POLITECHNIKA GDAŃSKA

AI TECH

Wprowadzenie do Sztucznej Inteligencji

testowanie modeli uczenia maszynowego

Adam Kurowski
Katedra Systemów Multimedialnych,
Wydział Elektroniki, Telekomunikacji i Informatyki PG

Fundusze Europejskie Polska Cyfrowa
Rzeczpospolita Polska
Unia Europejska Europejski Fundusz Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.

Wprowadzenie

Niezależnie od wybranej techniki uczenia maszynowego po wytrenowaniu modelu pojawia się zwykle **konieczność szerszego przetestowania skuteczności wybranego rozwiązania.**

Czasami też może być konieczne sprawdzenie jak dana technika uczenia maszynowego sprawdza się **w kontekście danego typu zbioru**.

Często konieczna jest też wiedza dotycząca tego, jak w ocenie jakości działania algorytmu uwzględnić fakt, że posiadamy zbiór danych o **nierównych sobie liczebnościach klas**.

Tego typu zagadnienia wymagają szerszego podejścia do analizy skuteczności modelu uczenia maszynowego niż prosty podział danych treningowych na trzy podzbiory.

Problematyczne sytuacje

Przykładowe sytuacje w których konieczne jest **specyficzne podejście do oceny działania algorytmu:**

- przygotowywany jest algorytm wykonujący zadanie klasyfikacji i **liczba przykładów nie jest podobna dla wszystkich klas,**
- przygotowywany jest algorytm wykonujący zadanie regresji i **liczba przykładów nie jest równomierna dla całej dziedziny**, dla której obliczana jest regresja,
- musimy wybrać **lepszy z dwóch modeli, które nieznacznie różnią się** osiąganą jakością klasyfikacji/regresji,
- chcemy przewidzieć, **jak dany model zachowa się po napotkaniu różnorodnych danych** – na przykład po to by przygotować się na zjawisko niedopasowania modelu do danych rzeczywistych (ang. *data shift*),
- przygotować się na fakt, że model **będzie dotrenowywany w trakcie eksploatacji** i chcemy, aby osiągane efekty treningu nie miały dużego rozrzutu jakości i były zawsze możliwie wysokie.

Testowanie modelu uczenia maszynowego

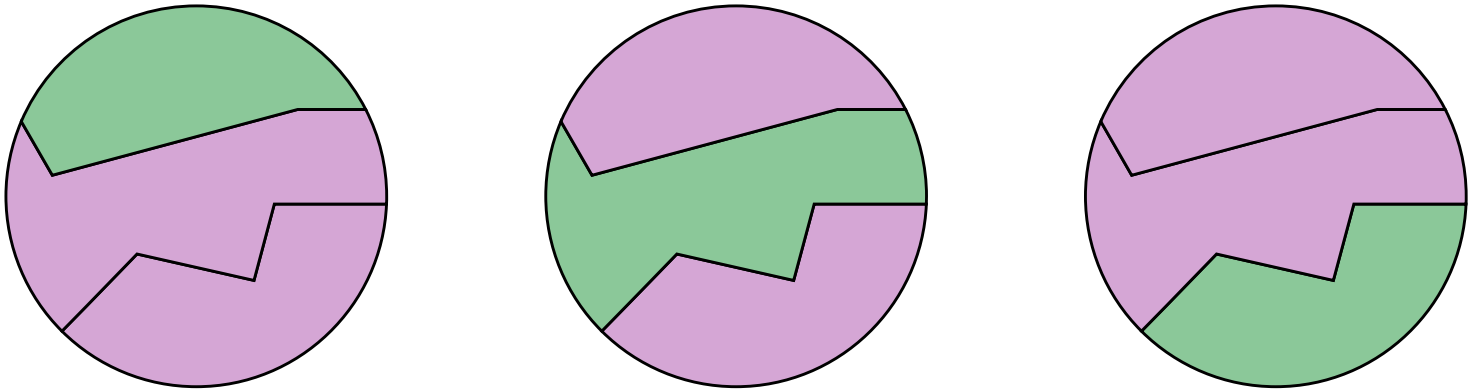
Jeżeli przygotowujemy duży, długo uczący się model, to mamy ograniczoną możliwość testowania jakości modelu poprzez jego wielokrotny trening. W takim przypadku pomocne jest:

- testowanie modelu za pomocą **kilku różnych zbiorów**,
- a także poprzez **testowanie na zbiorze podzielonym na kilka mniejszych** (równolicznych, ale często mogących mieć części wspólne) zbiorów.

