

## Entropia

### Zadanie 1.

Uzasadnić następujące własności ( $X \perp Y$  oznacza, że atrybuty są niezależne):

$$I_H(X,Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

$$X \perp Y \Rightarrow H(X,Y) = H(X) + H(Y)$$

$$X \perp Y \Rightarrow I_H(X,Y) = 0$$

$$0 \leq H(X) \leq \log_2(n)$$

$$0 \leq I(X,Y) \leq H(Y)$$

### Zadanie 2.

Pewna zmienna losowa  $X$  może przyjąć wartości  $\{0,1,2\}$ . Rzucamy symetryczną monetą. Jeśli wypadnie orzeł, to zmiennej przypisujemy wartość 0, jeśli nie, ponownie rzucamy monetą. Gdy ponownie wypadnie orzeł przypisujemy zmiennej wartość 1, w przeciwnym przypadku 2. Wyznaczyć entropię zmiennej  $X$ .

### Zadanie 3.

Tabela przedstawia prawdopodobieństwa łączne dwóch zmiennych  $X$  oraz  $Y$ . Wyznaczyć  $H(X)$ ,  $H(Y)$ ,  $H(X|Y)$ ,  $H(Y|X)$ ,  $I(X,Y)$ .

$p(x,y)$	$x=1$	$x=2$	$x=3$	$x=4$
$y=1$	$1/8$	$1/16$	$1/32$	$1/32$
$y=2$	$1/16$	$1/8$	$1/32$	$1/32$
$y=3$	$1/16$	$1/16$	$1/16$	$1/16$
$y=4$	$1/4$	$0$	$0$	$0$

### Zadanie 4.

Na podstawie danych z tabeli wyznaczyć ile informacji na temat udomowienia dostarcza informacja o upierzeniu zwierzęcia.

pióra\domowe	false	true	łącznie
false	71	10	81
true	17	3	20
łącznie	88	13	

## Entropia

Niech  $X$  i  $Y$  są zmiennymi losowymi, przyjmującymi odpowiednie wartości  $\{x_1, x_2, \dots, x_n\}$ ,  $\{y_1, y_2, \dots, y_m\}$  z prawdopodobieństwami  $\{p_1(x), p_2(x), \dots, p_n(x)\}$  i  $\{p_1(y), p_2(y), \dots, p_m(y)\}$ .

Prawdopodobieństwa łączne  $p(x, y)$ .

**Entropia** – miara niepewności występowania losowego zdarzenia:

$$H = -\sum p \log p = \sum p \log \frac{1}{p}.$$

W przypadku, gdy  $p=0$  dla pewnego zdarzenia losowego wynik logarytmowania  $\log 0$  jest nieokreślony, ale nie stanowi to ograniczenia, ponieważ wartość składnika  $0 \log 0$  jest przyjmowana jako 0, co jest zgodne z granicą  $\lim_{p \rightarrow 0^+} p \log p = 0$ .

## Entropia warunkowa

Entropia warunkowa zdarzenia  $Y$  pod warunkiem zajścia zdarzenia  $X$ :

$$H(Y | X) = \sum_{i=1}^n P(x_i) H(Y | x_i)$$

Entropia warunkowa zdarzenia  $Y$  pod warunkiem, że zdarzenie  $X$  przyjęło wartość  $x_i$ :

$$H(Y | x_i) = -\sum_{j=1}^m P(y_j | x_i) \log P(y_j | x_i)$$

$$H(XY) = H(X) + H(Y | X) = H(Y) + H(X | Y)$$

## Informacja wzajemna

**Informacja wzajemna** – miara zależności pomiędzy dwiema zmiennymi losowymi.

$$I_H(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)}$$

$$I_H(X; Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y)$$

### Zadanie 1

Uzasadnić następujące własności ( $X \perp Y$  oznacza, że zmienne są niezależne):

- a.  $I_H(X, Y) = H(X) + H(Y) - H(X, Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$
- b.  $X \perp Y \Rightarrow H(X, Y) = H(X) + H(Y)$
- c.  $X \perp Y \Rightarrow I_H(X, Y) = 0$
- d.  $X \perp Y \Rightarrow I_G(X, Y) = 0$
- e.  $0 \leq H(X) \leq \log_2(n)$
- f.  $0 \leq I(X, Y) \leq H(Y)$

## Zadanie 2

Pewna zmienna losowa  $X$  może przyjąć wartości  $\{0, 1, 2\}$ . Rzucamy symetryczną monetą. Jeśli wypadnie orzeł, to zmiennej przypisujemy wartość 0, jeśli nie, ponownie rzucamy monetą. Gdy ponownie wypadnie orzeł przypisujemy zmiennej wartość 1, w przeciwnym przypadku 2. Wyznaczyć entropię zmiennej  $X$ .

$$H(X) = -\sum p \log p = \sum p \log \frac{1}{p} = \frac{1}{2} \log 2 + \frac{1}{4} \log 4 + \frac{1}{4} \log 4 = \\ = \frac{1}{2} (\log 2 + \log 4) = \frac{1}{2} (1+2) = \frac{3}{2}$$

$$\Pr(X=0) = \frac{1}{2}$$

$$\Pr(X=1) = \frac{1}{4}$$

$$\Pr(X=2) = \frac{1}{4}$$



## Zadanie 3

Tabela przedstawia prawdopodobieństwa łączne dwóch zmiennych  $X$  oraz  $Y$ . Wyznaczyć  $H(X)$ ,  $H(Y)$ ,  $H(X|Y)$ ,  $H(Y|X)$ ,  $H(X,Y)$ ,  $I(X,Y)$ .

