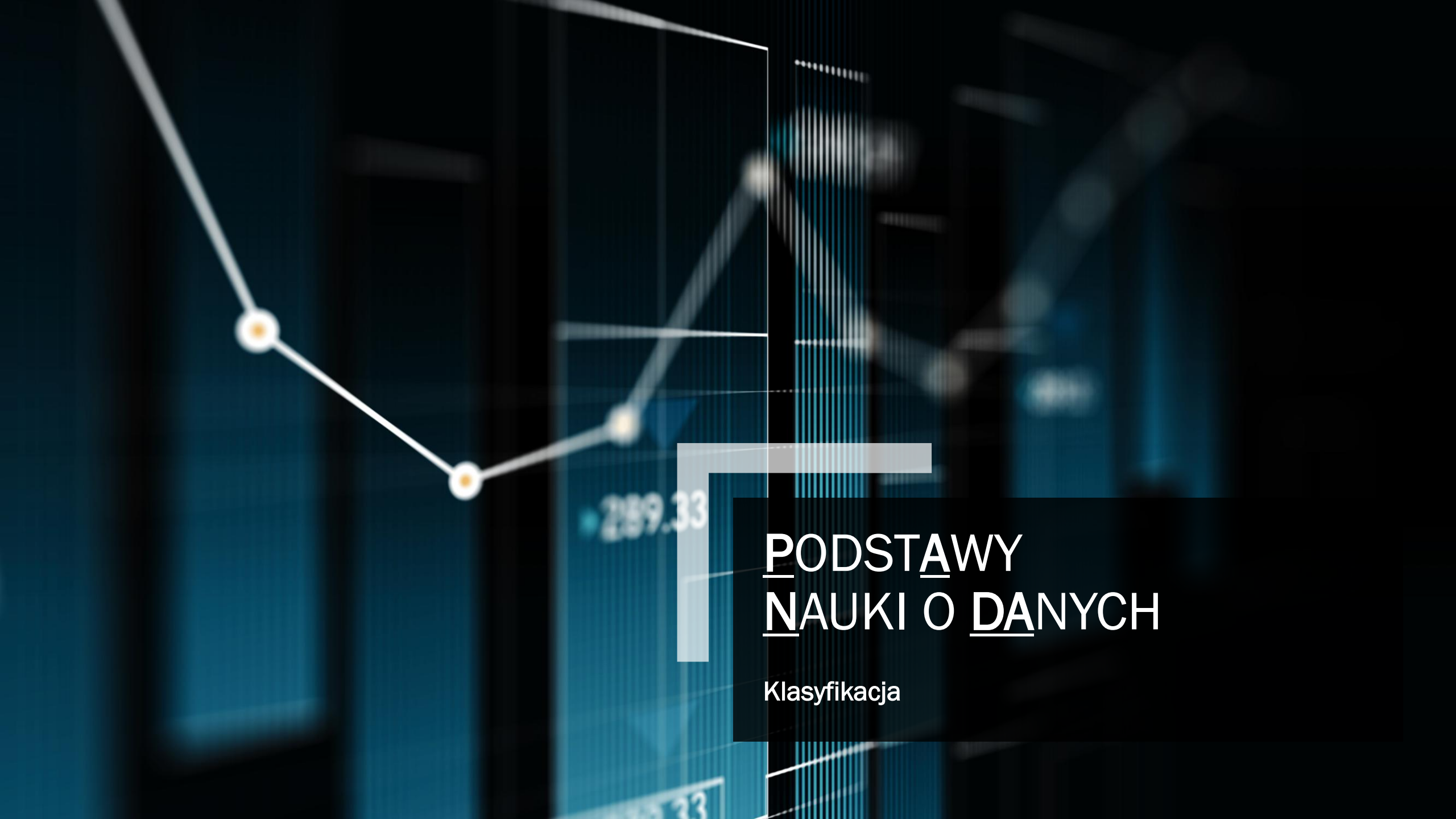




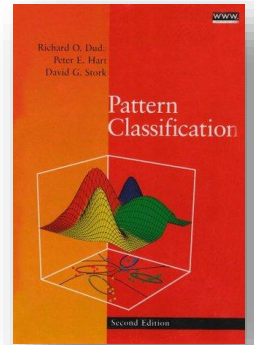
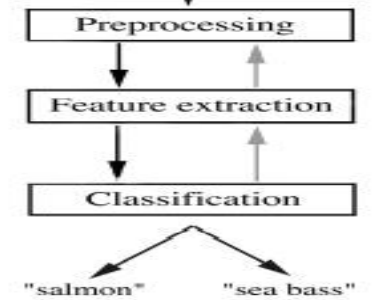
PODSTAWY NAUKI O DANYCH