Jeżeli mamy model, który możemy trenować wielokrotnie, to możemy posłużyć się następującymi metodami:

- ***k*-krotna walidacja krzyżowa** (*k-fold cross-validation*) – podział zbioru na *k* **rozłącznych, równolicznych** podzbiorów, trening modelu na każdym możliwym zestawie k-1 podzbiorów i test na jednym pozostawionym podzbiorze.
- **procedura 5x2CV** – 5-krotne powtórzenie procedury 2-krotnej walidacji – jest to optymalny ze statystycznego punktu widzenia dobór wartości *k*, oraz liczby powtórzeń procedury walidacji krzyżowej.

Przykład: 3-krotna walidacja krzyżowa

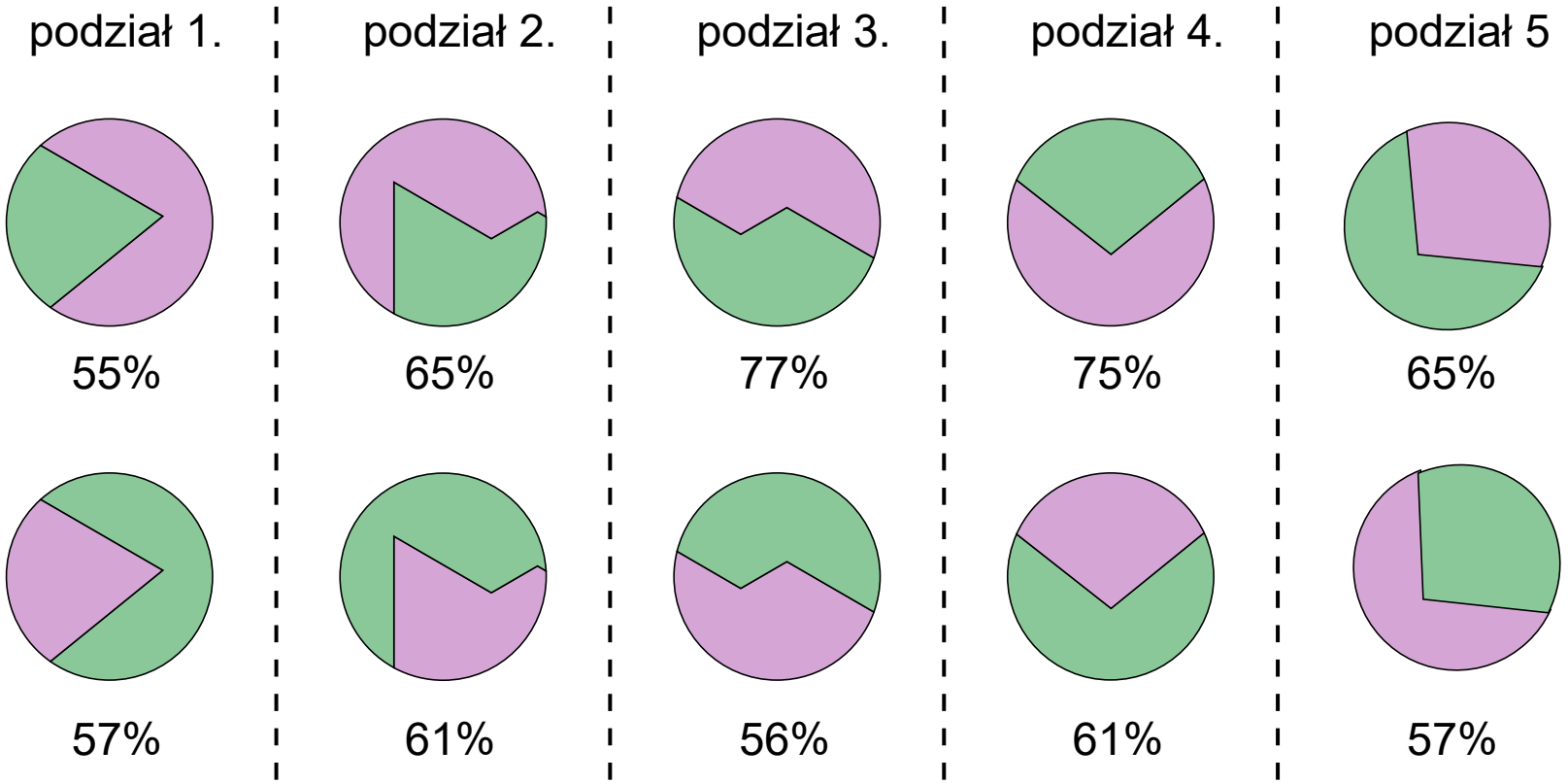


83% 91% 88%

zbiór treningowy
zbiór walidacyjny

wartościami procentowymi oznaczone są skuteczności klasyfikacji osiągnięte dla każdego sposobu przypisania zbiorów testowych i walidacyjnych