$p(x,y)$	$x=1$	$x=2$	$x=3$	$x=4$	$\Pr(Y)$
$y=1$	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
$y=2$	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
$y=3$	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
$y=4$	$1/4$	$0$	$0$	$0$	$1/4$
$\Pr(x)$	$1/2$	$1/4$	$1/8$	$1/8$	

$$H(X) = 1 \frac{3}{4}$$

$$H(Y) = 2$$

$$H(X|Y) = \frac{11}{8}$$

$$H(Y|X) = \frac{13}{8}$$

$$H(X,Y) = \frac{27}{8}$$

$$I(X,Y) = \frac{3}{8}$$

$$H(X) = \frac{1}{2} \log 2 + \frac{1}{4} \log 4 + \frac{1}{8} \log 8 + \frac{1}{8} \log 8 = \frac{1}{2} + \frac{1}{2} + \frac{6}{8} = 1\frac{3}{4}$$

$$H(Y) = 4 \cdot \left( \frac{1}{4} \log 4 \right) = \log 4 = 2$$

Entropie warunkowa Y pod warunkiem X

$$H(Y|X) = \sum p(x_i) H(Y|x_i)$$

$$H(Y|x_i) = \sum p(y_j|x_i) \log \frac{1}{p(y_j|x_i)}$$

warunkowy Tarcie

$$p(y_j|x_i) \neq p(y_i), x$$

$$p(y_j|x_i) = \frac{p(y_i, x_i)}{x_i}$$

kolumna 1 kolumna 2 kolumna 3, 4

$$\cancel{\frac{1}{8}/\frac{1}{2} = \frac{1}{4}} \quad \cancel{\frac{1}{16}/\frac{1}{4} = \frac{1}{4}} \quad \cancel{\frac{1}{32}/\frac{1}{8} = \frac{1}{4}}$$

$$\cancel{\frac{1}{16}/\frac{1}{2} = \frac{1}{8}} \quad \cancel{\frac{1}{8}/\frac{1}{4} = \frac{1}{2}} \quad \cancel{\frac{1}{32}/\frac{1}{8} = \frac{1}{4}}$$

$$\cancel{\frac{1}{16}/\frac{1}{2} = \frac{1}{8}} \quad \cancel{\frac{1}{16}/\frac{1}{4} = \frac{1}{4}} \quad \cancel{\frac{1}{16}/\frac{1}{8} = \frac{1}{2}}$$

$$\cancel{\frac{1}{16}/\frac{1}{2} = \frac{1}{2}} \quad 0 \quad 0 \quad 0$$

$$H(Y|x=1) = \frac{1}{4} \log 4 + \frac{1}{8} \log 8 + \frac{1}{8} \log 8 + \frac{1}{2} \log 2 = \\ = \frac{1}{2} + \frac{3}{8} + \frac{3}{8} + \frac{1}{2} = \frac{14}{8}$$

$$H(Y|x=2) = \frac{1}{4} \log 4 + \frac{1}{2} \log 2 + \frac{1}{4} \log 4 = \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = \frac{3}{2}$$

$$H(Y|x=3) = H(Y|x=4) = \frac{1}{4} \log 4 + \frac{1}{4} \log 4 + \frac{1}{2} \log 2 = \frac{3}{2}$$

$$H(Y|x) = \frac{1}{2} \cdot \frac{14}{8} + \frac{1}{4} \cdot \frac{3}{2} + \frac{1}{8} \cdot \frac{3}{2} + \frac{1}{8} \cdot \frac{3}{2} =$$

$H(X|Y)$  - korzystamy z własności z zad 1(a)

$$I_H(X,Y) = H(X) + H(Y) - H(X,Y) = H(Y) - H(Y|X) = H(X) - H(X|Y)$$

$$\bullet I_H(X,Y) = H(Y) - H(Y|X) = 2 - \frac{13}{8} = \frac{3}{8}$$

$$\bullet \frac{3}{8} = \frac{7}{4} + 2 - H(X,Y)$$

$$\frac{3}{8} - \frac{7}{4} - 2 = -H(X,Y)$$

$$\bullet H(X,Y) = \frac{27}{8}$$

$$\bullet \frac{3}{8} = \frac{7}{4} - H(X|Y)$$

$$H(X|Y) = \frac{14}{8} - \frac{3}{8} = \frac{11}{8}$$

#### Zadanie 4

Na podstawie danych z tabeli wyznaczyć ile informacji na temat udomowienia dostarcza informacja o upierzeniu zwierzęcia.

$$P(Y)$$

pióra\domowe	false $X_1$	true $X_2$	łącznie
$Y_1$ false	71/101	10/101	81/101
$Y_2$ true	17/101	3/101	20/101
łącznie $P(X)$	88/101	13/101	

Y - udomowienie

X - upierzenie

101

$$I(Y, X) = H(Y) + H(X) - H(Y|X) = H(Y) - H(Y|X) = H(X) + H(X|Y)$$

$$H(X) = \frac{88}{101} \log \frac{101}{88} + \frac{13}{101} \log \frac{101}{13} = 0,55888$$

0.87129      1.14473      0.12871      2.76923

$$H(Y) = \frac{81}{101} \log \frac{101}{81} + \frac{20}{101} \log \frac{101}{20} = 0.41785$$