Klasyfikacja

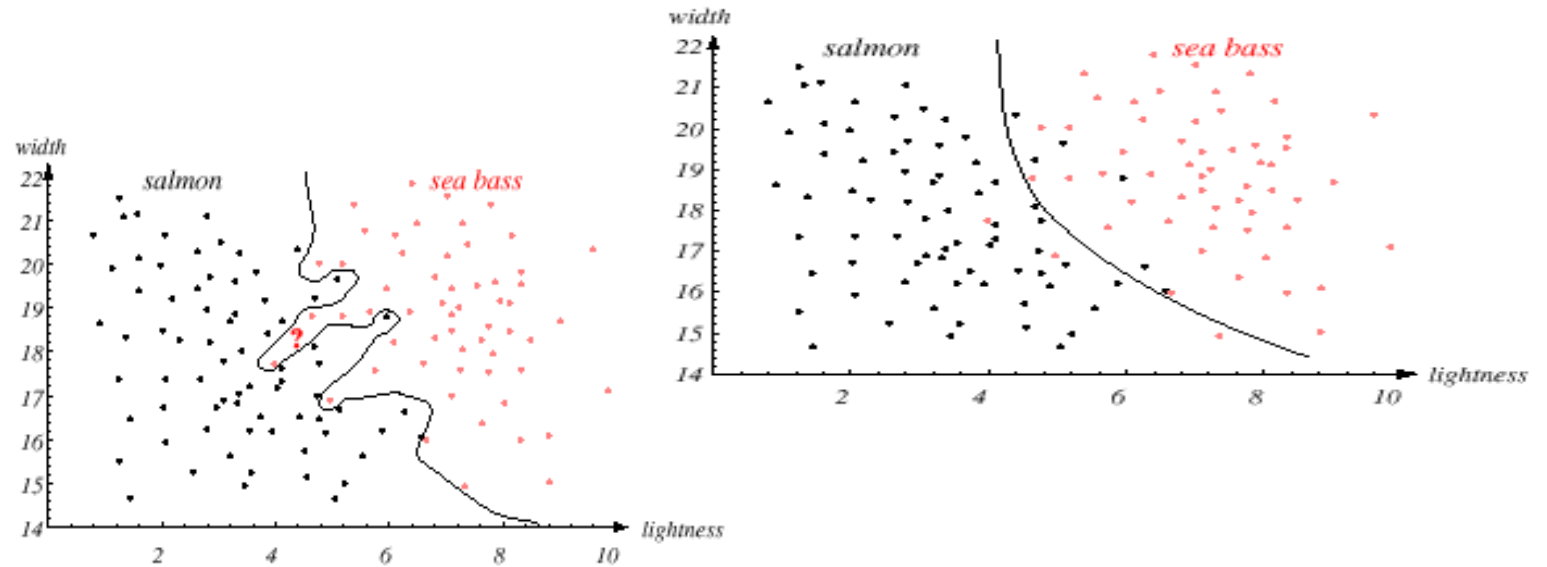
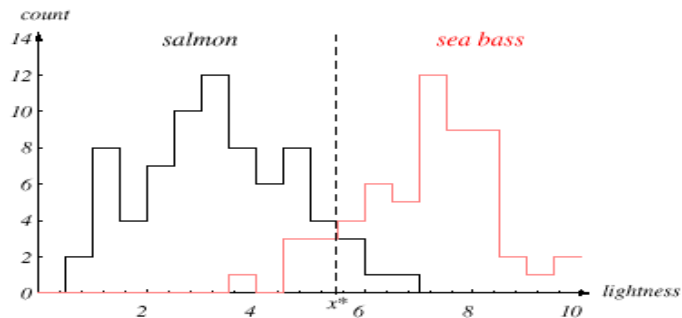
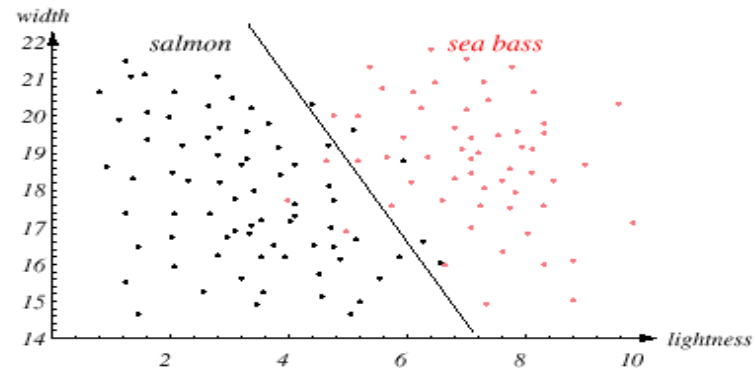
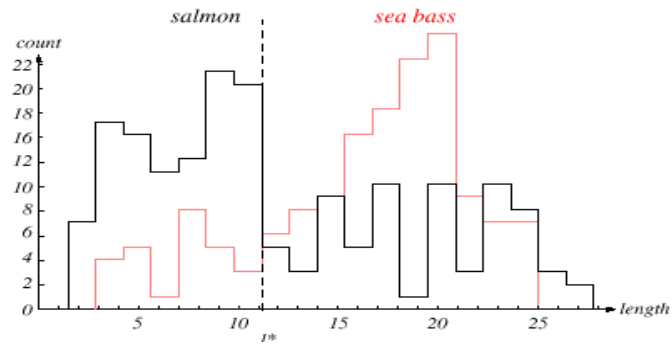
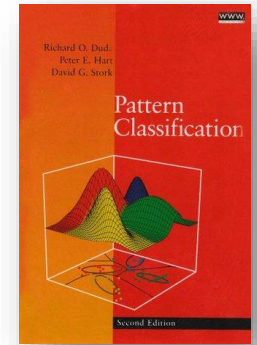
Klasyfikacja

Cechy:

- ✓ długość
- ✓ jasność
- ✓ szerokość
- ✓ liczba i kształt płetw
- ✓ położenie otworu gębowego



Klasyfikacja



Klasyfikacja

Dane
pomiarowe



$$\{(\mathbf{x}_i, C_i)\}_{i=1}^N$$

Wektor cech

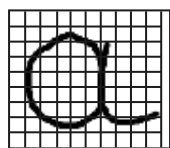
$\mathbf{x}$

algorytm klasyfikacji

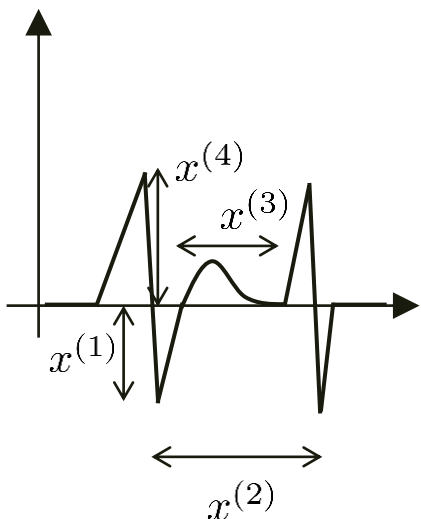
„Odgadnięta”
klasa

$\hat{C}$

■ Cechy



$$\begin{bmatrix} 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix} \equiv \mathbf{x}$$



$$\begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(S)} \end{bmatrix} \equiv \mathbf{x}$$

■ Przykład algorytmu klasyfikacji (modelu)

$$\hat{C} = \varphi(\theta^T \mathbf{x})$$

$$\phi(w) = \begin{cases} 1 & \text{jeżeli } w > 0 \\ 0 & \text{jeżeli } w \leq 0 \end{cases}$$

