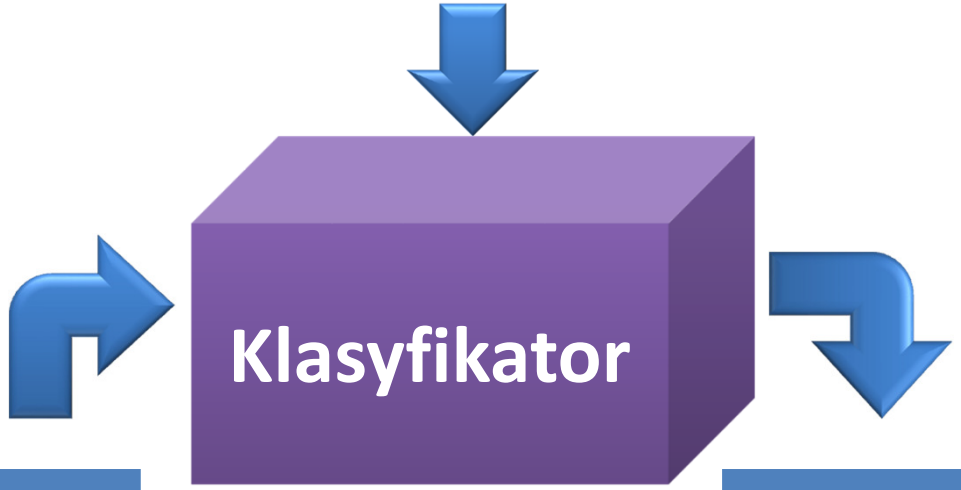

Klasyfikacja

Klasyfikacja

TEMP	BÓL	WYSYPKA	GARDŁO	DIAGNOZA
36.6	T	BRAK	NORMA	NIESTRAWNOŚĆ
37.5	N	MAŁA	PRZEKR.	ALERGIA
36.0	N	BRAK	NORMA	PRZECHŁODZENIE
39.5	T	DUŻA	PRZEKR.	CZARNA OSPA
…	…	…	…	…



TEMP	BÓL	WYSYPKA	GARDŁO
38.0	N	BRAK	NORMA

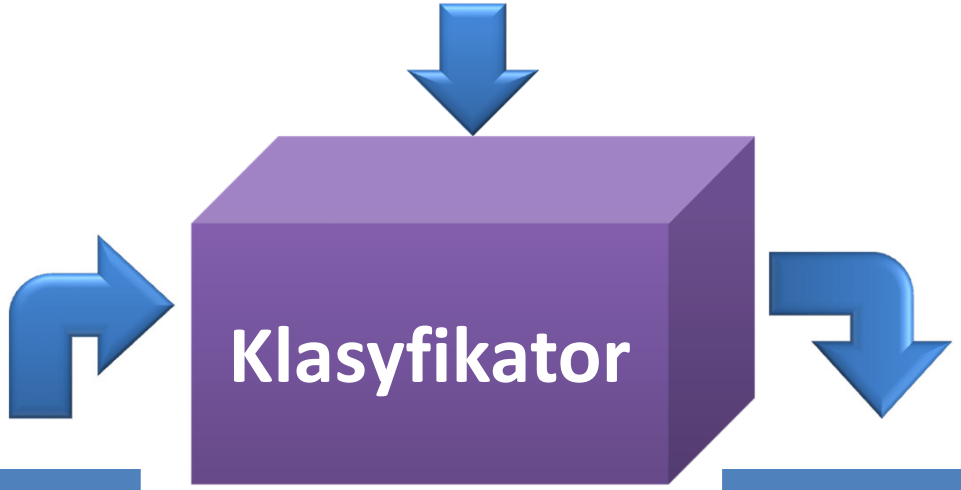
DIAGNOZA
ALERGIA

Klasyfikacja

TEMP	BÓL	WYSYPKA	GARDŁO	DIAGNOZA
36.6	N	BRAK	NORMA	N
37.5	N	MAŁA	PRZEKR.	ALERGIA
36.0	N	BRAK	NORMA	PRZECHŁODZENIE
39.5	T	DUŻA	PRZEKR.	CZARNA OSPA
…	…	…	…	…