0.80199      1.24691      0.19801      5.05

$$H(X|Y) = \frac{71}{101} \log \frac{101}{71} + \frac{17}{101} \log \frac{101}{17} + \frac{10}{101} \log \frac{101}{10} + \frac{3}{101} \log \frac{101}{3}$$

1.27115

# Lecture 1 - 1

$$\begin{aligned} I &= \sum \sum p(x,y) \log \frac{p(x,y)}{p(x)p(y)} = \\ &= \sum_x \sum_y p(x,y) \left( \log p(x,y) - \log p(x) - \log p(y) \right) = \\ &= \sum_x \sum_y \left[ p(x,y) \left( \log p(x,y) - p(x,y) \log p(x) - p(x,y) \log p(y) \right) \right] = \\ &= \sum_x \sum_y p(x,y) \log(p(x,y)) - \underbrace{\sum_x \sum_y p(x,y) \log p(x)}_{+ H(X)} - \underbrace{\sum_x \sum_y p(x,y) \log p(y)}_{+ H(Y)} = \\ &= -H(X,Y) - \underbrace{H(X)}_{+ H(X)} - \underbrace{H(Y)}_{+ H(Y)} = H(X,Y) \end{aligned}$$

## Przetwarzanie wstępne i wizualizacja

**Zadanie.** Dla zbioru danych challengerRing (zob tab.):

1. Narysować histogramy dla atrybutów (zmiennych)
2. Zdyskretyzować zmienną temperatura pod warunkiem uszkodzeń .
3. Napisać w postaci tabeli kontyngencji rozkład liczności tych dwóch zmiennych.
4. Znaleźć rozkład częstości i rozkłady brzegowe tych dwóch atrybutów.  
Porównać empiryczny rozkład częstości z rozkładem produktowym
5. Postawić test  $\chi^2$  o zależności zmiennych (znaleźć wartość statystyki).

Temporal order of flight	Number of O-rings at risk on a given flight	Number experiencing thermal distress	Launch temperature (degrees F)	Leak-check pressure (psi)
1	6	0	66	50
2	6	1	70	50
3	6	0	69	50
4	6	0	68	50
5	6	0	67	50
6	6	0	72	50
7	6	0	73	100
8	6	0	70	100
9	6	1	57	200
10	6	1	63	200
11	6	1	70	200
12	6	0	78	200
13	6	0	67	200
14	6	2	53	200
15	6	0	67	200
16	6	0	75	200
17	6	0	70	200
18	6	0	81	200
19	6	0	76	200
20	6	0	79	200
21	6	2	75	200
22	6	0	76	200
23	6	1	58	200

## Algorytmy geometryczne, grupowanie danych

**Zadanie 1.**

Dane są punkty  $x_1, x_2, \dots, x_n$ . Znaleźć taki punkt  $c$ , dla którego suma kwadratów odległości punktów  $x_i$  od punktu  $c$  jest najmniejsza.

**Zadanie 2.**

Dane są punkty  $x_1, x_2, \dots, x_n$ . Znaleźć taki punkt  $c$ , dla którego suma odległości punktów  $x_i$  od punktu  $c$  jest najmniejsza.

A2 - 2 Zdjęcie typowym dla zmiennej temperatury

<del>T</del>	U	0	1	$P(T)$
$T < 60$	0	0	$\frac{3}{23}$	$\frac{3}{23}$
$T > 60$	1	$\frac{16}{23}$	$\frac{4}{23}$	$\frac{20}{23}$
$P(U)$		$\frac{16}{23}$	$\frac{4}{23}$	

<del>T</del>	U	0	1	$P(T)$
$T < 70$	0	$\frac{6}{23}$	$\frac{4}{23}$	$\frac{10}{23}$
$T > 70$	1	$\frac{10}{23}$	$\frac{3}{23}$	$\frac{13}{23}$
$P(U)$		$\frac{16}{23}$	$\frac{4}{23}$	

<del>T</del>	U	0	1	$P(T)$
$T < 65$	0	0	$\frac{4}{23}$	$\frac{4}{23}$
$T > 65$	1	$\frac{16}{23}$	$\frac{3}{23}$	$\frac{19}{23}$
$P(U)$		$\frac{16}{23}$	$\frac{4}{23}$	

<del>T</del>	U	0	1	$P(T)$
$T < 75$	0	$\frac{10}{23}$	$\frac{6}{23}$	$\frac{16}{23}$
$T > 75$	1	$\frac{6}{23}$	$\frac{1}{23}$	$\frac{16}{23}$
$P(U)$		$\frac{16}{23}$	$\frac{4}{23}$	

$$H = \sum p \log \frac{p}{\bar{p}}$$

$$H(U) = \frac{16}{23} \log \frac{23}{16} + \frac{4}{23} \log \frac{23}{4} = 0,8865$$

$$H(\bar{T}_{60}) = \frac{3}{23} \log \frac{23}{3} + \frac{20}{23} \log \frac{23}{20} = 0,55863$$

$$H(\bar{T}_{65}) = \frac{4}{23} \log \frac{23}{4} + \frac{19}{23} \log \frac{23}{19} = 0,66658$$

$$H(\bar{T}_{70}) = \frac{10}{23} \log \frac{23}{10} + \frac{13}{23} \log \frac{23}{13} = 0,98769$$

$$H(\bar{T}_{75}) = \frac{6}{23} \log \frac{23}{6} + \frac{17}{23} \log \frac{23}{17} = 0,88654$$