Przykład: procedura 5x2CV



zbiór treningowy
zbiór walidacyjny

wartościami procentowymi oznaczone są skuteczności klasyfikacji osiągnięte dla każdego sposobu przypisania zbiorów testowych i walidacyjnych

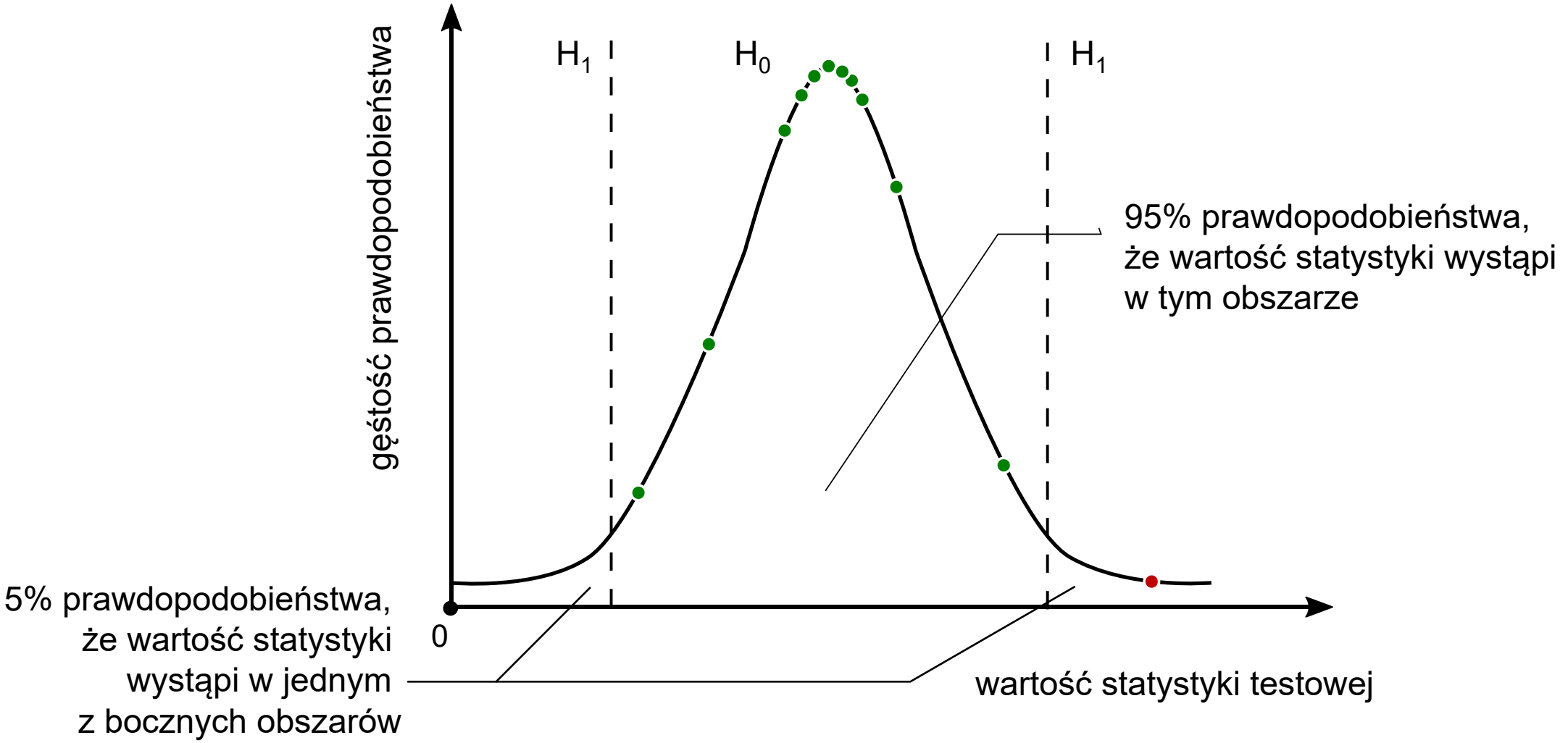
Analiza danych z procedur bazujących na walidacji krzyżowej

Walidacja krzyżowa i procedura 5x2CV dostarczają informacji nie tylko o dokładności modelu. Dzięki powtórzonemu wielokrotnie testowaniu modelu możliwe jest stwierdzenie jak bardzo powtarzalne są efekty treningu – pojawia się **informacja o rozrzucie uzyskiwanych wartości skuteczności**.