Deskryptory/
Atrybuty opisowe

Atrybut
decyzyjny



TEMP	BÓL	WYSYPKA	GARDŁO
38.0	N	BRAK	NORMA

DIAGNOZA
ALERGIA

Klasyfikacja

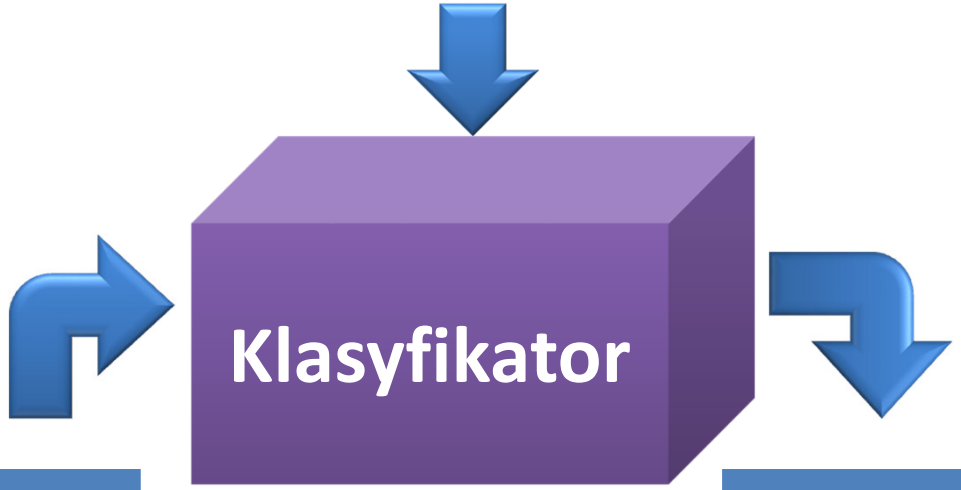
TEMP	BÓL	WYSYPKA	GARDŁO	DIAGNOZA
36.			NORMA	N
37.5	N	MAŁA	PRZEKR.	ALERGIA
36.			NORMA	P
39.			PRZEKR.	C
…			…	…

Deskryptory/ Atrybuty opisowe

Atrybuty kategoryczne i ciągłe

Atrybut decyzyjny

Atrybut kategoryczny



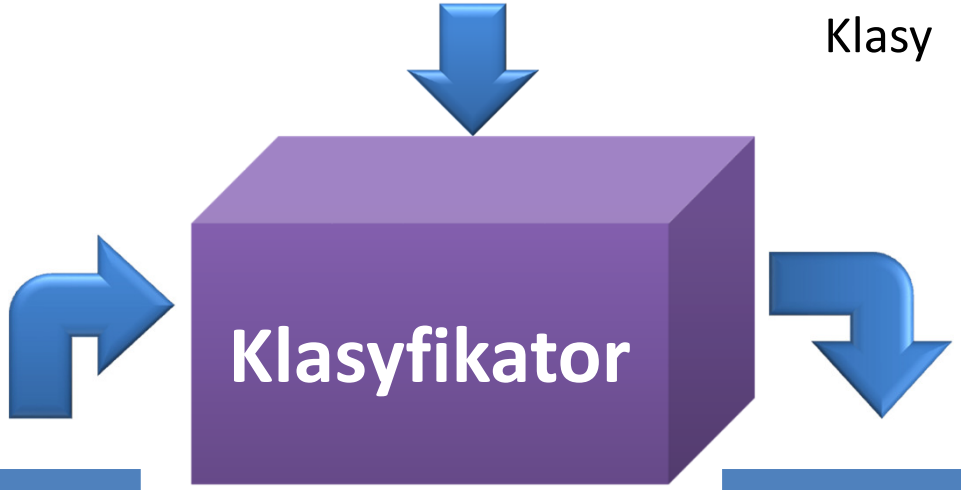
TEMP	BÓL	WYSYPKA	GARDŁO
38.0	N	BRAK	NORMA

DIAGNOZA
ALERGIA

Klasyfikacja

TEMP	BÓL	WYSYPKA	GARDŁO	DIAGNOZA
36.6	T	BRAK	NORMA	NIESTRAWNOŚĆ
37.5	N	MAŁA	PRZEKR.	ALERGIA
36.0	N	BRAK	NORMA	PRZECHŁODZENIE
39.5	T	DUŻA	PRZEKR.	CZARNA OSPA
…	…	…	…	…