$$H(T_{60}, U) = 0 + \frac{3}{23} \log \frac{23}{3} + \frac{16}{23} \log \frac{23}{16} + \frac{4}{23} \log \frac{23}{4} = 1,18639$$

$$H(T_{65}, U) = 0 + \frac{4}{23} \log \frac{23}{4} + \frac{16}{23} \log \frac{23}{16} + \frac{3}{23} \log \frac{23}{3} = 1,18639$$

$$H(T_{70}, U) = \frac{6}{23} \log \frac{23}{6} + \frac{4}{23} \log \frac{23}{4} + \frac{10}{23} \log \frac{23}{10} + \frac{3}{23} \log \frac{23}{3} = 1,85035$$

$$H(T_{75}, U) = \frac{10}{23} \log \frac{23}{10} + \frac{6}{23} \log \frac{23}{6} + \frac{6}{23} \log \frac{23}{6} + \frac{1}{23} \log 23 = 1,73057$$

$$I(T, U) = H(T) + H(U) - H(T, U)$$

$$I(T_{60}, U) = 0,55863 + 0,8865 - 1,18639 = 0,25849$$

$$I(T_{65}, U) = 0,66658 + 0,8865 - 1,18639 = 0,36669$$

$$I(T_{70}, U) = 0,98768 + 0,8865 - 1,85035 = 0,02384$$

$$I(T_{75}, U) = 0,88654 + 0,8865 - 1,73057 = 0,04244$$

Podejścia (dystretyczne) dokonujemy dla  
największej wartości - tu 65

Rozkład empiryczny a produktywny (przykład)

W rozkładzie produktywnym  $p(U, T) = p(U)p(T)$

$T \setminus U$	0	1	$p(T)$
$T < 65$	0	4	$\frac{4}{23}$
$T \geq 65$	16	3	$\frac{19}{23}$
$p(U)$	$\frac{16}{23}$	$\frac{4}{23}$	

$T \setminus U$	0	1	$p(T)$
$T < 65$	0	0	$\frac{4 \cdot 16}{23 \cdot 23}$
$T \geq 65$	16	$\frac{16}{23}$	$\frac{18 \cdot 16}{23 \cdot 23}$
$p(U)$	$\frac{16}{23}$	$\frac{3}{23}$	$\frac{19 \cdot 4}{23 \cdot 23}$
			$\frac{49}{23}$

m - empiryczne

## Hipoteza

$H_0$  - zmienne są niezależne

$H_1$  - zmienne są zależne

$$\alpha = 0,05$$

$$\chi^2 = \sum \frac{(n_{empiryczne} - n_{teoretyczne})^2}{n_{teoretyczne}}$$

$$\chi^2 = \frac{\left(0 - \frac{16 \cdot 4}{23}\right)^2}{\frac{16 \cdot 4}{23}} + \frac{\left(4 - \frac{4 \cdot 7}{23}\right)^2}{\frac{4 \cdot 7}{23}} + \frac{\left(16 - \frac{16 \cdot 19}{23}\right)^2}{\frac{16 \cdot 19}{23}} + \frac{\left(3 - \frac{4 \cdot 19}{23}\right)^2}{\frac{4 \cdot 19}{23}} =$$

$$= 11,07$$

$\chi^2_{\text{obszar krytyczny}}$

df? - liczba stopni...

$$df? = (N-1)(k-1) = 1$$

$$\chi_n(4,65,1) = 3,84$$

Z tabeli (3,8415) - przafiliszymy w przedziałie wyciągając  $\chi^2$

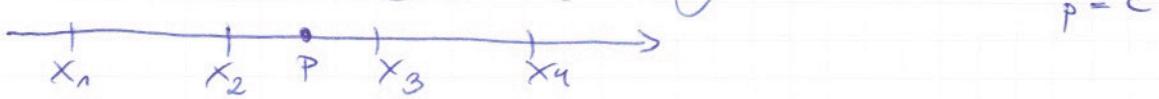
obszar krytyczny

$(3,84; +\infty)$  - odrzucajemy  $H_0$ , nie mały odstęp, zatem...  $H_1$

# Algorytmy geometryczne...

## Zadanie 1.

Dane są punkty  $x_1, x_2, \dots, x_n$ . Znaleźć punkt  $c$ , dla którego suma odległości punktów  $x_i$  od punktu  $c$  jest najmniejsza.



$$F = \sum_{i=1}^n (x_i - c)^2$$

$$\frac{dF}{dc} = \sum 2(x_i - c)(-1) = \sum -2(x_i - c) = 0$$

$$\sum (x_i - c) = 0 \Leftrightarrow \sum_{i=1}^n (x_i - c) = x_1 - nc + x_2 - nc + x_3 - nc \dots + x_n - nc$$

$$\sum x_i \cdot nc = 0$$

$$c = \frac{\sum x_i}{n}$$

## Wnioskowanie

### **Zadanie 1.**

Rzucamy  $n$  razy niesymetryczną monetą (prawdopodobieństwo orła-  $p$ ),  $k$  razy wypadł orzeł  $n - k$  - reszka. Znaleźć estymator parametru  $p$ , stosując metodę największej wiarygodności.

### **Zadanie 2.**

Wykorzystując metodę największej wiarygodności, ocenić parametr  $p$  dwumianowego rozkładu Bernoullego  $P_n(k) = C_n^k p^k (1-p)^{n-k}$ , jeśli przy  $n_1$  niezależnych doświadczeniach zdarzenie  $A$  występowało  $m_1$  razy, a przy  $n_2$  niezależnych doświadczeniach zdarzenie  $A$  występowało  $m_2$  razy.

