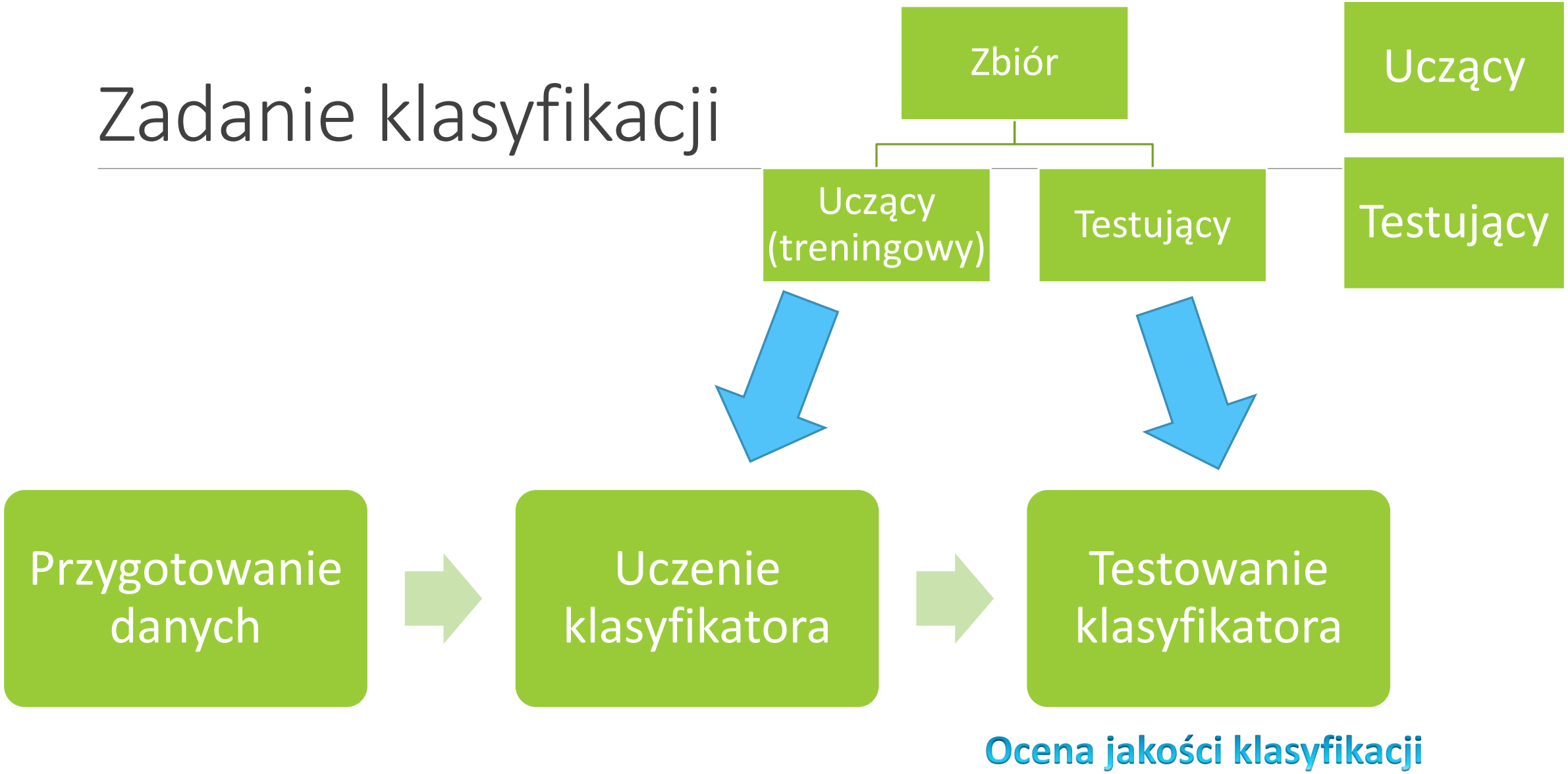
Uczenie maszynowe

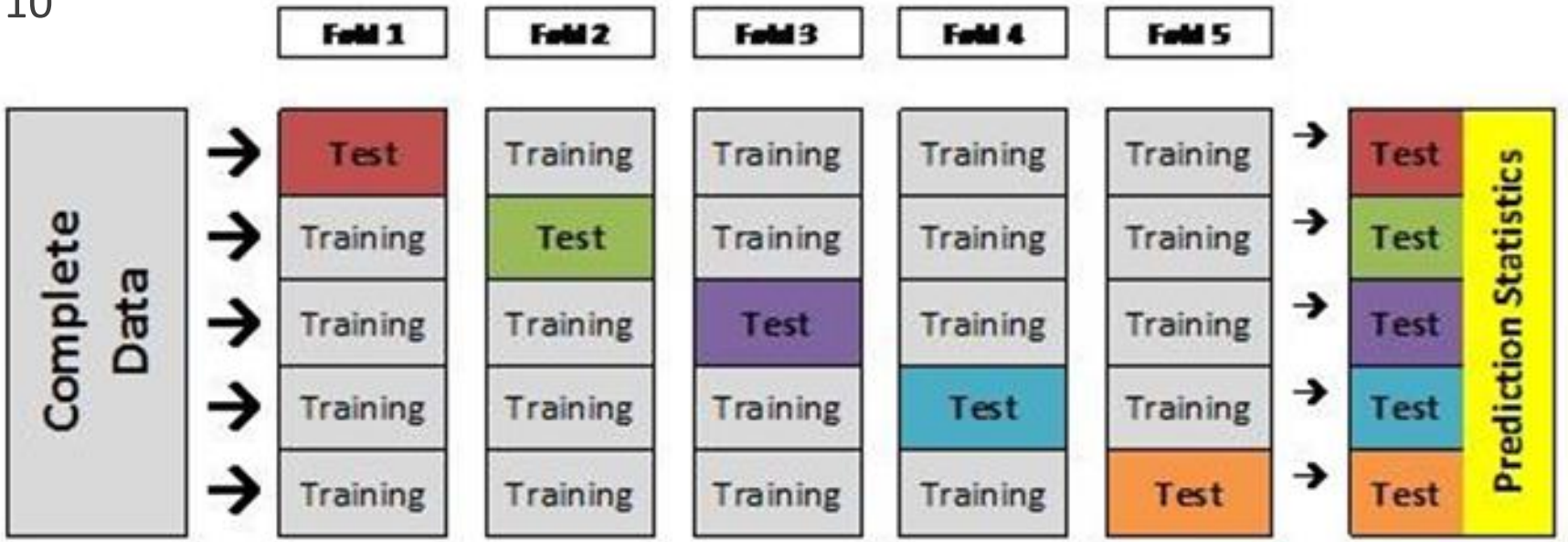
ZADANIE KLASYFIKACJI CZ. 2

Zadanie klasyfikacji



k-krotna walidacja krzyżowa (k-fold *cross-validation*)

- Podział oryginalnego zbioru na *k* podzbiorów (losowy): Każdy z nich wykorzystywany jest jako zbiór testujący, a pozostałe *k*-1 podzbiorów jako zbiór uczący. Dla każdej z *k* iteracji wykonuje się analizę, a na końcu wyniki są uśredniane.
- Najczęściej *k* równe 5 lub 10



Rysunek: https://stackoverflow.com/questions/40368467/cross-validation-extracting-the-model-values-out-per-row
http://edu.pjwstk.edu.pl/wyklady/adn/scb/rW9.htm

Naiwny klasyfikator Bayesa

- Klasyfikator statystyczny, oparty na twierdzeniu Bayesa:

> Przykład X klasyfikujemy jako pochodzący z tej klasy C_i, dla której wartość $P(C_i|X)$, i = 1, 2, ..., m, jest największa

gdzie X oznacza przykład z nieznaną klasą. Każdy przykład jest opisany jako wektor o n wymiarach $X = (x_1,\ldots,x_n)$;

P(C|X) prawdopodobieństwo a-posteriori, że przykład X należy do klasy C

