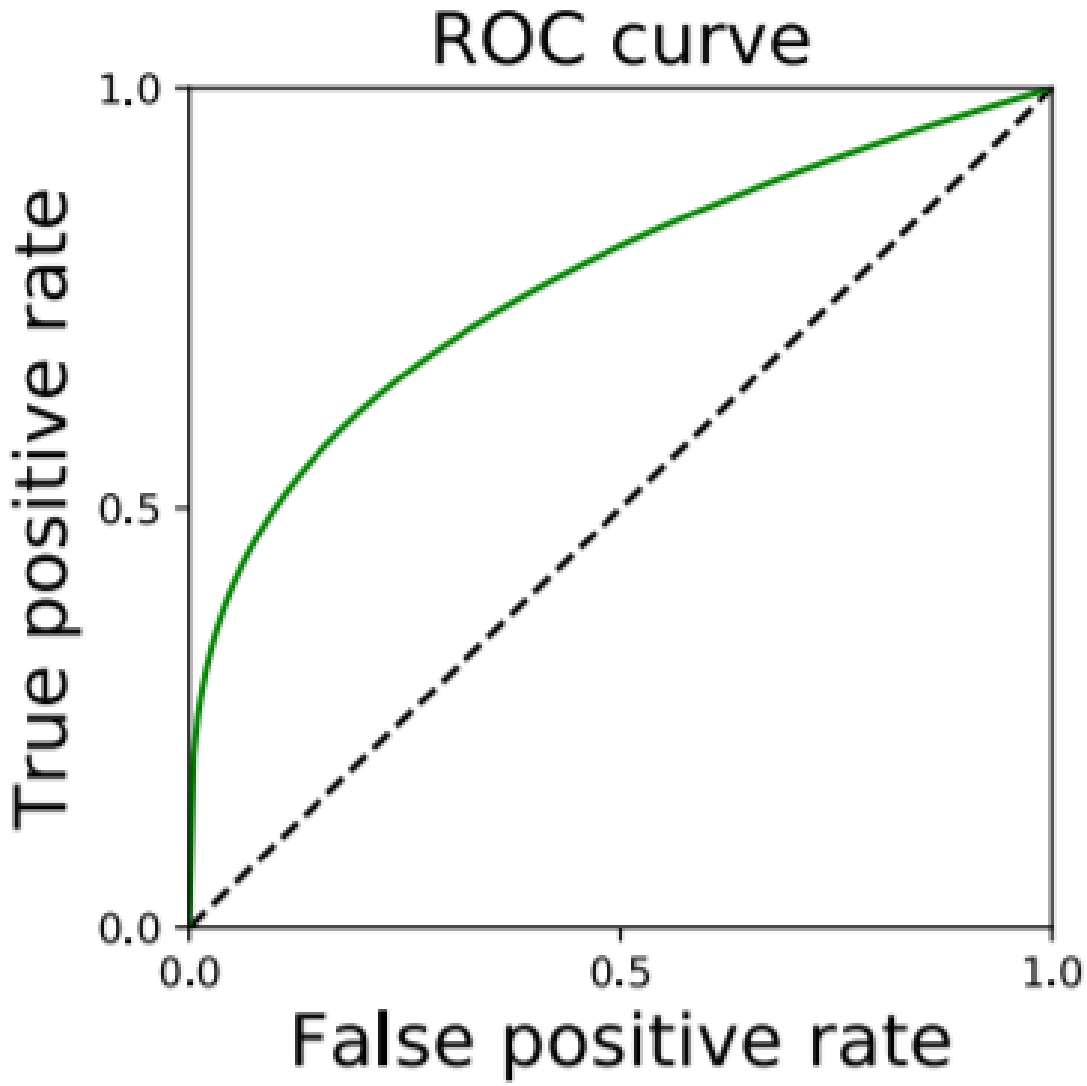
Przykład: Kryteria klasyfikacyjne

Chcemy stworzyć klasyfikator przewidujący czy dana osoba będzie brała udział w wypadku samochodowym w przyszłym roku.

Przykład: Kryteria klasyfikacyjne

Chcemy stworzyć klasyfikator przewidujący czy dana osoba będzie brała udział w wypadku samochodowym w przyszłym roku. Klasyfikator przewidujący, że nikt nigdy nie ma wypadku będzie dobrze sobie radził w tych metrykach pomimo, że nie wnosi żadnej wartościowej informacji.

Krzywa ROC


ROC curve
True positive rate
False positive rate
0.0
0.5
1.0
0.0
0.5
1.0

Czynniki wrażliwe

Definicja

Czynnki wrażliwe - wektor czynników, którego różne wartości identyfikują różne grupy społeczne

Czynniki wrażliwe

Definicja

Czynnki wrażliwe - wektor czynników, którego różne wartości identyfikują różne grupy społeczne