■ Wektor parametrów klasyfikatora

$$\theta \in \Re^S$$

UJĘCIE PROBABILISTYCZNE



Pełna informacja probabilistyczna

wektor cech

$$\mathbf{x} = \begin{bmatrix} x^{(1)} \\ \vdots \\ x^{(S)} \end{bmatrix}$$

etykiety klas

$$\{C_1, C_2, \dots, C_M\}$$

$$p(\mathbf{x}, C)$$

$$p(x^{(1)}|C)$$

$\vdots$

$$p(x^{(S)}|C)$$

$p(\mathbf{x})$ – rozkład cech

$p(C)$ – rozkład klas

$p(\mathbf{x}|C)$ – rozkład cech w klasach

$p(\mathbf{x}|C)$ – model generatywny

$p(C|\mathbf{x})$ – model dyskryminatywny

$$p(C|\mathbf{x}) = \frac{p(\mathbf{x}|C) \cdot p(C)}{p(\mathbf{x})}$$

Optymalny algorytm Bayesa

$$\hat{m} = \arg \max_m p(C_m | \mathbf{x})$$

$$p(C | \mathbf{x}) = \frac{p(\mathbf{x} | C) \cdot p(C)}{p(\mathbf{x})}$$

$$p(C_m | \mathbf{x}) \propto p(\mathbf{x} | C_m) \cdot p(C_m)$$

$$\log p(C | \mathbf{x}) = \log p(\mathbf{x} | C) + \log p(C) - \log p(\mathbf{x})$$

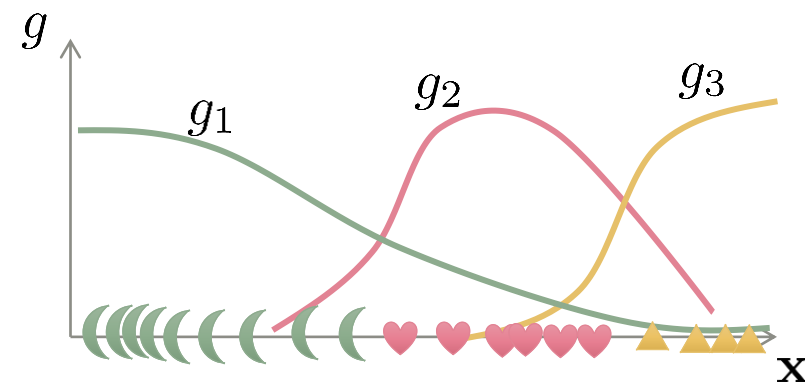
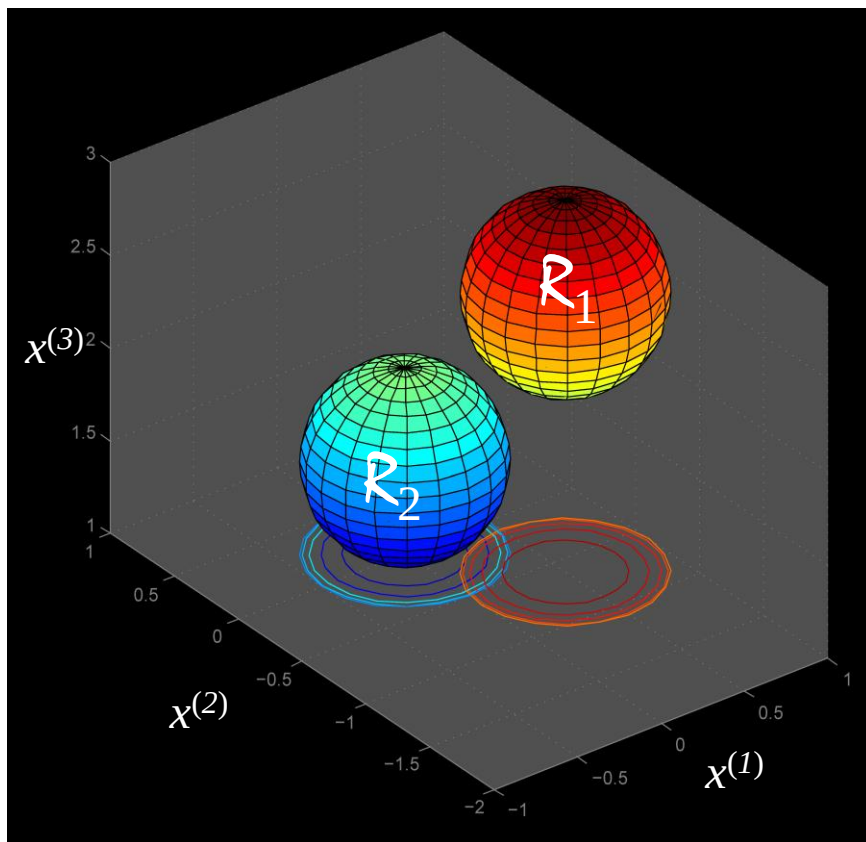
funkcja dyskryminująca :

$$g_m(\mathbf{x}) = \log p(\mathbf{x} | C_m) + \log p(C_m)$$

powierzchnia rozdzielająca
/ granica decyzyjna :

$$s_{rj}(\mathbf{x}) = g_r(\mathbf{x}) - g_j(\mathbf{x}) = \log \frac{p(\mathbf{x} | C_r)}{p(\mathbf{x} | C_j)} + \log \frac{p(C_r)}{p(C_j)} = 0$$

Co się dzieje w przestrzeni cech



Przypominajka statystyczna

$$p(AB) = p(A) \cdot p(B|A)$$



A – katar
B – kurs akcji

■ Bezwarunkowa niezależność

$$B \perp A$$

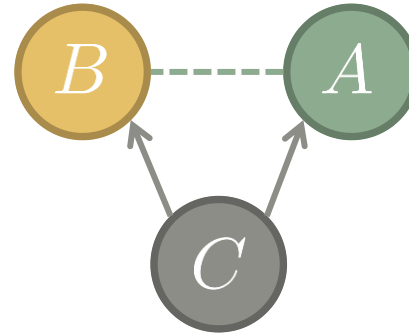


$$p(AB) = p(A) \cdot p(B)$$

lub

$$p(A|B) = p(A) \text{ oraz } p(B|A) = p(B)$$

$$p(AB|C) = p(A|C) \cdot p(B|AC)$$



A – gorączka
B – bóle mięśni
C – grypa

A – podbite oko
B – rozcięta warga
C – bójka

■ Warunkowa niezależność

$$A \perp\!\!\!\perp B \mid C$$



$$p(AB|C) = p(A|C) \cdot p(B|C)$$

lub

$$p(A|B, C) = p(A|C) \text{ oraz } p(B|A, C) = p(B|C)$$

Warunkowa niezależność cech

$$\begin{aligned} p(\mathbf{x}) &= p\left(x^{(1)}, \dots, x^{(S)}\right) = \\ &= p\left(x^{(2)}, \dots, x^{(K)} \mid x^{(1)}\right) \cdot p\left(x^{(1)}\right) = \\ &= p\left(x^{(3)}, \dots, x^{(S)} \mid x^{(1)}, x^{(2)}\right) \cdot p\left(x^{(2)} \mid x^{(1)}\right) \cdot p\left(x^{(1)}\right) = \\ &= \prod_{j=1}^S p\left(x^{(j)} \mid x^{(2)}, \dots, x^{(j-1)}\right) \end{aligned}$$



$$p\left(x^{(i)} \mid x^{(j)}\right) = p\left(x^{(i)}\right)$$

$$p(\mathbf{x}) = \prod_{j=1}^S p\left(x^{(j)}\right)$$

Naiwny algorytm Bayesa

Optymalny algorytm Bayesa: $\hat{m} = \arg \max_m p(C_m | \mathbf{x}) = \arg \max_m \boxed{p(\mathbf{x} | C_m)} \cdot p(C_m)$