Dzięki takiemu podejściu łatwiej **porównywać między sobą klasyfikatory**, bo posiadanie jednocześnie miary np. średniej skuteczności modelu i informacji o wariancji skuteczności **umożliwia wykorzystanie testu statystycznego do stwierdzenia czy jeden z modeli radzi sobie istotnie statystycznie lepiej niż drugi.**

Test statystyczny polega na obliczeniu specjalnie przystosowanej do danego przypadku statystyki testowej i porównaniu jej z progiem (obliczanym na podstawie parametru α **- tzw. poziomu istotności**). Każdy test ma w **dwie wynikowe hipotezy** – **zerową** (H_0 - gdy statystyka jest mniejsza niż próg) i **alternatywną** (H_1 - gdy statystyka jest większa niż próg). Hipotezy mogą być różne w zależności od potrzeb (a tym samym – wybranego testu).

Testy statystyczne



Skuteczność klasyfikacji i macierze pomyłek

Najprostszą miarą tego, jak dobrze klasyfikator realizuje swoje zadanie jest **obliczenie np. procentowej poprawności (skuteczności) rozpoznawania** przez niego klas na podstawie dostarczonych danych walidacyjnych i testowych.

Klasyfikator operujący na zbiorze zawierającym taką samą liczbę przykładów dla każdej z klas zaczyna być skuteczny w momencie, gdy skuteczność jego klasyfikacji jest statystycznie **istotnie lepsza niż klasyfikacja przykładów poprzez losowe kwalifikowanie ich do jednej z możliwych klas.**

W udowodnieniu takiej tezy pomocne będą takie **techniki statystyczne jak np. przedziały ufności, test dwumianowy, czy test *t*-Studenta.** Służą one do określania na przykład, czy

- model **częściej** klasyfikuje dane poprawnie niż niepoprawnie (test **dwumianowy**),
- model A ma **wyższą średnią** dokładność klasyfikacji/regresji niż model B (test ***t*-Studenta**.

Skuteczność klasyfikacji i macierze pomyłek

Jeżeli zbiór zawiera **więcej niż 2 klasy,** to często zamiast pojedynczą liczbą operujemy za pomocą tzw. **macierzy pomyłek** (ang. *confusion matrix*), która pozwala na określenie między innymi na to, które klasy są dla algorytmu problematyczne (np. najczęściej są ze sobą mylone). Przykład macierzy pomyłek dla algorytmu realizującego rozpoznawanie pojazdów na drodze:

		Efekt klasyfikacji		
		ciężarowy	**osobowy**	**motocykl**
Klasa wejściowa (oczekiwana)	**ciężarowy**	20	2	3
	osobowy	3	19	3
	motocykl	5	4	16

Macierz pomyłek to przykład tzw. tablicy kontyngencji i stąd w udowodnieniu **istotności statystycznej wyników** pomocne będą np. test McNemara, test χ^2, czy dokładny test Fishera.

Miary skuteczności dla przypadku gdy klasy mają niezrównoważone liczności

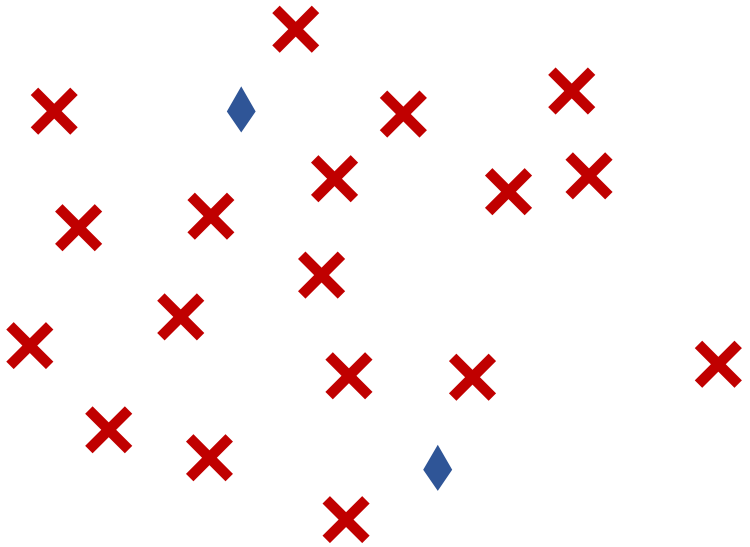
Czasami sama informacja o skuteczności klasyfikacji jest niewystarczająca. Może się tak zdarzyć nawet w przypadku, gdy skuteczność jest przedstawiona w postaci macierzy pomyłek.

Częstą sytuacją wymuszającą rezygnację z prostej miary, jaką jest skuteczność klasyfikacji, jest przypadek w którym realizowane jest zadanie klasyfikacji i **w wejściowym zbiorze danych liczebności przykładów powiązanych z poszczególnymi klasami bardzo od siebie odbiegają.**

Mówimy w takim przypadku, że mamy do czynienia ze **zbiorem danych o niezrównoważonej liczności klas** (ang. *imbalanced dataset*).