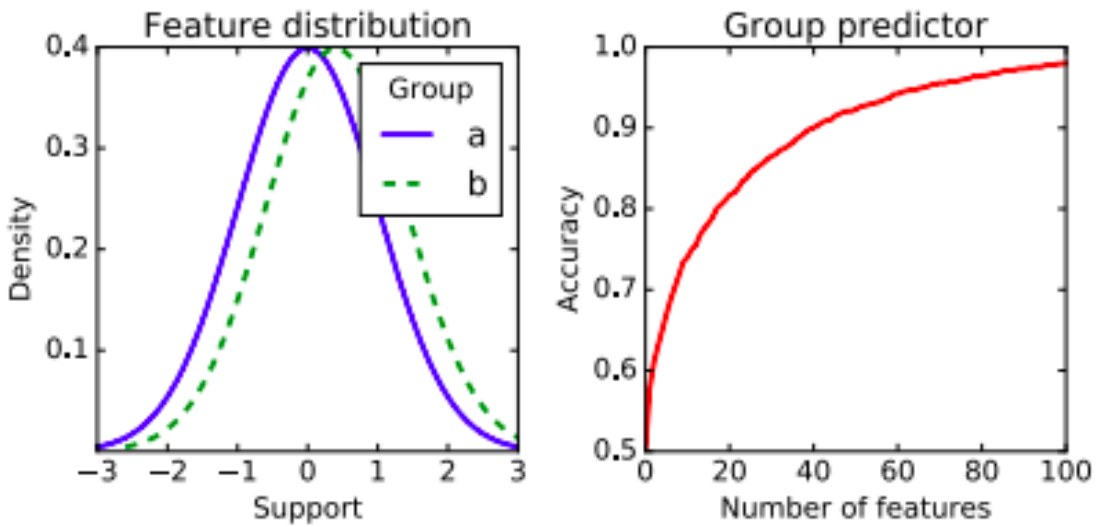
Mając wiele czynników skorelowanych z czynnikiem wrażliwym nasz klasyfikator naturalnie może stać się stronniczy względem czynnika wrażliwego.

Przykład : Czynniki wrażliwe

Pinterest Demographics

- 71% of Pinterest users are actually Females.
- 40% of New Signups are Men; 60% New Signups are Women.
- Men account for only 7% of total pins on Pinterest.
- 40% of US dads use Pinterest.
- 50%+ Pinners live outside the US.
- 34% of Users aged 18-29 in the US uses Pinterest.
- 42% US women adults who uses Pinterest.
- 80% of US mothers who use the internet uses Pinterest.
- 1 out of 2 **US millennials** uses Pinterest every month.
- The median age of a Pinterest user is 40, however, the majority of active pinners are below 40.
- **Half of Pinterest users earn $50K or greater per year**, with 10 percent of Pinteresting households making greater than $125K.
- 28% of marketers are already using Pinterest.
- 144.5 million – Number of people that pinterest reports can be reached with adverts on pinterest
- 28% of all US social media users are Pinterest users.

No fairness through to unawareness



Kryteria niedyskryminacji

Definicja/przypomnienie

A: wrażliwy czynnik
R: klasyfikator/ocena
Y: stan faktyczny

Formalne kryteria niedyskryminacji:

1. niezależność (*independence*): $R \perp A$
2. oddzielność (*separation*): $R \perp A \mid Y$
3. wystarczalność (*sufficiency*): $Y \perp A \mid R$

Niezależność

Warunek niezależności (dla wszystkich par grup $a, b \in A$):

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b)$$

Niezależność

Warunek niezależności (dla wszystkich par grup $a, b \in A$):

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b)$$

Złagodzony:

$$\frac{P(R = 1 \mid A = a)}{P(R = 1 \mid A = b)} \geqslant 1 - \epsilon$$

Niezależność

Warunek niezależności (dla wszystkich par grup $a, b \in A$):

$$P(R = 1 \mid A = a) = P(R = 1 \mid A = b)$$

Złagodzony:

$$\frac{P(R = 1 \mid A = a)}{P(R = 1 \mid A = b)} \geqslant 1 - \epsilon$$

Dla $\epsilon = 0.2$ to tzw. *„Zasada 80 procent”*

Problemy z niezależnością

Firma zatrudnia pracowników z grup etnicznych a i b.

- z grupy a wybiera $\frac{p}{|a|}$ najlepszych kandydatów,
- z grupy b wybiera $\frac{p}{|b|}$ kandydatów losowo.

Problemy z niezależnością

Firma zatrudnia pracowników z grup etnicznych a i b.

- z grupy a wybiera $\frac{p}{|a|}$ najlepszych kandydatów,
- z grupy b wybiera $\frac{p}{|b|}$ kandydatów losowo.

Niezależność jest zachowana!

Problemy z niezależnością

Firma zatrudnia pracowników z grup etnicznych a i b.

- z grupy a wybiera $\frac{p}{|a|}$ najlepszych kandydatów,
- z grupy b wybiera $\frac{p}{|b|}$ kandydatów losowo.

Niezależność jest zachowana!

Ale są problemy. Jakie?

Problemy z niezależnością

Firma zatrudnia pracowników z grup etnicznych a i b.

- z grupy a wybiera $\frac{p}{|a|}$ najlepszych kandydatów,
- z grupy b wybiera $\frac{p}{|b|}$ kandydatów losowo.

Niezależność jest zachowana!

Ale są problemy. Jakie?

Czy może się to zdarzyć nieintencjonalnie?

W jaki sposób stowrzyć sprawiedliwy klasyfikator

1. Preprocessing atrybutów, żeby wyeliminować korelacje z wrażliwym czynnikiem
2. Zmodyfikować proces trenowania aby wprowadzić ograniczyć korelację
3. Ograniczyć korelację po treningu

Koniec