Etykiety klas
Klasy decyzyjne
Klasy



TEMP	BÓL	WYSYPKA	GARDŁO
38.0	N	BRAK	NORMA

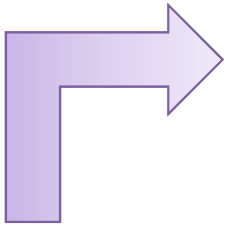
DIAGNOZA
ALERGIA

Czym jest klasyfikacja?

- Cel: znalezienie dla każdego zbioru danych D *funkcji klasyfikacyjnej* (klasyfikatora), która odwzorowuje każdy rekord danych w etykietę klasy.
- Funkcja klasyfikacyjna służy do predykcji wartości atrybutu decyzyjnego rekordów, dla których wartość ta nie jest znana.
- 2 etapy:
  - **Etap 1**: budowa modelu (klasyfikatora) opisującego predefiniowany zbiór klas danych lub zbiór pojęć
  - **Etap 2**: zastosowanie opracowanego modelu do klasyfikacji nowych danych

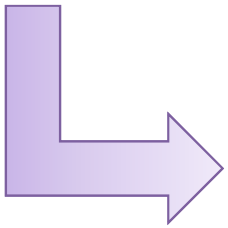
Proces budowy klasyfikatora

TEMP	…	GARDŁO	DIAGNOZA
36.6	…	NORMA	NIESTR.
37.5	…	PRZEKR.	ALERGIA
36.0	…	NORMA	PRZECHŁ.
39.5	…	PRZEKR.	CZ. OSPA
…	…	…	…



Zbiór uczący

TEMP	…	GARDŁO	DIAGNOZA
36.6	…	NORMA	NIESTR.
37.5	…	PRZEKR.	ALERGIA
36.0	…	NORMA	PRZECHŁ.
…	…	…	…



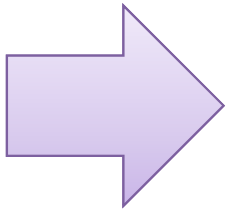
Zbiór testujący

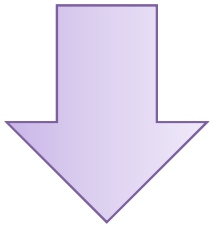
TEMP	…	GARDŁO	DIAGNOZA
39.5	…	PRZEKR.	CZ. OSPA
…	…	…	…