- **Założenie: wzajemna niezależność atrybutów (naiwność …)**

http://wazniak.mimuw.edu.pl/images/4/46/ED-4.2-m09-1.0.pdf

P-stwo warunkowe

- P-stwo warunkowe zajścia zdarzenia B pod warunkiem zajścia zdarzenia A:

$$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$$

- Zadanie:

70% twoich koleżanek/kolegów lubi czekoladę, a 35% czekoladę i truskawki. Ile procent z tych, którzy lubią czekoladę, lubi również truskawki?

$$P(truskawki|czekolada) = \frac{P(czekolada \cap truskawki)}{P(czekolada)} = \frac{0{,}35}{0{,}7} = 0{,}5$$

https://www.mathsisfun.com/data/probability-events-conditional.html

Naiwny klasyfikator Bayesa

Wyznaczanie prawdopodobieństwa a-posteriori, że przykład X należy do klasy C na podstawie twierdzenia Bayesa:

$$P(C|X) = (P(X|C) * P(C))/P(X),$$

gdzie: **P(C)** oznacza prawdopodobieństwo a-priori wystąpienia klasy C (tzn. prawdopodobieństwo, że dowolny przykład należy do klasy C);

P(X|C) oznacza prawdopodobieństwo a-posteriori, że X należy do klasy C;

P(X) oznacza prawdopodobieństwo a-priori wystąpienia przykładu X.