### **Zadanie 3.**

Wykorzystując metodę największej wiarygodności, ocenić parametr  $\lambda$  rozkładu Poissona  $P_m(X = x_i) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$  zmiennej losowej  $X$  na podstawie próby  $x_1, x_2, \dots, x_n$ , gdzie  $x_i$  - liczba wystąpień niezależnych zdarzeń  $A$  w  $i$ -tym doświadczeniu.

### **Zadanie 4.**

Wykorzystując metodę największej wiarygodności, ocenić parametr  $w$  na podstawie próby: 1,4 1,5 3,2 1,4 2,5 3,4 3,1 2,4 3,8 2,6, jeśli funkcja gęstości ma postać:

$$f(x) = \frac{2x^3}{\sqrt{2\pi}} e^{-\frac{(x^4-w)^2}{2}}$$

### **Zadanie 5. (samodzielnie 😊)**

Wykorzystując metodę największej wiarygodności, ocenić parametr  $p$  rozkładu geometrycznego  $P(X = x_i) = (1-p)^{x_i-1} p$  zmiennej losowej  $X$  na podstawie próby  $x_1, x_2, \dots, x_n$ , gdzie  $x_i$  - liczba doświadczeń wykonanych do występowania zdarzenia,  $p$  - prawdopodobieństwo występowania zdarzenia w jednym doświadczeniu.

### **Zadanie 6.**

Wykorzystując metodę największej wiarygodności, ocenić parametr  $\lambda$  na podstawie próby:

$X$	1-3	3-5	5-7	7-9	9-11	11-13	13-15	15-17	17-19
$n$	5	6	7	15	22	27	30	34	35

pod warunkiem, że ciągła zmienna losowa ma gęstość prawdopodobieństwa:

$$f(x) = \begin{cases} \lambda e^{\lambda(x-20)}, & x \leq 20 \\ 0, & x > 20 \end{cases}$$

# WNIOSKI NIANIE

## FUNKCJA WIARYGODNOŚCI

$$L(x, w) = \prod_{i=1}^n p(x_i, w) \quad / \text{iloczyn prawdopodobieństwa}$$

Funkcja  $L(x, w)$  osiąga maksimum w tych samych punktach, co jej logarytm i w praktyce często wykorzystuje się LOGARYTMICZNĄ FUNKcję WIARYGODNOŚCI

$$l(x, w) = \ln(L(x, w)) = \ln\left(\prod_{i=1}^n p(x_i, w)\right) = \sum_{i=1}^n \ln(p(x_i, w))$$

$$\frac{\partial l}{\partial w_k} = 0$$

### Zadanie 1

$L$  - iloczyn wszystkich prawdopodobieństw

$$L = \prod_{i=1}^n p = \cancel{p^k \cdot p^{n-k}} = p^k \cdot (1-p)^{n-k}$$

$$l = \ln L = \ln p^k \cdot (1-p)^{n-k} = \ln p^k + \ln(1-p)^{n-k} - \\ - k \ln p + (n-k) \ln(1-p)$$

$$\frac{\partial l}{\partial p} = \frac{k}{p} + \frac{n-k}{1-p} \cdot (-1) = 0$$

$$\frac{k}{p} = \frac{n-k}{1-p} \quad / \cdot p$$

$$k = \frac{(n-k)p}{1-p} \quad \Rightarrow p = \frac{k}{n}$$

### Zadanie 3

$$L = \prod_{i=1}^n p_i = \prod_{i=1}^n \frac{\lambda^{x_i} e^{-\lambda}}{x_i!} = e^{-\lambda n} \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!}$$

$$\lambda = \ln L = \ln \left( e^{-\lambda n} \prod_{i=1}^n \frac{\lambda^{x_i}}{x_i!} \right) = \ln e^{-\lambda n} + \sum_{i=1}^n \frac{\lambda^{x_i}}{x_i!} =$$

$$= -\lambda n + \sum_{i=1}^n \left[ x_i \ln \lambda - \ln(x_i!) \right] = -\lambda n + \ln \lambda \sum_{i=1}^n - \sum_{i=1}^n \ln(x_i!)$$

$$\frac{\partial \lambda}{\partial \lambda} = -n + \frac{1}{\lambda} \sum_{i=1}^n = 0$$

$$\sum_{i=1}^n x_i = n\lambda$$

$$\lambda = \frac{\sum_{i=1}^n x_i}{n}$$

### Zadanie 4

$$L = \prod_{i=1}^n \frac{2x_i^3}{\sqrt{2\pi}} e^{-\frac{(x_i^4 - w)^2}{2}} = \left( \frac{2}{\sqrt{2\pi}} \right)^n \prod_{i=1}^n x_i^3 e^{-\frac{(x_i^4 - w)^2}{2}}$$

$$\lambda = \ln \left( \frac{2}{\sqrt{2\pi}} \right)^n + \sum_{i=1}^n \ln(x_i^3) + \ln \left( e^{-\frac{(x_i^4 - w)^2}{2}} \right) \Rightarrow (\ln e)$$

$$= n \ln \left( \frac{2}{\sqrt{2\pi}} \right) + \sum_{i=1}^n 3 \ln(x_i) + \sum_{i=1}^n -\frac{(x_i^4 - w)^2}{2}$$

$$\frac{\partial \lambda}{\partial w} = -\frac{1}{2} \sum_{i=1}^n 4(x_i^4 - w) - (-1) = 0 \Rightarrow$$

$$\sum_{i=1}^n (x_i^4 - w)$$

$$w = \frac{1,4^4 + 1,5^4 + 3,2^4 + 1,4^4 + 2,5^4 + 3,4^4 + 3,1^4 + 2,1^4 + 3,8^4 + 2,6^4}{10} = 67,01$$