Proces budowy klasyfikatora

Zbiór uczący

TEMP	…	GARDŁO	DIAGNOZA
36.6	…	NORMA	NIESTR.
37.5	…	PRZEKR.	ALERGIA
36.0	…	NORMA	PRZECHŁ.
…	…	…	…

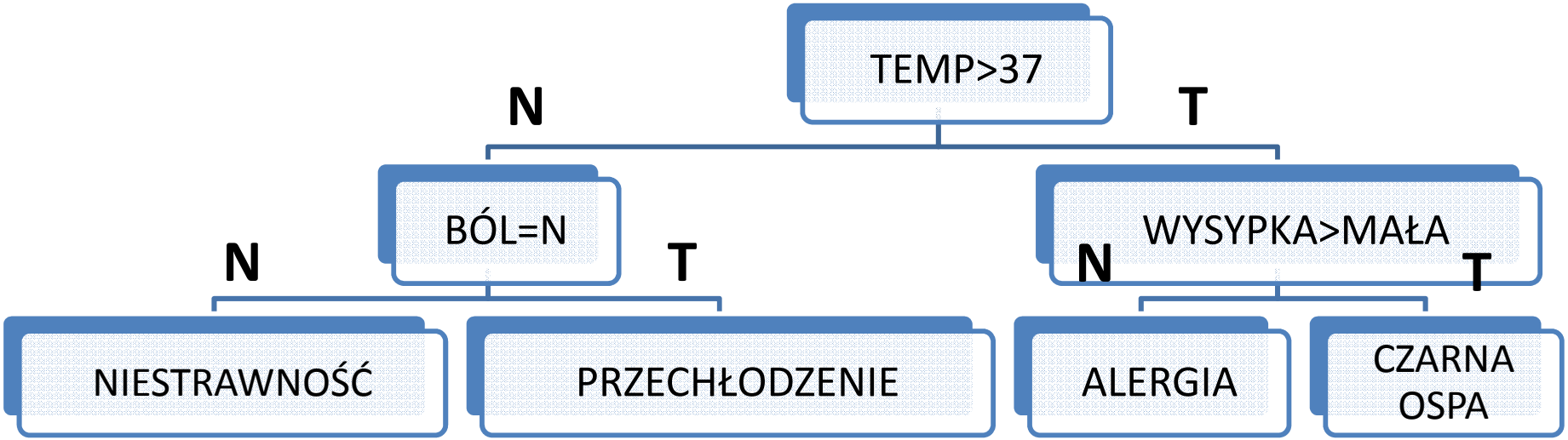




Zbiór testujący

TEMP	…	GARDŁO	DIAGNOZA	DIAGNOZA
39.5	…	PRZEKR.	CZ. OSPA	NIESTR.
…	…	…	…	

Drzewa decyzyjne



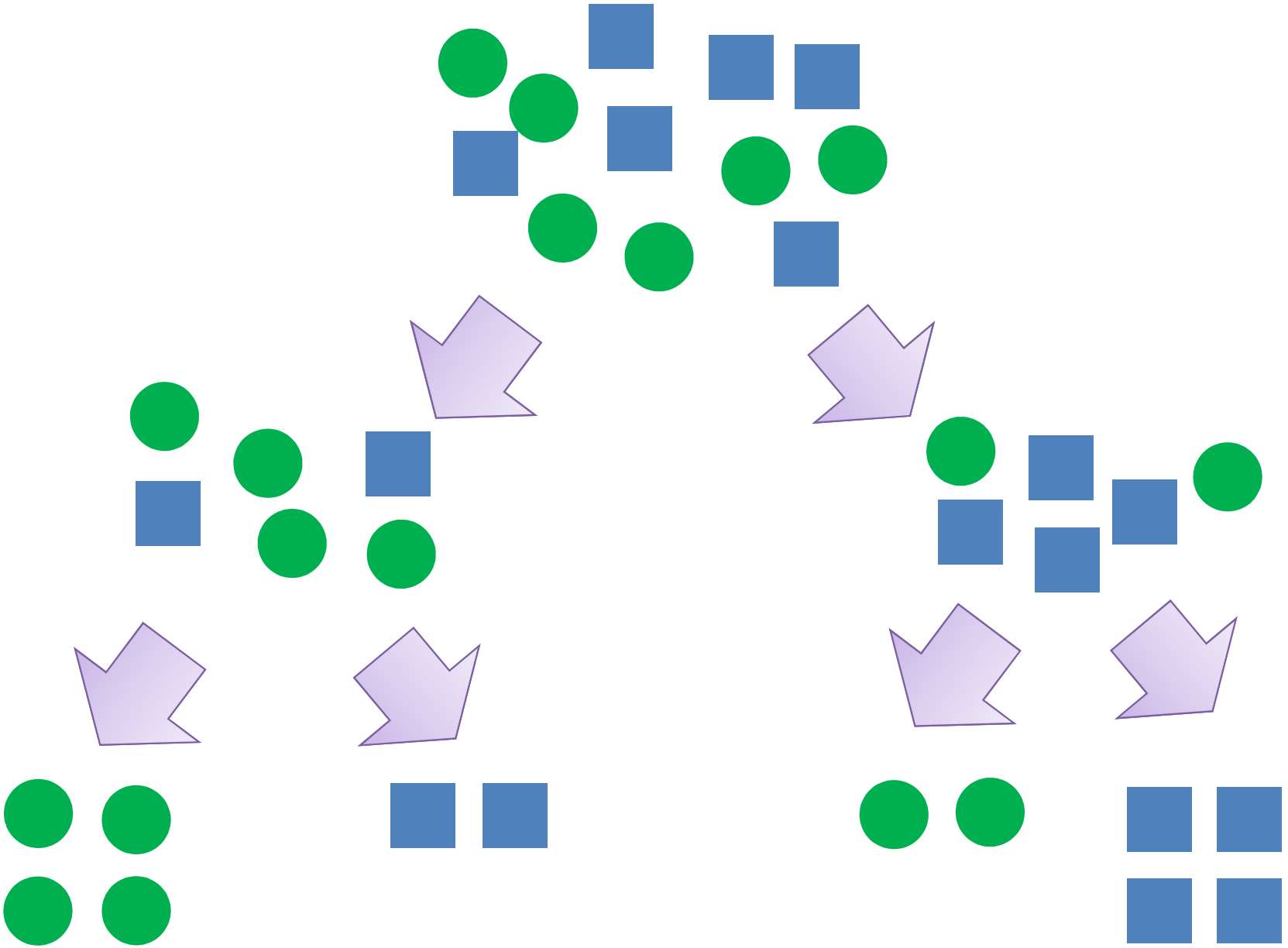
Indukcja drzew decyzyjnych

1. Konstrukcja drzewa decyzyjnego w oparciu o zbiór treningowy
2. Przycinanie drzewa w celu poprawy dokładności, interpretowalności i uniezależnienia się od efektu przetrenowania

Indukcja drzew decyzyjnych

- Algorytm podstawowy: algorytm zachłanny, który konstruuje rekurencyjnie drzewo decyzyjne metodą top-down
- Wiele wariantów algorytmu podstawowego (źródła):
  - uczenie maszynowe (ID3, C4.5)
  - statystyka (CART)
  - rozpoznawanie obrazów (CHAID)
- Podstawowa różnica: kryterium podziału

Indukcja drzew decyzyjnych



Jak wybrać najlepszy podział?

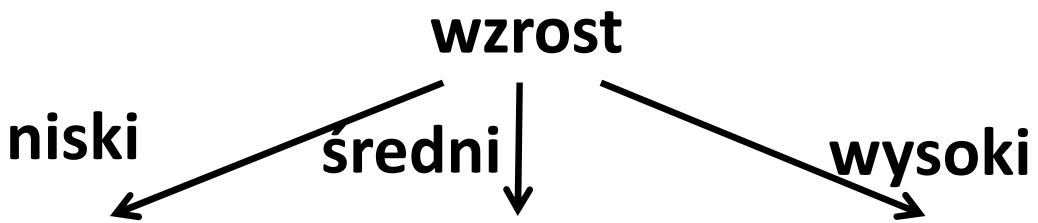
- Indeks Giniego (Gini indeks) (algorytmy CART, SPRINT)
  - Wybieramy atrybut, który minimalizuje indeks Giniego
- Zysk informacyjny (algorytmy ID3, C4.5)
  - Wybieramy atrybut, który maksymalizuje redukcję entropii
- Indeks korelacji χ^2 (algorytm CHAID)
  - Mierzymy korelację pomiędzy każdym atrybutem i każdą klasą (wartością atrybutu decyzyjnego)