Wyprowadzenie:

$$P(C|X) = \frac{P(C \cap X)}{P(X)} = \frac{P(X|C) \times P(C)}{P(X)}, bo\ P(X|C) = \frac{P(C \cap X)}{P(C)}$$

http://wazniak.mimuw.edu.pl/images/4/46/ED-4.2-m09-1.0.pdf

Naiwny klasyfikator Bayesa

- Dany jest zbiór treningowy D składający się z n przykładów
- Załóżmy, że atrybut decyzyjny przyjmuje m różnych wartości definiując m różnych klas C_i, i = 1, ..., m
- Niech s_i oznacza liczbę przykładów z D należących do klasy C_i
- Klasyfikator Bayesa przypisuje nieznany przykład X do tej klasy C_i, dla której wartość P(C_i|X) jest największa

- Prawdopodobieństwo P(X) jest stałe dla wszystkich klas - klasa C_i, dla której wartość $P(C_i|X)$ jest największa, to klasa C_i, dla której wartość $P(X| C_i) * P(C_i)$ jest największa
- Wartości $P(C_i)$ zastępujemy estymatorami s_i/n (względną częstością klasy C_i), lub zakładamy, że wszystkie klasy mają to samo prawdopodobieństwo $P(C_1) = P(C_2) = ... = P(C_m)$

$$P(C|X) = (P(X|C) * P(C))/P(X),$$

Naiwny klasyfikator Bayesa

- W jaki sposób obliczyć $P(X|C_i)$?
- Dla dużych zbiorów danych, o dużej liczbie deskryptorów, obliczenie $P(X|C_i)$ będzie bardzo kosztowne
- Przyjmujemy założenie o **niezależności atrybutów**
- Założenie o niezależności atrybutów prowadzi do następującej formuły:

$$P(X|C_i) = \prod_{j=1}^{n} P(x_j \mid C_i)$$

- Prawdopodobieństwa $P(x_1|C_i)$, $P(x_2|C_i)$, ..., $P(x_n|C_i)$ można estymować w oparciu o zbiór treningowy następująco:

jeżeli j-ty atrybut jest atrybutem kategorycznym, to $P(x_j|C_i)$ estymujemy względną częstością występowania przykładów z klasy C_i posiadających wartość x_j dla j-tego atrybutu, (s_{ij}/s_i)

jeżeli j-ty atrybut jest atrybutem ciągłym, to $P(x_j|C_i)$ estymujemy **funkcją gęstości Gaussa**

$$f(x) \frac{1}{\sigma \sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\cdot\sigma^2}}$$

(zakładając rozkład normalny wartości atrybutów)

$$P(C|X) = (P(X|C) * P(C))/P(X),$$

http://wazniak.mimuw.edu.pl/images/4/46/ED-4.2-m09-1.0.pdf

Zbiór czyKupiKomputer

- Atrybuty:
- wiek {≤30, 31...40, >40}
- dochód {duży, średni, mały}
- czy student? {tak, nie}
- zdolność kredytowa {umiark, doskonala}
- czy kupi komputer? {tak, nie} → KLASA

wiek	dochod	stud	zdolKred	czyKupi
≤ 30	duży	nie	umiark	nie
≤ 30	duży	nie	doskonala	nie
$31 \ldots 40$	duży	nie	umiark	tak
> 40	sredni	nie	umiark	tak
> 40	mały	tak	umiark	tak
> 40	mały	tak	doskonala	nie
$31 \ldots 40$	mały	tak	doskonała	tak
≤ 30	sredni	nie	umiark	nie
≤ 30	mały	tak	umiark	tak
> 40	sredni	tak	umiark	tak
≤ 30	sredni	tak	doskonala	tak
$31 \ldots 40$	sredni	nie	doskonala	tak
$31 \ldots 40$	duży	tak	umiark	tak
> 40	sredni	nie	doskonala	nie

Dane: https://inf.ug.edu.pl/~jj/materialydyd/_CI/przykladCzyKupi.pdf

Zadanie czyKupiKomputer

Nowy przypadek: **X = (wiek = ≤30, dochód = duży, student = tak, zdolność kredytowa = doskonała)**