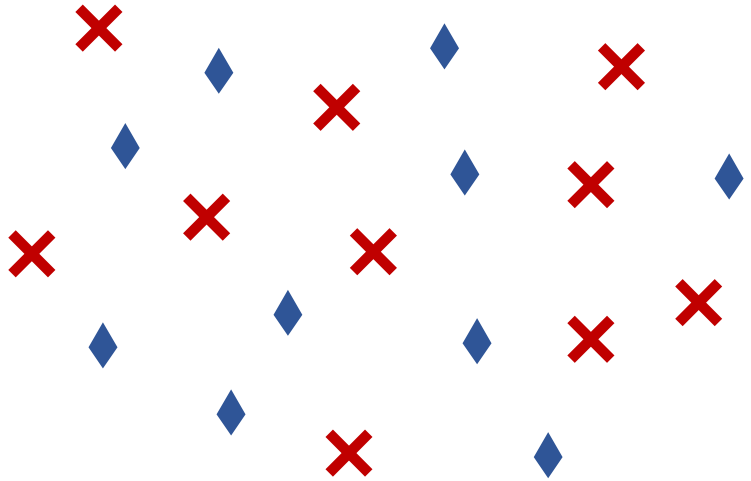
Oczywiście możliwe jest także **poradzenie sobie z tym problemem** poprzez przekształcenie zbioru danych, na przykład za pomocą **operacji upsamplingu i downsamplingu poszczególnych klas.**

Miary skuteczności dla przypadku gdy klasy mają niezrównoważone liczności



Jeśli zawsze zgaduję ✕, to mam 90% szans na to, że wybiorę poprawną odpowiedź.

Jeśli zawsze zgaduję ♦, to mam 10% szans na to, że wybiorę poprawną odpowiedź.



Jeśli zawsze zgaduję ✕, to mam 50% szans na to, że wybiorę poprawną odpowiedź.

Jeśli zawsze zgaduję ♦, to mam 50% szans na to, że wybiorę poprawną odpowiedź.

Miary skuteczności dla przypadku gdy klasy mają niezrównoważone liczności

Zwykła skuteczność może być myląca w przypadku danych, w których **klasy nie są charakteryzowane przez podobne liczby przykładów**. Stąd konieczne jest posłużenie się **specyficznymi miarami dokładności**.

- częstość prawdziwych pozytywów (ang. *true positive rate,* ***TPR***),
- częstość prawdziwych negatywów (ang. *true negative rate,* ***TNR***),
- częstość fałszywych pozytywów (ang. *false positive rate,* ***FPR***),
- częstość fałszywych negatywów (ang. *false negative rate,* ***FNR***),
- **precyzja** (ang. *precision*), która jest tym mniejsza, im większa jest szansa na pomyłkę modelu, gdy ten dokonuje detekcji danej klasy. Oblicza się ją jako:

$$\text{precyzja} = \frac{\text{TPR}}{\text{TPR} + \text{FPR}},$$

Miary skuteczności dla przypadku gdy klasy mają niezrównoważone liczności (c.d.)

- **czułość** (ang. *recall*), która jest tym mniejsza im większej liczby przykładów model nie wykryje:

$$\text{czułość} = \frac{\text{TPR}}{\text{TPR} + \text{FNR}}$$

- **miara F** (ang . *F-score, F-score*), która łączy w jedną liczbę wartości precyzji oraz czułości:

$$F = \frac{2 \cdot \text{precyzja} \cdot \text{czułość}}{\text{precyzja} + \text{czułość}}$$

Przytoczone metryki mogą być **obliczane zarówno dla każdej z klas osobno, jak i dla całego modelu** poprzez sumowanie składowych wartości np. częstości fałszywych pozytywów otrzymanych dla wszystkich rozpoznawanych przez model klas.

Krzywa ROC

Miarą skuteczności stosowaną w sytuacji, gdy mamy do czynienia z systemem dokonującym **klasyfikacji za pomocą progu** jest krzywa ROC (ang. *receiver operation characteristic*).

Krzywą tę uzyskuje się poprzez **wykreślenie częstości poprawnych detekcji** systemu (ang. *true positives rate, TPR*) **w funkcji częstości fałszywych** detekcji (ang. *false positives rate, FPR*).