  - Wybieramy atrybut o maksymalnej korelacji

Kiedy przestać dzielić?

- Podstawowe warunki zatrzymania
  - Kiedy wszystkie przykłady w podzbiorze należą do jednej klasy
  - Wykorzystano wszystkie możliwe podziały
- Dodatkowe warunki zatrzymania
  - Jeżeli wartości wszystkich atrybutów w podzbiorze są identyczne
  - Liczba przykładów w podzbiorze jest mniejsza od zadanego progu
  - inne

Algorytm ID3

- Wszystkie atrybuty opisowe muszą być nominalne
- Podział zawsze grupuje przykłady wg. wartości atrybutu, np. dla atrybutu wzrost, o dziedzinie: $\{niski, średni, wysoki\}$ podział:



- Do wyboru atrybutu stosuje miarę *zysku informacyjnego*

Oznaczenia

- S – zbiór $|S|$ przykładów
- C_i - i-ta klasa decyzyjna ($i = 1, \ldots, m$)
- s_i - liczba przykładów zbioru S należących do klasy C_i
- p_i - prawdopodobieństwo, że dowolny przykład należy do klasy C_i, $p_i \approx s_i/|S|$
- S_j - partycje zbioru S powstałe przez test na atrybucie A ($j = 1, \ldots, v$, gdzie v to rozmiar dziedziny atrybutu A) .
  - Np., jeżeli dziedzina atrybutu A posiada v=3 różne wartości to powstaną 3 partycje

Entropia informacyjna

- $I(s_1, s_2, \ldots, s_m) = -\sum_{i=1}^{m} p_i log_2 p_i$
  - Pamiętamy, że $p_i \approx s_i/|S|$
- Entropia jest miarą nieuporządkowania
- Im większa entropia, tym większy „bałagan” w danych.