Naiwnie zakładamy niezależność cech w klasach

$$\boxed{p(\mathbf{x} | C_m)} = p(x^{(1)} | C_m) \cdot p(x^{(2)} | C_m) \cdot \dots \cdot p(x^{(S)} | C_m)$$

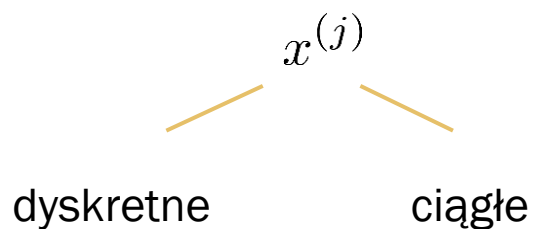
$$\hat{m} = \arg \max_m p(C_m) \prod_{j=1}^S \boxed{p(x^{(j)} | C_m)}$$

jaki rozkład cech w klasach?

GaussianNB

BernoulliNB

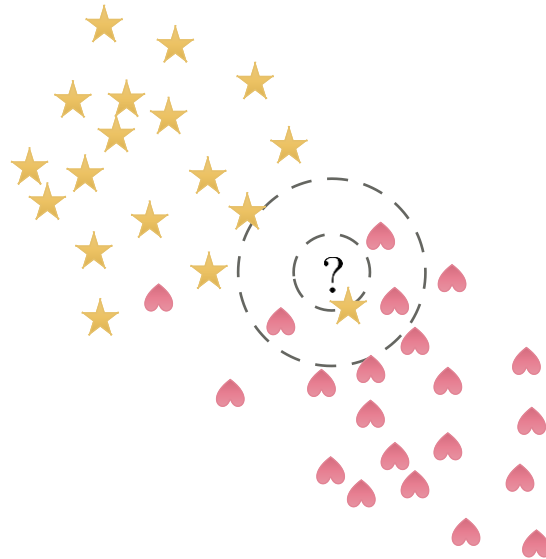
MultinomialNB



	wiek	dochód	studiuje	kupił sztabkę złota
1	młody	wysoki	nie	nie
2	młody	wysoki	nie	nie
3	produkcyjny	wysoki	nie	tak
4	senior	przeciętny	nie	tak
5	senior	niski	tak	tak
6	senior	niski	tak	nie
7	produkcyjny	niski	tak	tak
8	młody	przeciętny	nie	nie
9	młody	niski	tak	tak
10	senior	przeciętny	tak	tak
11	młody	przeciętny	tak	tak
12	produkcyjny	przeciętny	nie	tak
13	produkcyjny	wysoki	tak	tak
14	senior	przeciętny	nie	nie