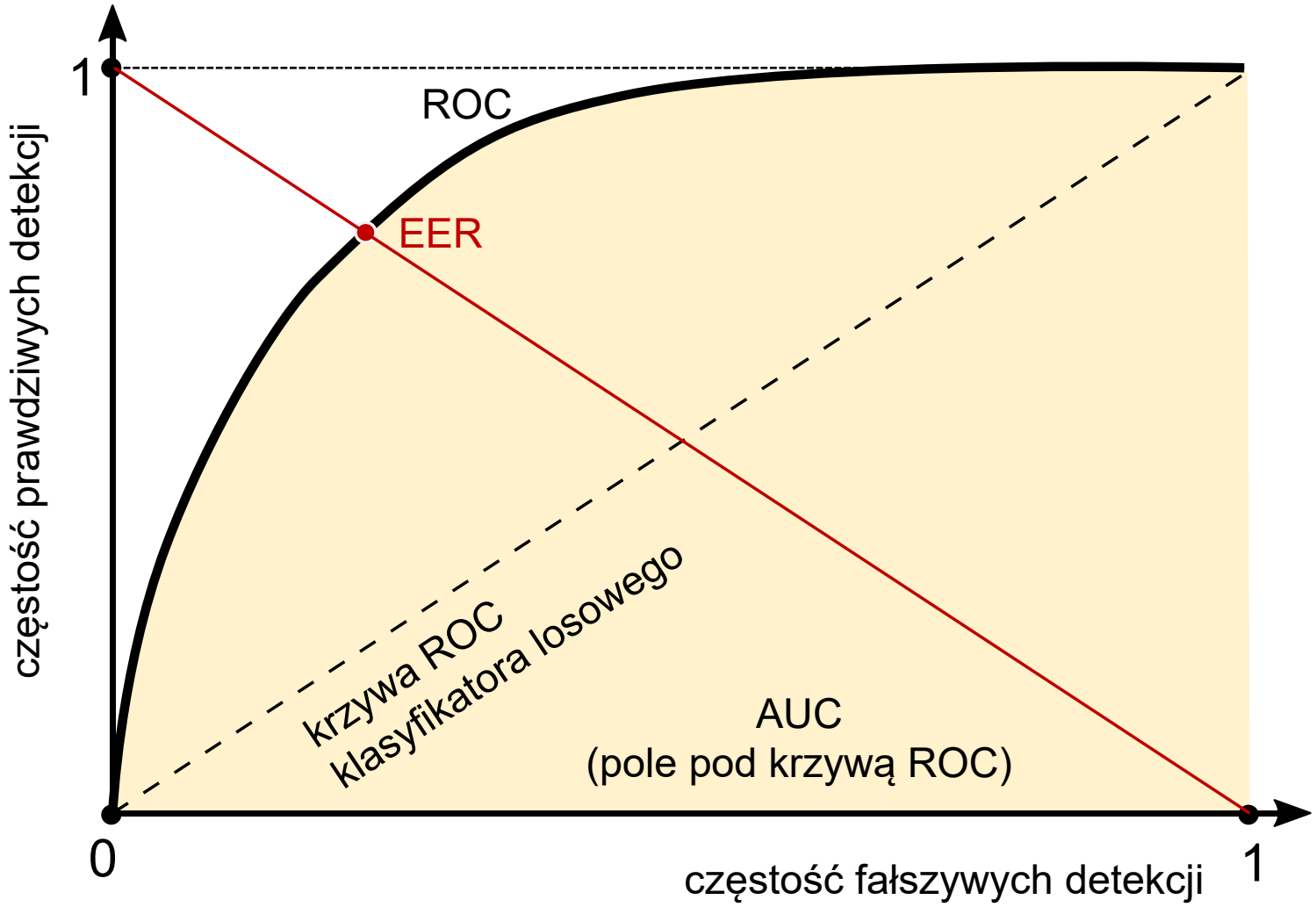
Krzywa ta jest szczególnie **popularna w przypadku oceny systemów biometrycznych**, gdzie sprawdza się podobieństwo badanej cechy biometrycznej pobranej od osoby weryfikowanej z wzorcami w bazie danych. **Obliczana jest miara podobieństwa, która następnie jest porównywana z progiem** w celu stwierdzenia, czy wartość danej cechy należy do danej osoby z bazy, czy nie.