Zbiór S	s_1	p_1	s_2	p_2	Entropia
	2	0.5	2	0.5	-(0.5*log$_2$0.5+0.5*log$_2$0.5)= -(0.5*-1+0.5*-1)= -(-1)=1
	3	0.75	1	0.25	-(0.75*log$_2$0.75+0.25*log$_2$0.25)= -(-0.31+0.25*-2)=0.81
	4	1	0	0	-(1*log$_2$1+0*log$_2$0)=0

Entropia po podziale wg. atrybutu A

- Każdy przykład ze zbioru S posiada klasę C_i
- Podział wg. atrybutu dzieli zbiór S na partycje S_j
- s_{ij} to liczba przykładów z partycji S_j które posiadają klasę C_i
- Entropia podziału zbioru:

$$-E(A) = \sum_{j=1}^{v} \frac{|S_j|}{|S|} * I(s_{1j}, s_{2j}, \dots, s_{mj})$$

- Im mniejsza entropia, tym większy porządek wprowadza podział

Zysk informacyjny

- **Zysk informacyjny**, wynikający z podziału zbioru S na partycje według atrybutu A, definiujemy następująco:

$$Gain(A) = I(s_1, s_2, \ldots, s_m) - E(A)$$

- Oblicza oczekiwaną redukcję entropii (nieuporządkowania) uzyskaną dzięki testowi na atrybucie A.

Przykład

outlook	temp	humidity	windy	play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

$|S| = 14 \quad s_1 = 9 \quad p_1 = 9/14 = 0.64$ (yes)

$s_2 = 5 \quad p_2 = 5/14 = 0.36$ (no)

$$I(s_1, s_2) = -\frac{9}{14} log_2 \frac{9}{14} - \frac{5}{14} log_2 \frac{5}{14} = 0.94$$

Przykład

outlook	temp	humidity	windy	play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Outlook

overcast $s_{11} = 4 \quad p_{11} = 4/4 = 1$

$s_{21} = 0 \quad p_{21} = 0/4 = 0$

rainy $s_{12} = 3 \quad p_{12} = 3/5 = 0.6$

$s_{22} = 2 \quad p_{22} = 2/5 = 0.4$

sunny $s_{13} = 2 \quad p_{13} = 2/5 = 0.4$

$s_{23} = 3 \quad p_{23} = 3/5 = 0.6$

$$E(outlook) =$$
$$= \frac{4}{14} I(4,0) + \frac{5}{14} I(3,2) + \frac{5}{14} I(2,3) =$$
$$= \frac{4}{14} (-1 log_2 1 - 0 log_2 0) +$$
$$+ \frac{10}{14} (-0.6 log_2 0.6 - 0.4 log_2 0.4) =$$
$$= \frac{4}{14} * 0 + \frac{10}{14} * 0.97 = 0.69$$

$$\mathrm{GA}IN(outlook) = 0.94 - 0.69 = 0.25$$

Przykład

outlook	temp	humidity	windy	play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Temperature

hot $s_{11} = 2 \quad p_{11} = 2/4 = 0.5$

$s_{21} = 2 \quad p_{21} = 2/4 = 0.5$

cool $s_{12} = 3 \quad p_{12} = 3/4 = 0.75$

$s_{22} = 1 \quad p_{22} = 1/4 = 0.25$

mild $s_{13} = 4 \quad p_{13} = 4/6 = 0.67$

$s_{23} = 2 \quad p_{23} = 2/6 = 0.33$

$$E(temperature) =$$

$$= \frac{4}{14} I(2,2) + \frac{4}{14} I(3,1) + \frac{6}{14} I(4,2) =$$

$$= \frac{4}{14}(-0.5 log_2 0.5 - 0.5 log_2 0.5) +$$

$$+ \frac{4}{14}(-0.75 log_2 0.75 - 0.25 log_2 0.25) +$$

$$+ \frac{6}{14}(-0.67 log_2 0.67 - 0.33 log_2 0.33) = \frac{4}{14} + \frac{4}{14} * 0.81 + \frac{6}{14} * 0.92 = 0.91$$