Ladungse S

$$\varphi(X=x_i) = (1-p)^{x_i-1}$$

$$L = \prod_{i=1}^n (1-p)^{x_i-1} \cdot p^n = p^n \prod_{i=1}^n (1-p)^{x_i-1}$$

$$l = \ln L = n \ln(p) + \sum_{i=1}^n (1-p)^{x_i-1} = \ln p + \sum_{i=1}^n \ln(1-p)^{x_i-1}$$
$$n \ln p + \sum_{i=1}^n (x_i - 1) \ln(1-p)$$

$$\frac{\partial l}{\partial p} = \frac{n}{p} - \frac{1}{1-p} \sum_{i=1}^n (x_i - 1) = 0$$

$$\frac{n}{p} = \frac{1}{1-p} \cdot \sum_{i=1}^n (x_i - 1) \quad | \cdot \frac{1-p}{n}$$

$$\frac{1-p}{p} = \frac{\sum_{i=1}^n (x_i - 1)}{n}$$

$$\frac{1-p}{p} = a$$

$$p \cdot a = 1-p$$

$$\frac{1-p}{p} = \frac{\sum_{i=1}^n (x_i - 1)}{n}$$

$$\frac{(\sum x_i) - n}{n} = \frac{\sum x_i - 1}{n}$$

$$\frac{1-p}{p} = \frac{\sum x_i - 1}{n} \cdot p$$

$$1-p = p \left( \frac{\sum x_i}{n} - 1 \right)$$

$$1 = p \frac{\sum x_i}{n} - p + p \quad \left| \frac{n}{\sum x_i} \right.$$

$$\boxed{p = \frac{n}{\sum x_i}}$$

### Zadanie 6

$$L = \prod_{i=1}^n \lambda e^{\lambda(x_i - 20)} = \lambda^n \prod_{i=1}^n e^{\lambda(x_i - 20)}$$

$$l = \ln L = \ln \left( \lambda^n \prod_{i=1}^n e^{\lambda(x_i - 20)} \right) = n \ln \lambda + \sum_{i=1}^n \ln(e^{\lambda(x_i - 20)})$$

$$= n \ln \lambda + \sum_{i=1}^n \ln e^{\lambda(x_i - 20)} = n \ln \lambda + \sum_{i=1}^n \lambda(x_i - 20) \cdot 1 =$$

$$= n \ln \lambda + \sum_{i=1}^n \lambda x_i - \sum_{i=1}^n 20\lambda$$

$$\frac{\partial l}{\partial \lambda} = \frac{n}{\lambda} + \sum x_i - \underbrace{\sum}_{20n} = 0$$

$$\frac{n}{\lambda} = - \sum_{i=1}^n x_i + 20n$$

$$\lambda = \frac{n}{\sum x_i - 20n}$$

- liczba z szeregu przedziałów (śr. przedziału)

~~$\lambda$~~

~~$x_i$~~

X	1-3	3-5	5-7	7-9	9-11	11-13	13-15	15-17	17-19
$x_i$	2	4	6	8	10	12	14	16	18
$n_i$	5	6	4	15	22	24	30	34	35

$$m = \sum n_i = 181$$

$$\text{Mak} \sum x_i = \sum_{i=1}^n (y_i \cdot n_i)$$

$$\lambda = \frac{181}{\frac{1}{2}(121+223)} = 0,141$$

## Miary jakości klasyfikacji

Możliwe wyniki klasyfikacji:

- *True-Positive (TP - prawdziwie pozytywna): przewidywanie pozytywne, faktycznie zaobserwowana klasa pozytywna;*
- *True-Negative (TN - prawdziwie negatywna): przewidywanie negatywne, faktycznie zaobserwowana klasa negatywna*
- *False-Positive (FP - fałszywie pozytywna): przewidywanie pozytywne, faktycznie zaobserwowana klasa negatywna*
- *False-Negative (FN - fałszywie negatywna): przewidywanie negatywne, faktycznie zaobserwowana klasa pozytywna*

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

$$\text{Błąd klasyfikacji: } \frac{FN + FP}{TN + TP + FN + FP} = 1 - \text{dokładność}$$

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

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

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

$$F_1: \frac{2TP}{2TP + FP + FN} = \frac{2 \cdot \text{czułość} \cdot \text{precyzja}}{\text{czułość} + \text{precyzja}}$$

$$\text{Zbalansowana dokładność: } \frac{1}{2} \text{czułość} + \frac{1}{2} \text{specyficzność}$$

### Zadanie 1.

Zinterpretuj TN, FN, TP, FP na podstawie przykładu klasyfikatora wykrywającego pewną chorobę (tj. 0 – osoba zdrowa, 1 – osoba chora).

### Zadanie 2.

Zinterpretuj miary: dokładność klasyfikacji, błąd klasyfikacji, czułość, specyficzność, precyzja na przykładzie klasyfikatora wykrywającego chorobę (tj. 0 – osoba zdrowa, 1 – osoba chora).

