

# Techniki eksploracji danych

**Krzysztof Gajowniczek**

Rok akademicki: 2020/2021

- 1 Wykorzystywana notacja matematyczna
- 2 Miary jakości modeli
- 3 Sprawdzian krzyżowy
- 4 Optymalizacja parametrów
- 5 Literatura

## Section 1

# Wykorzystywana notacja matematyczna

## Wyjaśnienie symboli

- $n$  - liczba obserwacji.
- $i$  - iterator obserwacji,  $i \in \{1, \dots, n\}$ .
- $y$  - zmienna celu.
  - dla problemu regresji:  $y \in \mathbb{R}$ .
  - dla problemu klasyfikacji:  $y \in \{1, \dots, l, \dots, k\}$ .
  - $y_i$  -  $i$ -ta wartość zmiennej celu.
- $\hat{y}$  - wartość teoretyczna zmiennej celu.
  - dla problemu regresji:  $\hat{y} \in \mathbb{R}$ .
  - dla problemu klasyfikacji:  $\forall l, \hat{y} \in [0, 1], \sum_{l=1}^k \hat{y}^{(l)} = 1$ .
- $x$  - atrybut,  $x \in X$ .
- $j$  - iterator atrybutu,  $j \in \{1, \dots, p\}$ .
  - $\mathbf{x}^T \in X^{(p)}$ .

## Section 2

# Miary jakości modeli

## Subsection 1

### Miary dla problemów regresyjnych

## Mean absolute error

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i|$$

## Mean squared error

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

## Mean absolute percentage error

$$MAPE = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|$$

## Subsection 2

### Miary dla problemów klasyfikacyjnych



- W zależności o wykorzystanej miary, pracować będziemy albo na wektorach prawdopodobieństwa:

$$\hat{\mathbf{y}}^T = \{\hat{y}^{(1)}, \hat{y}^{(2)}\} = \{0.35, 0.65\}$$

- albo na finalnych etykietach klas, wyznaczanych jako:

$$\hat{y} = \arg \max_l \hat{y}^{(l)} = \{2\}$$

- w niektórych jednak przypadkach dla klasyfikacji binarnej finalna etykieta wyznaczana jest jako:

$$\hat{y} = \arg_{l>t} \hat{y}^{(l)} = \{2\}$$

- etykieta, której prawdopodobieństwo jest większe od pewnego progu  $t$ , najczęściej ustawionego na wartość 0.5.

# Klasyfikacja binarna

- Klasa pozytywna (P) jest to ta klasa, która nas najbardziej interesuje.
- Najczęściej jest niedoreprezentowana, tj. 5%-95%.

		Przewidywane	
		P (2)	N (1)
Prawdziwe	P (2)	TP	FN
	N (1)	FP	TN

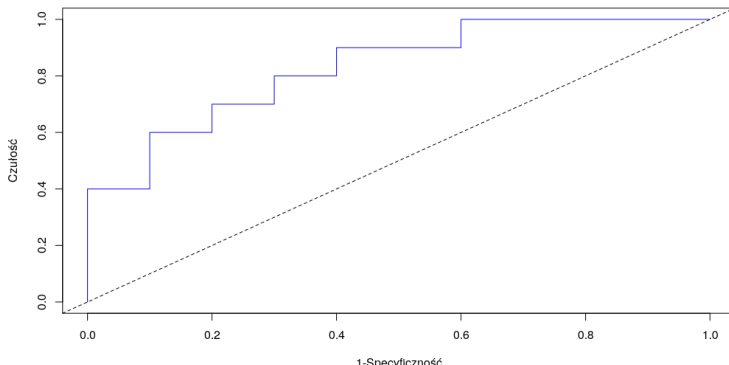
$$\text{Czułość} = \frac{TP}{TP + FN}$$

$$\text{Specyficzność} = \frac{TN}{TN + FP}$$

$$\text{Jakość} = \frac{TN + TP}{TN + TP + FP + FN}$$

# Klasyfikacja binarna

- Krzywa ROC (ang. Receiver Operating Characteristic):  
graficzna reprezentacja efektywności modelu poprzez  
wykreślenie miar *Czułości* i *Specyficzności* powstałych z  
modelu przy zastosowaniu wielu różnych punktów odcięcia  $t$ .



## Pole pod krzywą ROC

- AUC (ang. Area Under Curve) - wielkość pola pod krzywą ROC mieści się w przedziale  $[0; 1]$ .

### Interpretacja jak pole

$$AUC = \int_0^1 ROC(t) dt$$

### Interpretacja jako prawdopodobieństwo

$$AUC = P(\hat{y}^{(2)} > \hat{y}^{(1)})$$

## Metody estymacji AUC

- Istnieje kilka różnych sposobów oszacowania tej miary.
- Obserwacje muszą być posortowane nierosnąco względem  $\hat{y}^{(2)}$ .

### Interpretacja jak pole

$$AUC = -\frac{1}{2} \sum_{i=2}^n (Czułość_{i-1} Specyficzność_i - Czułość_i Specyficzność_{i-1})$$

### Interpretacja jako prawdopodobieństwo

$$AUC = \frac{R_2 - \frac{n_2(n_2-1)}{2}}{n_2 n_1}$$

gdzie:  $n_2$  - liczba obserwacji z klasy 2,  $n_1$  - liczba obserwacji z klasy 1,  $R_2$  - suma rang obserwacji z klasy 2.