$$P(\text{wiek} = \textit{młody} \mid \text{kupił} = \textit{tak})$$

Algorytm k -najbliższych sąsiadów

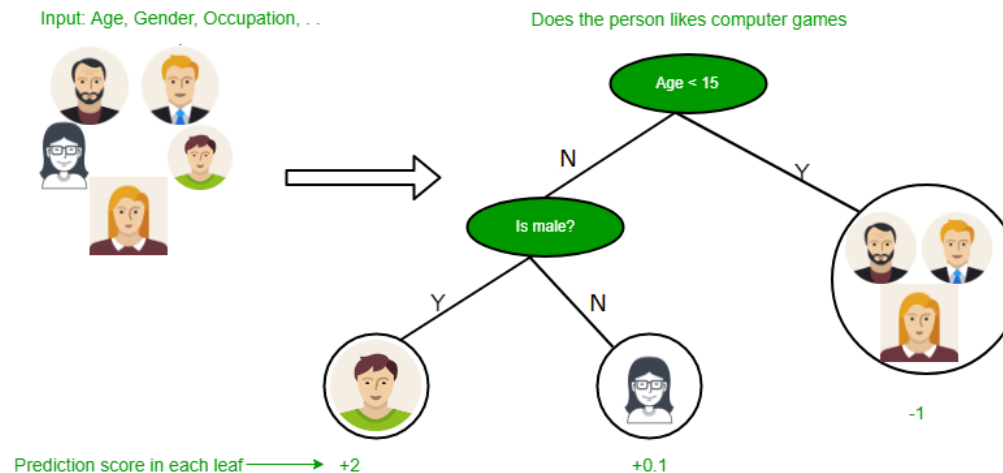


- stały dostęp do bazy danych
- drzewo poszukiwań



Drzewa decyzyjne

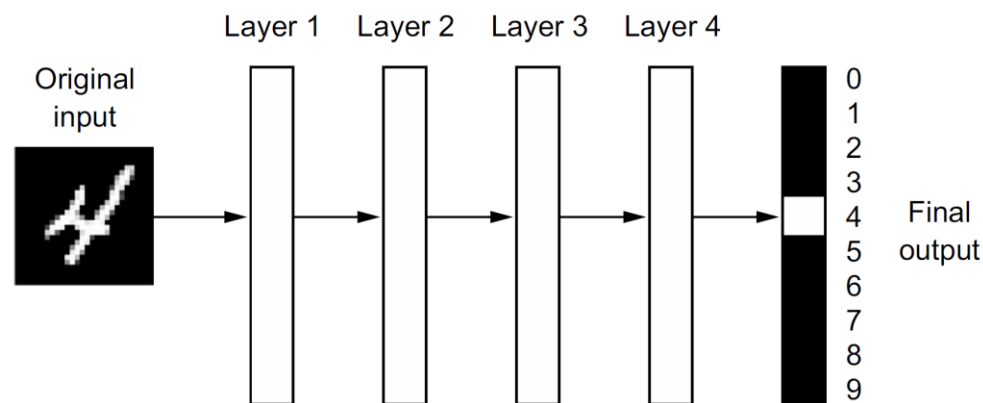
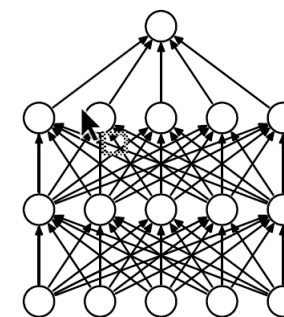
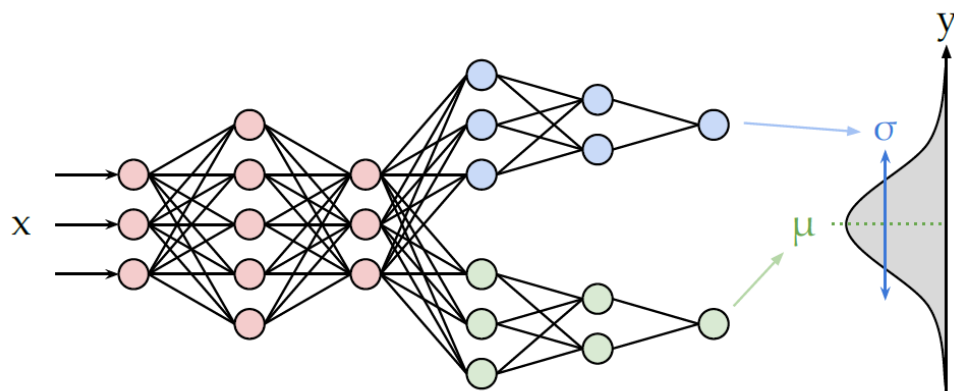
- Drzewa decyzyjne – dzielą dane na bloki
- Lasy losowe – dekorelacja drzew przez losowy wybór wierszy i kolumn (pakowanie w worki, ang. *bagging*) + uśrednianie predykcji
- Drzewa wzmacniane gradientowo – sekwencyjne uczenie na błędach poprzednika



Sieci neuronowe

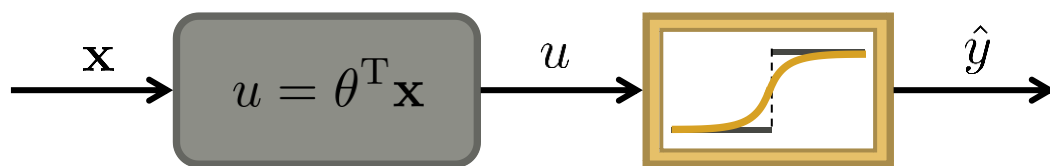
Model warstwowy: sekwencja (mnożenie macierzy $\rightarrow$ nieliniowa transformacja), DAG

„Neuron”: model liniowy + nieliniowa funkcja (tzw. *funkcja aktywacji*)



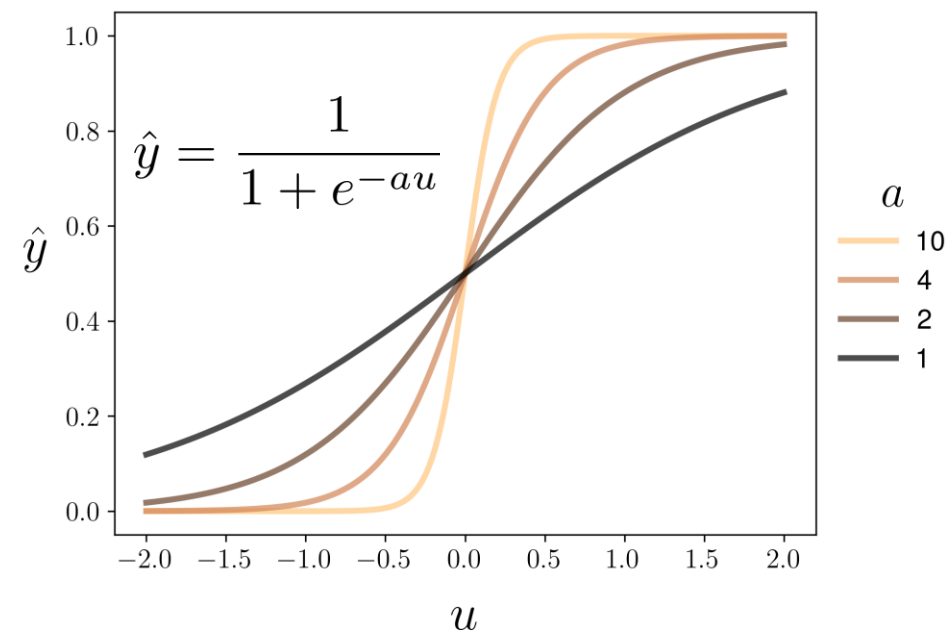
Regresja logistyczna jako klasyfikator

 $\left\{(\mathbf{x}_i, y_i)\right\}_{i=1}^N$ ale $y_i \in \{0, 1\}$ $y_i = \begin{cases} 1, & \text{jeśli } x_i \text{ reprezentuje klasę } C, \\ 0, & \text{w przeciwnym przypadku.} \end{cases}$