$P(czyKupi = tak) = P(C1) = 9/14 \approx 0{,}643$

$P(czyKupi = nie) = P(C2) = 5/14 \approx 0{,}357$

$P(wiek = \leq 30 | czyKupi = tak) = 2/9 \approx 0{,}222$

$P(wiek = \leq 30 | czyKupi = nie) = 3/5 \approx 0{,}6$

$P(dochód = duży | czyKupi = tak) = 2/9 \approx 0{,}222$

$P(dochód = duży | czyKupi = nie) = 2/5 \approx 0{,}4$

$P(student = tak | czyKupi = tak) = 6/9 \approx 0{,}667$

$P(student = tak | czyKupi = nie) = 1/5 \approx 0{,}2$

$P(zdolKred = doskonała | czyKupi = tak) = 3/9 \approx 0{,}333$

$P(zdolKred = doskonała | czyKupi = nie) = 3/5 \approx 0{,}6$

wiek	dochod	stud	zdolKred	czyKupi
≤ 30	duży	nie	umiark	nie
≤ 30	duży	nie	doskonala	nie
$31 \ldots 40$	duży	nie	umiark	tak
> 40	sredni	nie	umiark	tak
> 40	mały	tak	umiark	tak
> 40	mały	tak	doskonala	nie
$31 \ldots 40$	mały	tak	doskonała	tak
≤ 30	sredni	nie	umiark	nie
≤ 30	mały	tak	umiark	tak
> 40	sredni	tak	umiark	tak
≤ 30	sredni	tak	doskonala	tak
$31 \ldots 40$	sredni	nie	doskonala	tak
$31 \ldots 40$	duży	tak	umiark	tak
> 40	sredni	nie	doskonala	nie

Zadanie czyKupiKomputer

$$P(X|czyKupi = tak) = P(wiek = \leq 30|czyKupi = tak) \times P(dochód = duży|czyKupi = tak) \times$$
$$\times P(student = tak|czyKupi = tak) \times P(zdolKred = doskonała|czyKupi = tak) =$$
$$= 0{,}222 \times 0{,}222 \times 0{,}667 \times 0{,}333 \approx 0{,}011$$

$$P(X|czyKupi = nie) = 0{,}6 \times 0{,}4 \times 0{,}2 \times 0{,}6 \approx 0{,}029$$

$$P(X|czyKupi = tak) \times P(czyKupi = tak) = 0{,}011 \times 0{,}643 \approx 0{,}007$$

$$P(X|czyKupi = nie) \times P(czyKupi = nie) = 0{,}029 \times 0{,}357 \approx 0{,}010$$

Przypadkowi X zostanie przypisana etykieta klasy:

$$czyKupi = nie$$

Problem zmiennych numerycznych

- Gra w golfa w zależności od pogody
- Atrybuty:

- Temperatura

- Wilgotność

- czyGramy? Play {yes, no}

Jak policzyć p-stwo wystąpienia poszczególnych wartości np. temperatury 69?

Rozwiązanie – dyskretyzacja!

Temperature	Humidity	Play
85	85	no
80	90	no
65	70	no
72	95	no
71	80	no
83	78	yes
70	96	yes
68	80	yes
64	65	yes
69	70	yes
75	80	yes
75	70	yes
72	90	yes
81	75	yes

http://www.simafore.com/blog/bid/107702/2-ways-of-using-Naive-Bayes-classification-for-numeric-attributes

Zadanie 1

Na podstawie zbioru czyKupiKomputer zaklasyfikuj dwa następujące przypadki:

1) **X = (wiek = 31 … 40, dochód = sredni, student = nie, zdolność kredytowa = umiark)**
2) **swój własny przypadek**

wiek	dochod	stud	zdolKred	czyKupi
≤ 30	duży	nie	umiark	nie
≤ 30	duży	nie	doskonala	nie
$31 \ldots 40$	duży	nie	umiark	tak
> 40	sredni	nie	umiark	tak
> 40	mały	tak	umiark	tak
> 40	mały	tak	doskonala	nie
$31 \ldots 40$	mały	tak	doskonała	tak
≤ 30	sredni	nie	umiark	nie
≤ 30	mały	tak	umiark	tak
> 40	sredni	tak	umiark	tak
≤ 30	sredni	tak	doskonala	tak
$31 \ldots 40$	sredni	nie	doskonala	tak
$31 \ldots 40$	duży	tak	umiark	tak
> 40	sredni	nie	doskonala	nie