Krzywa ROC

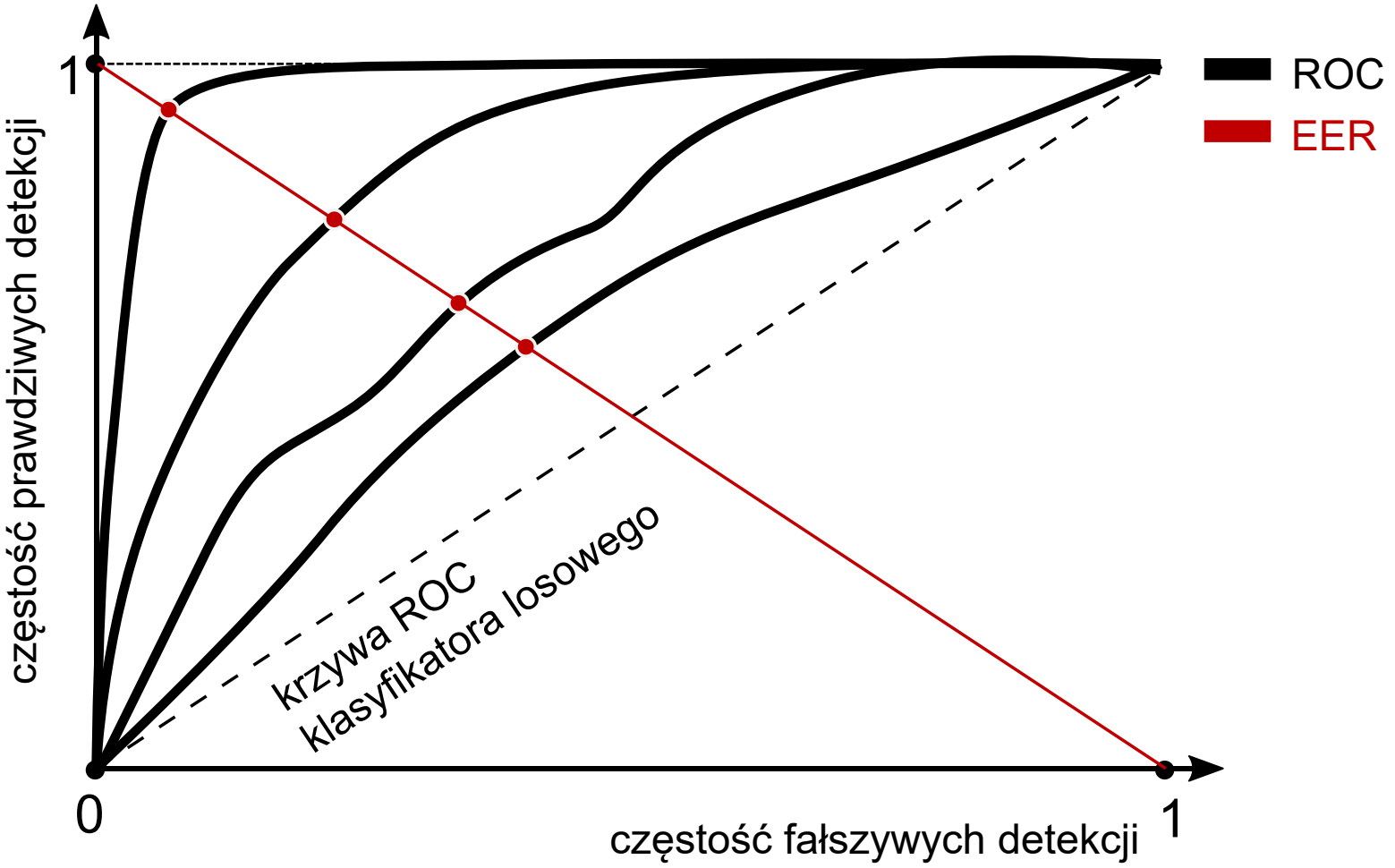
Na podstawie krzywej ROC można wyróżnić dwie miary pochodne:

- **AUC (ang. area under curve)**, czyli **pole pod wykresem** krzywej ROC, jest to wygodna miara która za pomocą pojedynczej liczby charakteryzuje jak bardzo dokładny jest system. Dokładność jest tym wyższa im wyższe jest pole pod wykresem,
- **EER (ang. equal error rate)**, którą definiuje się poprzez **znalezienie takiego punktu** dla krzywej ROC, dla którego **częstość fałszywych pozytywów** (ang. false positives rate, FPR) **jest równa częstości fałszywych negatywów** (ang. false negatives rate, FNR). System jest tym lepszy im niższa jest ta miara.

Krzywa ROC i miary pokrewne



Krzywa ROC i miary pokrewne



Przykładowe miary skuteczności dla regresji

Błąd średniokwadratowy (ang. *mean squared error, MSE*) – prosta miara skuteczności, która jest bardzo powszechnie stosowana. Z powodu zastosowania kwadratu błędu jest ona bardziej czuła na przypadku, gdy model popełnia błędy o znacznych wartościach:

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}\left(y_{true,i} - y_{pred,i}\right)^2$$

gdzie:

N oznacza liczbę przykładów w zbiorze danych,

i to indeks przykładu pochodzącego ze zbioru danych,

$y_{true,i}$ to przykład oczekiwanej wartości, jaką powinien zwrócić algorytm dla i-tego przykładu uczącego,

$y_{pred,i}$ to wartość zwrócona przez algorytm dla i-tego przykładu uczącego.