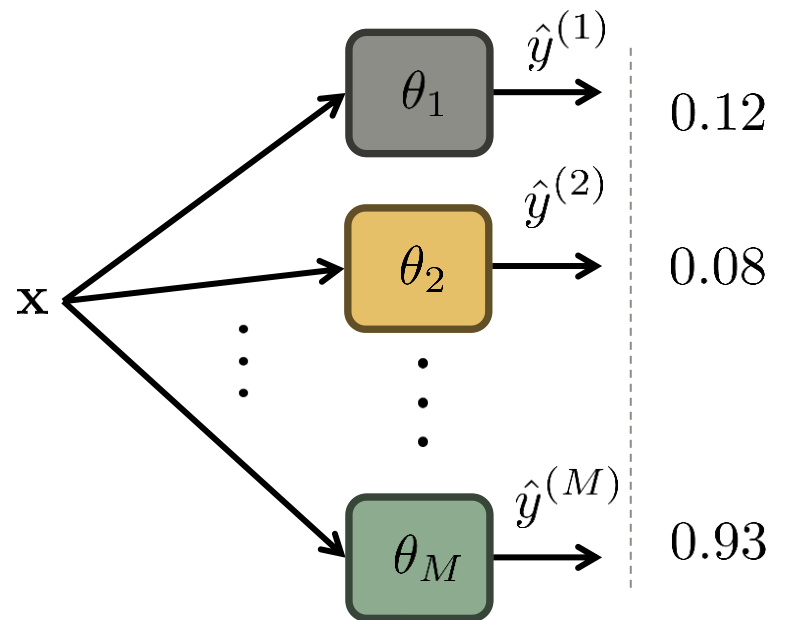


$$\hat{y}_i = p(y_i = 1 | \mathbf{x}_i, \theta)$$

$$\hat{y}_i = \sigma(u) = \frac{1}{1 + e^{-u}}$$



Funkcja softmax dla wielu klas



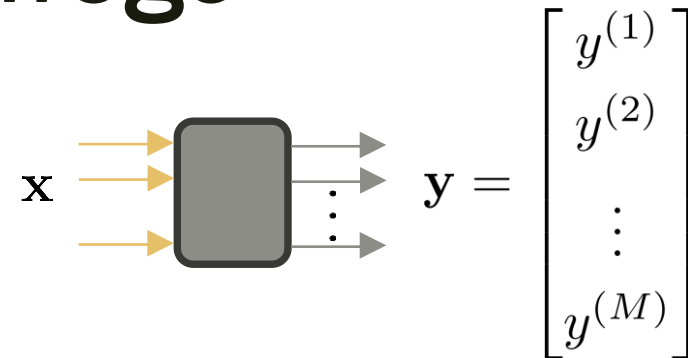
$$\hat{y}^{(m)} = \frac{1}{1 + e^{-\theta_m^T \mathbf{x}}} \quad (m = 1, 2, \dots, M)$$

vs

Softmax

$$\hat{y}^{(m)} = \frac{e^{\theta_m^T \mathbf{x}}}{\sum_{l=1}^M e^{\theta_l^T \mathbf{x}}} \quad (m = 1, 2, \dots, M)$$

Kryterium jakości dla klasyfikatora wieloklasowego



Kodowanie gorąco-jedynkowe klas (ang. *one-hot encoding*)

Jeżeli i -ta obserwacja $\mathbf{x}_i$ jest z klasy C_k to tylko $y_i^{(k)} = 1$ a pozostałe są zerowe.

Funkcja wiarygodności:

$$L(\theta) = \prod_{i=1}^N \prod_{m=1}^M p(y_i^{(m)} = 1 | \mathbf{x}_i, \theta)^{y_i^{(m)}} \Rightarrow -\log L(\theta) = -\sum_{i=1}^N \sum_{m=1}^M y_i^{(m)} \log p(y_i^{(m)} = 1 | \mathbf{x}_i, \theta)$$

kategorialna entropia krzyżowa
(ang. *categorical cross-entropy*)

Zał.: klasy wzajemnie się wykluczają

Kryterium jakości dla klasyfikatora **binarnego**

Jeżeli i -ta obserwacja $\mathbf{x}_i$ jest z klasy C to $y_i = 1$ w przeciwnym przypadku $y_i = 0$.

Funkcja wiarygodności:

$$L(\theta) = \prod_{i=1}^N p(y_i = 1 | \mathbf{x}_i, \theta)^{y_i} [1 - p(y_i = 1 | \mathbf{x}_i, \theta)]^{1-y_i}$$

$\Downarrow$

$$-\log L(\theta) = -\sum_{i=1}^N \left\{ y_i \log p(y_i = 1 | \mathbf{x}_i, \theta) + (1 - y_i) \log [1 - p(y_i = 1 | \mathbf{x}_i, \theta)] \right\}$$

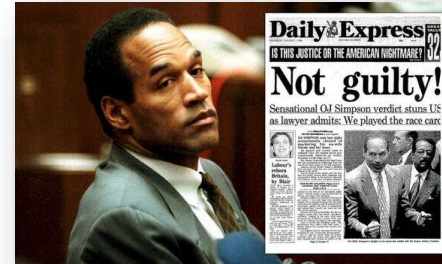
binarna entropia krzyżowa (ang. *binary cross-entropy*)

Błąd I rodzaju



- Hipoteza zerowa jest prawdziwa, ale ją odrzuca
- Wykrywam efekt, gdy go nie ma
- **Fałszywy alarm** (ang. *False Positive*)
- $P(\text{błąd I rodz.})$ to **poziom istotności**
 - Mylnie stwierdzona choroba
 - Dobry lek nie dopuszczony do obrotu
 - Ukarano niewinną osobę

Błąd II rodzaju



- Hipoteza zerowa jest fałszywa, ale ją akceptuję
- Nie zauważam efektu
- **Przeoczenie** (ang. *False Negative*)
- $1 - P(\text{błąd II rodz.})$ to **moc testu**
 - Nie rozpoznana choroba
 - Zły lek dopuszczony do obrotu
 - Uniewinniono sprawcę