### Zadanie 3.

Dla danych z tabeli:

klasa	0	1	1	0	1	0	0	1	0	0
predykcja $\Pr(d=1 x)$	0,95	0,8	0,75	0,6	0,55	0,4	0,3	0,25	0,2	0,1

wykonać następujące działania:

1. Przyjmując próg 0,5 wyznaczyć macierz konfuzji, dokładność, błąd, czułość, specyficzność, precyzję, zbalansowaną dokładność oraz miarę  $F_1$ .
2. Dla kilku wartości progu wyznaczyć czułość oraz specyficzność, a następnie narysować krzywą ROC.
3. Na podstawie wyznaczonych punktów na krzywej ROC oszacować minimalną i maksymalną wartość miary AUC.

# MILKY JAKSU KLASYFIKACJI

Zad 3.

$P(d=1|x)$

Próbka  
(klasy)

→ Klasifikator  
(predyktory)

0	0	1	1	0	1	0	0	1	0	0
0,95	0,8	0,75	0,6	0,55	0,4	0,3	0,25	0,2	0,1	
0,5	1 FP	1 TP	1 TP	1 FP	1 TP	TN	TN	FN	TN	TN
0	1 FP	1 TP	1 TP	1 FP	1 TP	1 FP	1 FP	1 TP	FP	FP
0,25	1 FP	1 TP	1 TP	1 FP	1 TP	1 FP	1 FP	1 TP	TN	TN
0,75	1 FP	1 TP	1 TP	TN	FN	TN	TN	FN	TN	TN
1	0 TN	FN	FN	TN	FN	TN	TN	FN	TN	TN

PR<sub>BA</sub>

0	0	1	0,25	0	1	0,5	0	1	0,75	0	1	1	0	1
0	TN	FN	0	0	TN	2	FN	0	0	TN	5	2	0	TN
1	FP	TP	4	1	FP	4	TP	4	1	FP	2	1	1	0

rob accuracy dla 0,5

$$\text{DOKIADNOŚĆ KLASYFIKACJI} = \frac{TN + TP}{TN + TP + FN + FP} = \frac{4+3}{4+3+1+2} = \frac{7}{10}$$

$$\text{BLAD KLASYFIKACJI} = 1 - \text{dokiadność}' = 1 - \frac{7}{10} = \frac{3}{10}$$

$$\text{CZUOSĆ} = \frac{TP}{TP + FN} = \frac{3}{3+1} = \frac{3}{4}$$

$$\text{SPECYFICZNOŚĆ} = \frac{TN}{TN + FP} = \frac{4}{4+2} = \frac{4}{6} = \frac{2}{3}$$

$$\text{PRECYZJA} = \frac{TP}{TP + FP} = \frac{3}{3+2} = \frac{3}{5}$$

$$F_1 = \frac{2 \cdot \text{czuosc} \cdot \text{precyzja}}{\text{czuosc} + \text{precyzja}} = \frac{2 \cdot \frac{3}{4} \cdot \frac{3}{5}}{\frac{3}{4} + \frac{3}{5}} = \frac{\frac{18}{20}}{\frac{15+12}{20}} = \frac{\frac{18}{20}}{\frac{27}{20}} = \frac{2}{3}$$

$$\text{%BALANSOWANA} = \frac{1}{2} \cdot \text{czuosc} + \frac{1}{2} \text{specyficzność} = \frac{1}{2} \cdot \frac{3}{4} + \frac{1}{2} \cdot \frac{2}{3} = \frac{3}{8} + \frac{1}{3} =$$

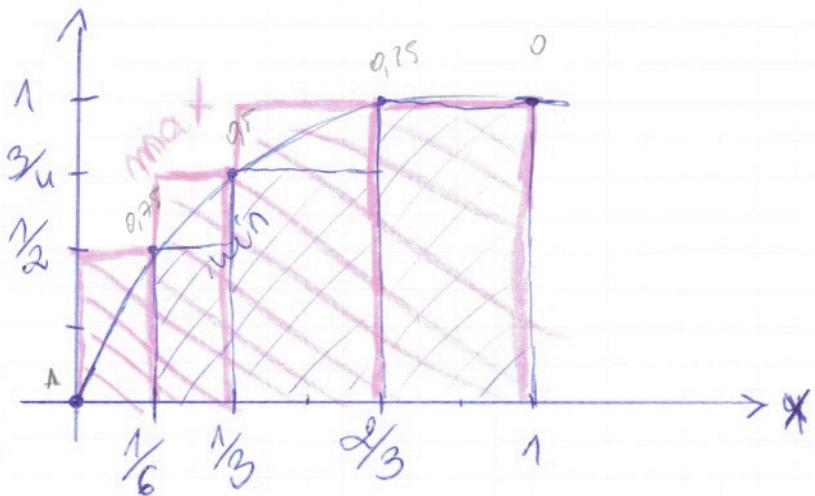
$$= \frac{9+8}{24} = \frac{17}{24}$$

# KWADRAT ROC

$$x = 1 - \text{specyficzność} = 1 - \frac{TN}{TN+FP}$$

$$y = \text{czułość} = \frac{TP}{TP+FN}$$

	czułość	specyficzność	1-specyficzność
0	$\frac{4}{4+0} = 1$	$0 + 6 = 0$	$1 - 0 = 1$
0,25	$\frac{1}{1+0} = 1$	$\frac{2}{2+4} = \frac{1}{3}$	$1 - \frac{1}{3} = \frac{2}{3}$
0,5	$\frac{2}{2+2} = \frac{1}{2}$	$\frac{5}{5+1} = \frac{5}{6}$	$1 - \frac{5}{6} = \frac{1}{6}$
1	$\frac{0}{0+4} = 0$	$\frac{6}{6+0} = 1$	$1 - 1 = 0$
0,15	$\frac{3}{4}$	$\frac{2}{3}$	$\frac{1}{3}$