# Klasyfikacja wieloklasowa

		Przewidywane					
		1	2	...	j	...	k
Prawdziwe	1	T	F	...	F	...	F
	2	F	T	...	F	...	F
	...	...	...	...	...	...	...
	j	F	F	...	T	...	F
	...	...	...	...	...	...	...
	k	F	F	...	F	...	T

$$Jakość = \frac{1}{n} \sum_{i=1}^n I(y_i = \arg \max_l \hat{y}_i^{(l)})$$

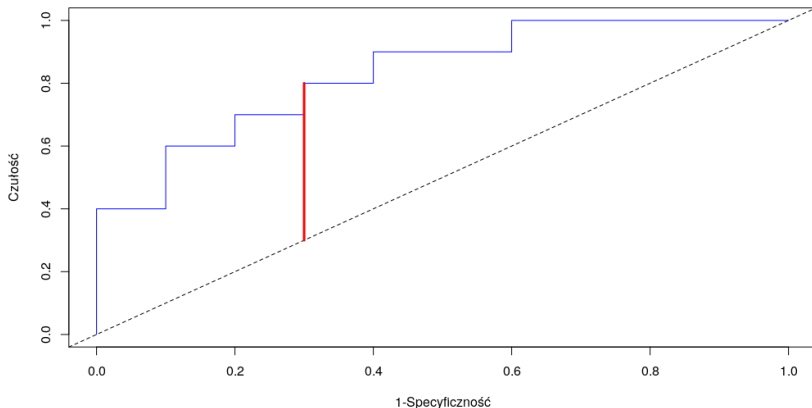
$$multiAUC = \frac{1}{k(k-1)} \sum_{l,k} AUC_{l,k}$$



## Youden index (J)

- Wyznaczenie optymalnej wartości  $t$  wyznacza się na podstawie:

$$J = \arg_t \text{Czułość} + \text{Specyficzność} - 1$$



## Section 3

# Sprawdzian krzyżowy

- Sprawdzian krzyżowy (walidacja krzyżowa, kroswalidacja, sprawdzanie krzyżowe) – metoda statystyczna, polegająca na podziale próby statystycznej na podzbiory, a następnie przeprowadzaniu wszelkich analiz na niektórych z nich (zbiór uczący), podczas gdy pozostałe służą do potwierdzenia wiarygodności jej wyników (zbiór testowy, zbiór walidacyjny).

## Prosta walidacja

- Najbardziej typowy rodzaj walidacji, w którym próbę dzieli się losowo na rozłączne zbiory:
  - Uczący i testowy najczęściej w proporcjach 80%-20% lub 70%-30%.
  - Uczący, walidacyjny i testowy najczęściej w proporcjach 60%-30%-10%.

## k-krotna walidacja

- Oryginalna próba jest dzielona na  $k$  podzbiorów, gdzie każdy z nich bierze się jako zbiór testowy, a pozostałe razem jako zbiór uczący. Analiza jest więc wykonywana  $k$  razy.  $K$  rezultatów jest następnie uśrednianych (lub łączonych w inny sposób) w celu uzyskania jednego wyniku.
  - Proporcja zbiorów wynosi  $(1-1/k)\%-(1/k)\%$ , np. dla  $k = 4$ , 75%-25%;  $k = 10$ , 90%-10%.

## Kroswalidacja stratyfikowana

- Nie jest to w zasadzie osobna odmiana kroswalidacji, a odnosi się do wszystkich jej rodzajów wymienionych powyżej.
- Kroswalidacja stratyfikowana (ang. stratified cross-validation) polega na takim podziale obiektów pomiędzy zbiór uczący i zbiór testowy, aby zachowane były oryginalne proporcje pomiędzy klasami decyzyjnymi.
- Zastosowanie kroswalidacji stratyfikowanej jest szczególnie ważne w przypadku, gdy w oryginalnym zbiorze danych występują znaczne dysproporcje w liczebności przykładów należących do poszczególnych klas decyzyjnych.

## Section 4

# Optymalizacja parametrów

- Modele uczenia maszynowego są sparametryzowane, aby ich zachowanie można było dostosować do danego problemu.
- Modele mogą mieć wiele parametrów, a znalezienie najlepszej kombinacji parametrów może być traktowane jako problem wyszukiwania.
- Proces ten nazywany jest optymalizacją hiperparametrów, gdzie parametry algorytmu są określane jako hiperparametry, podczas gdy współczynniki znalezione przez sam algorytm uczenia maszynowego są określane jako parametry.
- Optymalizacja wskazuje na poszukiwawczy charakter problemu.
- Określając ten proces jako problem wyszukiwania, można użyć różnych strategii wyszukiwania, aby znaleźć dobry i niezawodny parametr lub zestaw parametrów dla algorytmu dla danego problemu.

## Dostrajanie parametrów poprzez przeszukiwanie siatki (ang. grid search)

- Przeszukiwanie siatki to podejście do strojenia parametrów, które metodycznie buduje i ocenia model dla każdej kombinacji parametrów algorytmu określonych w siatce.

## Losowe dostrajanie parametrów (ang. random search)

- Wyszukiwanie losowe to podejście do dostrajania parametrów, które będzie próbkować parametry algorytmu z losowego rozkładu jednostajny) dla ustalonej liczby iteracji.
- Model jest konstruowany i oceniany dla każdej kombinacji wybranych parametrów.



## Section 5

### Literatura

- *Friedman, J., Hastie, T., Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). New York: Springer series in statistics.*
- *David J. Hand and Robert J. Till (2001). A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems. Machine Learning 45(2), 171-186.*