88%

3000 kobiet

re3data.org

4 ma
nowotwór

2996 nie ma
nowotworu

88%

3000 kobiet

re3data.org

4 ma
nowotwór

2996 nie ma
nowotworu

 **2636**
wynik ujemny

 **360**
wynik dodatni

88%

3000 kobiet

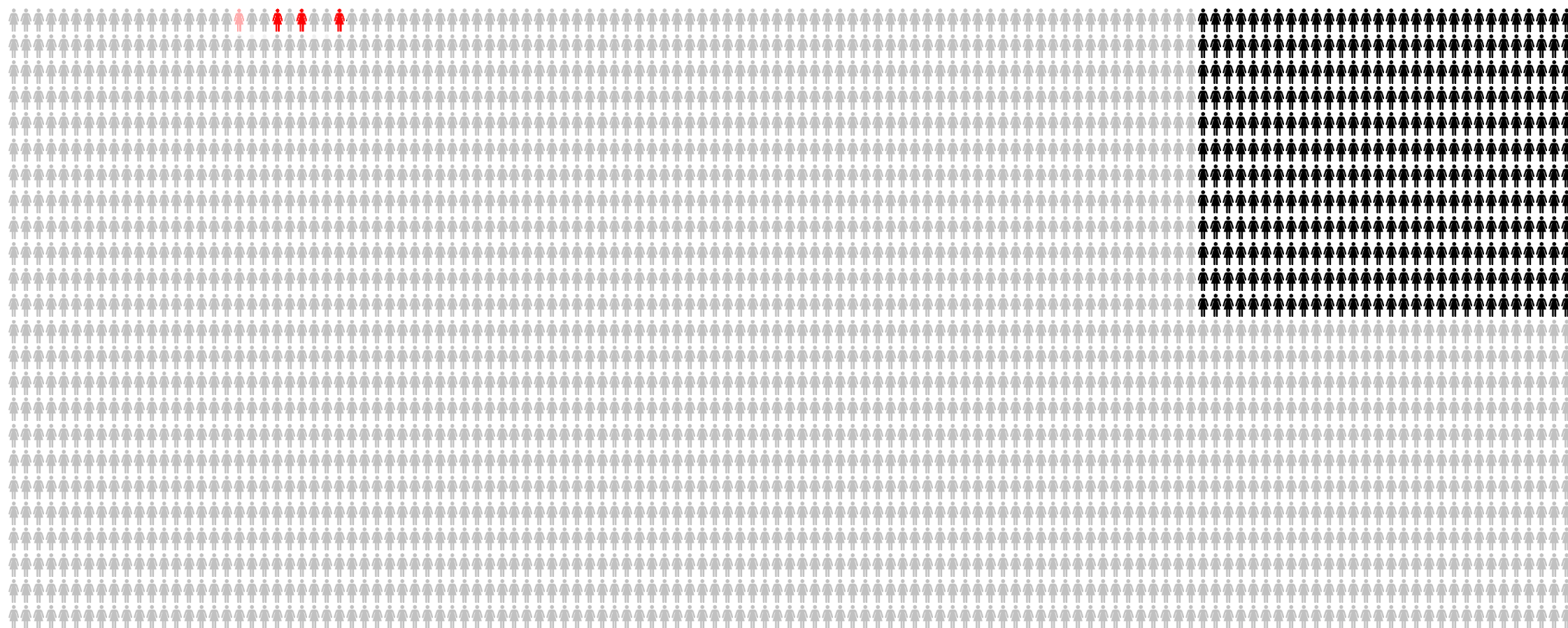
re3data.org



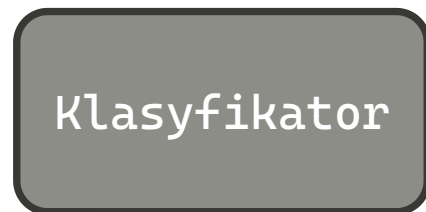
88%

3000 kobiet

re3data.org



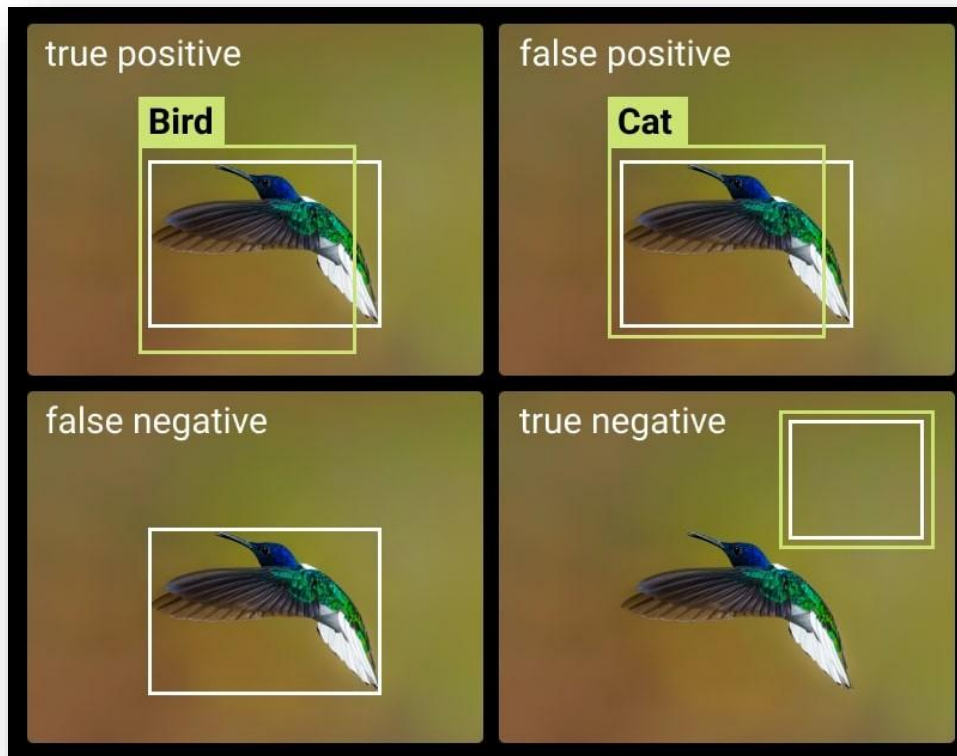
System klasyfikacji



+



Ocena jakości klasyfikacji



<https://labelyourdata.com/articles/object-detection-metrics>

		Stan faktyczny	
		Pozytywna	Negatywna
Predykcja	Pozytywna	✓ TP	✗ FP <i>falszywy alarm</i>
	Negatywna	✗ FN <i>przeoczenie</i>	✓ TN

Macierz pomyłek
(ang. *confusion matrix*)

Ocena jakości klasyfikacji

$$\text{Precyzja} = \frac{TP}{FP + TP}$$

Jak często przewidywania **P** są trafne?
Czy poprawnie wykrywa?

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

(ang. recall)

Jak często **P** są trafnie przewidziane?

$$F_1 = \frac{2}{\frac{1}{\text{Precyzja}} + \frac{1}{\text{Pełność}}}$$

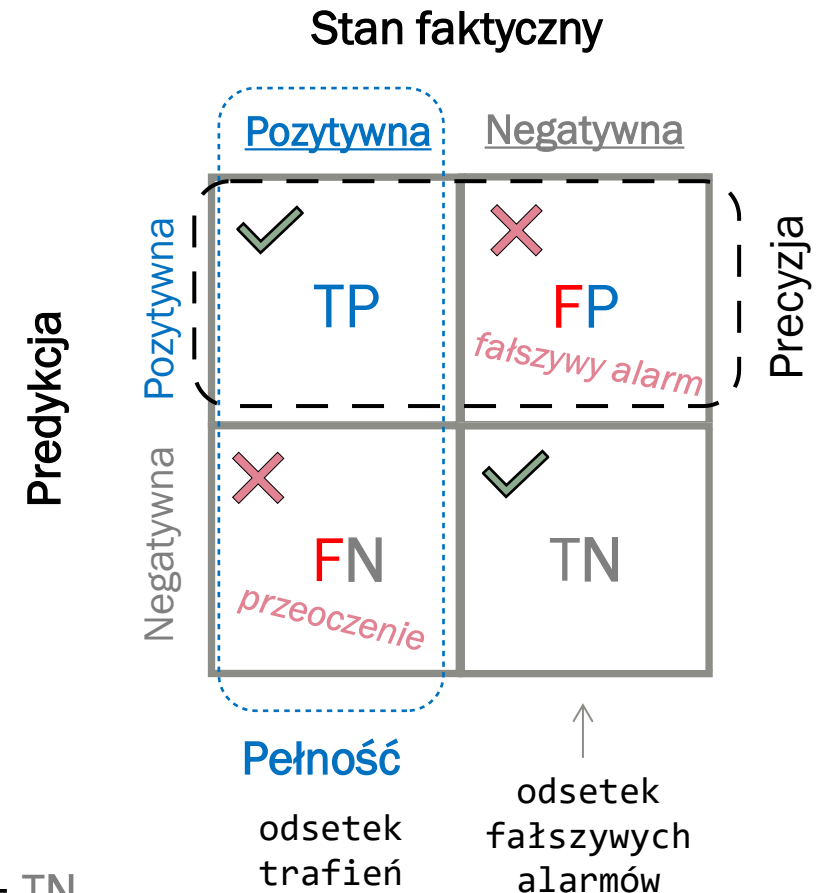
?wzór?