$$AUC_{\min} = \frac{1}{6} \cdot \frac{1}{2} + \frac{1}{3} \cdot \frac{3}{4} + \frac{1}{3} \cdot 1 = \frac{1}{12} + \frac{3}{12} + \frac{1}{3} = \frac{8}{12} = \frac{2}{3}$$

$$\begin{aligned} AUC_{\max} &= \frac{1}{6} \cdot \frac{1}{2} + \frac{1}{6} \cdot \frac{3}{4} + \frac{1}{3} \cdot 1 + \frac{1}{3} \cdot 1 = \frac{1}{12} + \frac{3}{24} + \frac{2}{3} = \\ &= \frac{2}{24} + \frac{3}{24} + \frac{16}{24} = \frac{21}{24} = \frac{7}{8} \end{aligned}$$

#### **Zadanie 4.**

Dla danych z tabel policzyć następujące miary oceny klasyfikacji: dokładność, precyzję, czułość oraz  $F_1$ . Który z klasyfikatorów działa najlepiej?

Tabela 1. Macierz konfuzji – klasyfikator I

klasa/predykcja	A	B
A	140	60
B	20	22

Tabela 2. Macierz konfuzji – klasyfikator II

klasa/predykcja	A	B
A	100	100
B	2	40

#### **Zadanie 5.**

Pewna choroba występuje u 2 osób na 100. Wiedząc, że czułość klasyfikatora wykrywającego taką chorobę wynosi 0,8, a specyficzność 0,7, wykonać następujące zadania:

1. Wytlumaczyć znaczenie wskaźników czułość i specyficzność dla podanego przykładu.
2. Wyznaczyć zbalansowana dokładność.
3. Uzupełnić macierz konfuzji.
4. Wyznaczyć dokładność klasyfikacji.
5. Wyznaczyć prawdopodobieństwo poprawnego wykrycia choroby, wiedząc, że klasyfikator podjął decyzję o chorobie.
6. Wyznaczyć prawdopodobieństwo poprawnego wskazania osoby zdrowej wiedząc, że klasyfikator podjął decyzję o braku choroby.

# Zadanie 4

Tabela 1.

		KLASYFIKATOR	
		A (0)	B (1)
PROBKA	A (0)	TN 140	FP 60
	B (1)	FN 20	TP 22

$$\text{Dokładność} = \frac{TN + TP}{TN + TP + FN + FP} = \frac{140 + 22}{140 + 22 + 60 + 20} = \frac{82}{122} = 66\%$$

$$\text{Precyzyj} = \frac{TP}{TP + FP} = \frac{22}{60 + 22} = \frac{22}{82} = 27\%$$

$$\text{Czułość} = \frac{TP}{TP + FN} = \frac{22}{22 + 20} = \frac{22}{42} = \frac{11}{21} = 52\%$$

$$F_1 = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} = \frac{2 \cdot \text{czułość} \cdot \text{precyzyj}}{\text{czułość} + \text{precyzyj}} = \frac{44}{44 + 60 + 20} = \frac{44}{124} = \frac{11}{31} = 0,3434\%$$

		KLASYFIKATOR	
		A (0)	B (1)
PROBKA	A (0)	TN 100	FP 100
	B (1)	FN 2	TP 40

$$\text{Dokładność} = \frac{100 + 40}{100 + 40 + 2 + 100} = \frac{140}{242} = 58\%$$

$$\text{Precyzyj} = 40 / 40 + 100 = 40 / 140 = 28\%$$

$$\text{czułość} = \frac{40}{40 + 2} = \frac{40}{42} = 95\%$$

$$F_1 = \frac{2 \cdot 40}{2 \cdot 40 + 100 + 2} = \frac{80}{182} = 44\%$$

Im wyższe  $F_1$  tym lepsze

# Zadanie 5

1 - choroba

$N$  - liczba osób

0 - zdrowie

0,8 - czułość klasyfikatora

0,7 - specyficzność

pkt. 2 ZBALANSOWANA DOKŁADNOŚĆ =

$$= \frac{1}{2} \text{czułość} + \frac{1}{2} \text{specyficzność} = 0,4 + 0,35 = 0,75$$

p3. MATERIAŁ KONFYZJI

		Klasa 1	
		0	1
Klasa 0	0	TN 0,686N	FN 0,004N
	1	FP 0,234N	TP 0,016N
		0,98-N	0,02-N

0,98-0,7

0,98-TN

0,02-0,8

0,02-0,016

p4. DOKŁADNOŚĆ KLASYFIKACJI

$$= \frac{TN + TP}{TN + TP + FN + FP} = \frac{0,686N + 0,016N}{N} = \frac{0,702N}{N} = 0,702$$

p5. PRANDOPODBIENSTWO WŁKRYWIA CHOROBY

$$P(\text{R}^{\text{pred}} = 1) = \frac{TP}{TP + FN} = \frac{0,016}{0,016 + 0,004} = 0,8 \quad \text{lub} \quad \frac{IP}{IP + FP} = \frac{0,016}{0,016 + 0,234} = 0,05$$

85%? 5%

p6. PRANDOPDOB. WŁKRYWIA 0. DZIĘKUJĘ