Przykładowe miary skuteczności dla regresji

Średni błąd absolutny (ang. *mean absolute error, MAE*) – miara skuteczności regresji, która charakteryzuje się tym, że nie faworyzuje błędów o wysokich wartościach:

$$\mathrm{MAE} = \frac{1}{N}\sum_{i=1}^{N}\left|y_{true,i} - y_{pred,i}\right|$$

gdzie:

N oznacza liczbę przykładów w zbiorze danych,

i to indeks przykładu pochodzącego ze zbioru danych,

$y_{true,i}$ to przykład oczekiwanej wartości, jaką powinien zwrócić algorytm dla i-tego przykładu uczącego,

$y_{pred,i}$ to wartość zwrócona przez algorytm dla i-tego przykładu uczącego.

Przykładowe miary skuteczności dla regresji

Podobieństwo kosinusowe (ang. *cosine similarity, CS*) – miara skuteczności regresji, która nie uwzględnia tak silnie jak MSE, czy MAE długości wektora danych, co jest przydatne np. w analizie danych tekstowych:

$$\mathrm{CS} = \sum_{i=1}^{N} y_{true,i} \cdot y_{pred,i} = \boldsymbol{Y_{true}} \cdot \boldsymbol{Y_{pred}}$$

gdzie

N oznacza liczbę przykładów w zbiorze danych,

i to indeks przykładu pochodzącego ze zbioru danych,

$\boldsymbol{Y_{true}}$ to wektor zawierający oczekiwane wartości, jaką powinien zwrócić algorytm, pojedyncza i-ta wartość z tego wektora to $y_{true,i}$,

$\boldsymbol{Y_{pred}}$ to wektor zawierający wartości zwrócony przez algorytm , pojedyncza i-ta wartość z tego wektora to $y_{pred,i}$,

Wartość miary może być także wyrażona jako $\mathrm{CS} = |\boldsymbol{Y_{true}}||\boldsymbol{Y_{pred}}|\cos(\alpha)$, gdzie α to kąt pomiędzy wektorami $\boldsymbol{Y_{true}}$ i $\boldsymbol{Y_{pred}}$, stąd nazwa tej miary.

Literatura

1. Bostanci , B., Bostanci, E., An evaluation of classification algorithms using Mc Nemar's test, Adv. Intell. Syst. Comput. (201) 1,15–26, 2013, doi: 10.1007/978-81-322-1038-2_2.
2. Bouckaert, R. R., Choosing between two learning algorithms based on calibrated tests, Proceedings, Twent. Int. Conf. Mach. Learn. (1), 51–58, 2003.
3. Demšar. J., Statistical comparisons of classifiers over multiple data sets., J. Mach. Learn. Res. (7), 1–30, 2006.
4. Dietterich, T. G., Approximate Statistical Tests for Comparing Supervised Classification Learning Algorithms, Neural Comput. (10) 7, 1895–1924, 1998.
5. García , S., Fernández , A., Luengo , J., Herrera, F., Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power, Inf. Sci., (180) 10, 2044–2064, 2010, doi: 10.1016/j.ins.2009.12.010.
6. Geron, A., Hands-On Machine Learning with Scikit-Learn & TensorFlow. Concepts, Tools, and Techniques to Build Intelligent Systems, O'Reilly Media, Inc., 2019.

Literatura

7. Goodfellow, I., Bengio, J., Courville, A., Deep Learning, The MIT Press, 2016.
8. Heaven, W. D., The way we train AI is fundamentally flawed, MIT Technology Review, artykuł dostępny w sieci Internet pod adresem: https://www.technologyreview.com/2020/11/18/1012234/training-machine-learning-broken-real-world-heath-nlp-computer-vision/
9. Herrera , F., García , S., An Extension on ‘Statistical Comparisons of Classifiers over Multiple Data Sets’ for all Pairwise Comparisons, J. Mach. Learn. Res. (9), 2677–2694, 2008.
10. Raschka, S., Model evaluation, model selection, and algorithm selection in machine learning, arXiv preprint, arXiv id: 1811.12808, 2018.

POLITECHNIKA GDAŃSKA

AI TECH

Dziękuję

Adam Kurowski

Rzeczpospolita Polska

Unia Europejska
Europejski Fundusz
Rozwoju Regionalnego

Projekt współfinansowany ze środków Unii Europejskiej w ramach Europejskiego Funduszu Rozwoju Regionalnego
Program Operacyjny Polska Cyfrowa na lata 2014-2020.
Oś priorytetowa nr 3 „Cyfrowe kompetencje społeczeństwa”, działanie nr 3.2 „Innowacyjne rozwiązania na rzecz aktywizacji cyfrowej”.
Tytuł projektu: „Akademia Innowacyjnych Zastosowań Technologii Cyfrowych (AI Tech)”.