Czy poprawnie ignoruje?

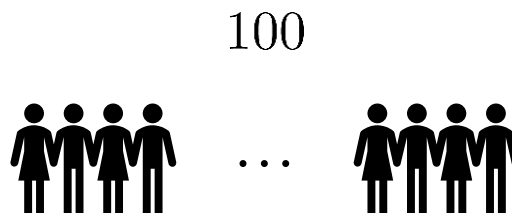
$$\text{Dokładność} = \frac{TP + TN}{TP + TN + FP + FN}$$

(ang. accuracy)

Odsetek prawidłowych decyzji



Przykład



- Czy klient zrezygnuje z subskrypcji?

5 zrezygnowało, model wykrył 4 z nich

20 innych model wskazał, ale jednak nie zrezygnowali

$$\text{Precyzja} = ? \quad \frac{4}{4 + 20}$$

$$\text{Pełność} = ? \quad \frac{4}{4 + 1}$$

$$\text{Dokładność} = ? \quad \frac{4 + 75}{4 + 75 + 20 + 1}$$

		Stan faktyczny	
		Pozytywna	Negatywna
Predykcja	Pozytywna	✓ TP=4	✗ FP=20
	Negatywna	✗ FN=1	✓ TN=75

Pełność: odsetek trafień

odsetek fałszywych alarmów

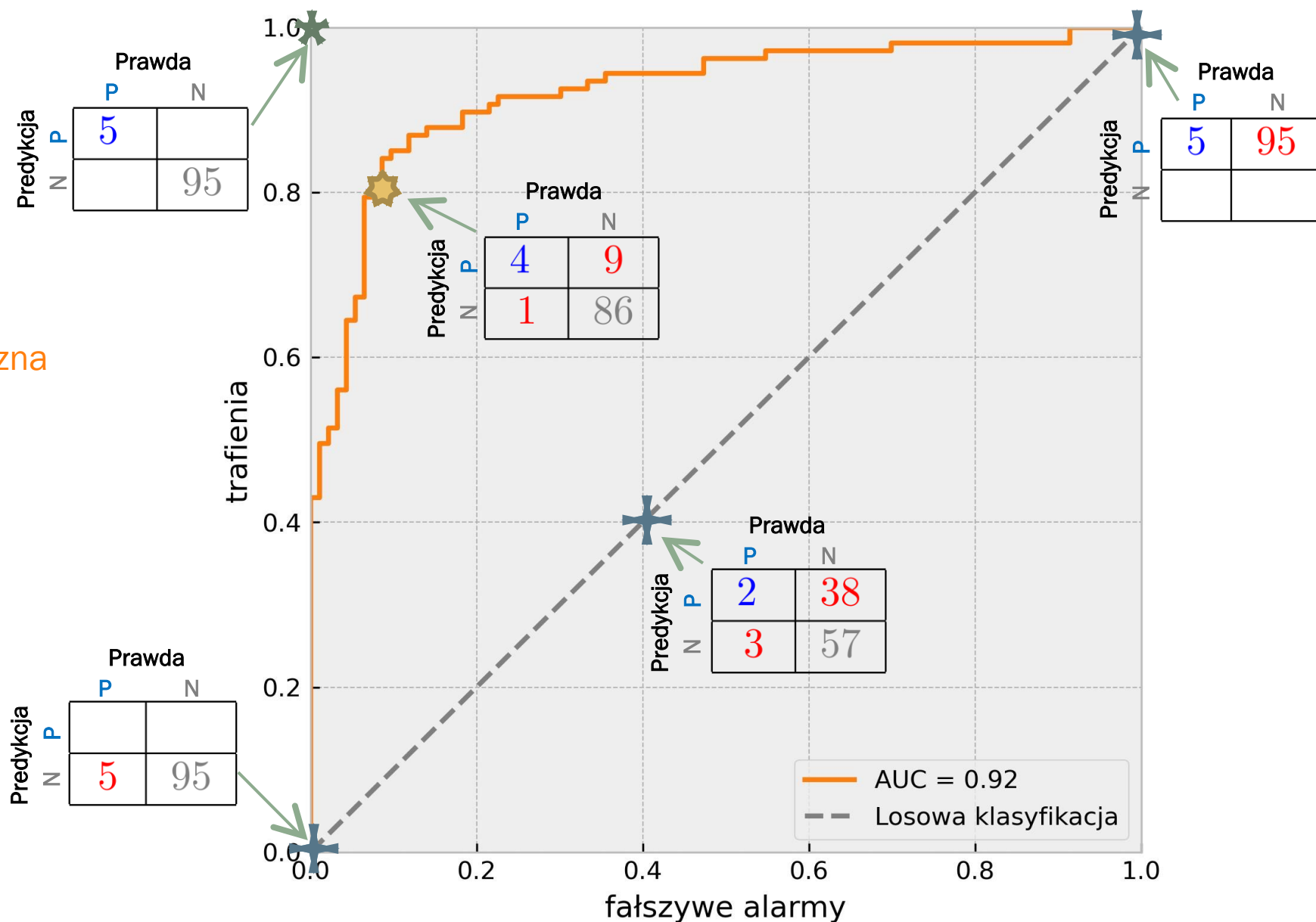
ROC

*Receiver
Operating
Characteristic*

krzywa
operacyjno-charakterystyczna

AUC

*Area
Under the
Curve*





THE END