$$\mathrm{GA}IN(outlook) = 0.94 - 0.91 = 0.03$$

Przykład

outlook	temp	humidity	windy	play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Humidity

high $s_{11} = 3 \quad p_{11} = 3/7 = 0.43$

$s_{21} = 4 \quad p_{21} = 4/7 = 0.57$

normal $s_{12} = 6 \quad p_{12} = 6/7 = 0.86$

$s_{22} = 1 \quad p_{22} = 1/7 = 0.14$

$$E(humidity) = \frac{7}{14} I(3,4) + \frac{7}{14} I(6,1) =$$

$$= \frac{7}{14}\left(-\frac{3}{7} log_2 \frac{3}{7} - \frac{4}{7} log_2 \frac{4}{7}\right) +$$

$$+ \frac{7}{14}\left(-\frac{6}{7} log_2 \frac{6}{7} - \frac{1}{7} log_2 \frac{1}{7}\right) = 0.79$$

$$\mathrm{GA}IN(humidity) = 0.94 - 0.79 = 0.15$$

Przykład

outlook	temp	humidity	windy	play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Windy

true

$$s_{11} = 3 \quad p_{11} = 3/6 = 0.5$$
$$s_{21} = 3 \quad p_{21} = 3/6 = 0.5$$

false

$$s_{12} = 6 \quad p_{12} = 6/8 = 0.75$$
$$s_{22} = 2 \quad p_{22} = 2/8 = 0.25$$

$$E(windy) =$$
$$= \frac{6}{14} I(3,3) + \frac{8}{14} I(6,2) =$$
$$= \frac{6}{14} * 1 + \frac{8}{14} * 0.81 = 0.89$$

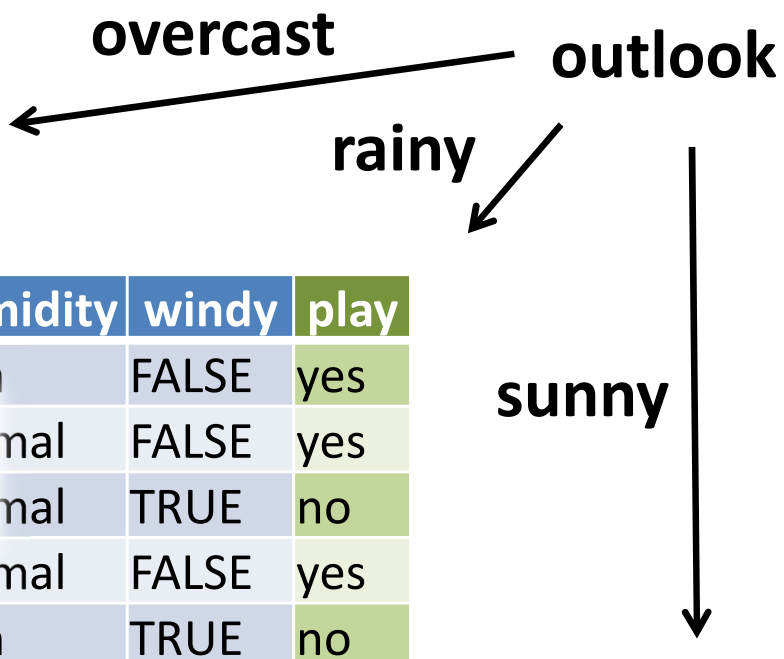
$$\mathrm{GA}IN(humidity) = 0.94 - 0.89 = 0.05$$

Przykład

Atrybut	Gain
Outlook	0.25
Temperature	0.03
Humidity	0.15
Windy	0.05

Podział wg. outlook

outlook	temp	humidity	windy	play
overcast	hot	high	FALSE	yes
over	Koniec	rmal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes



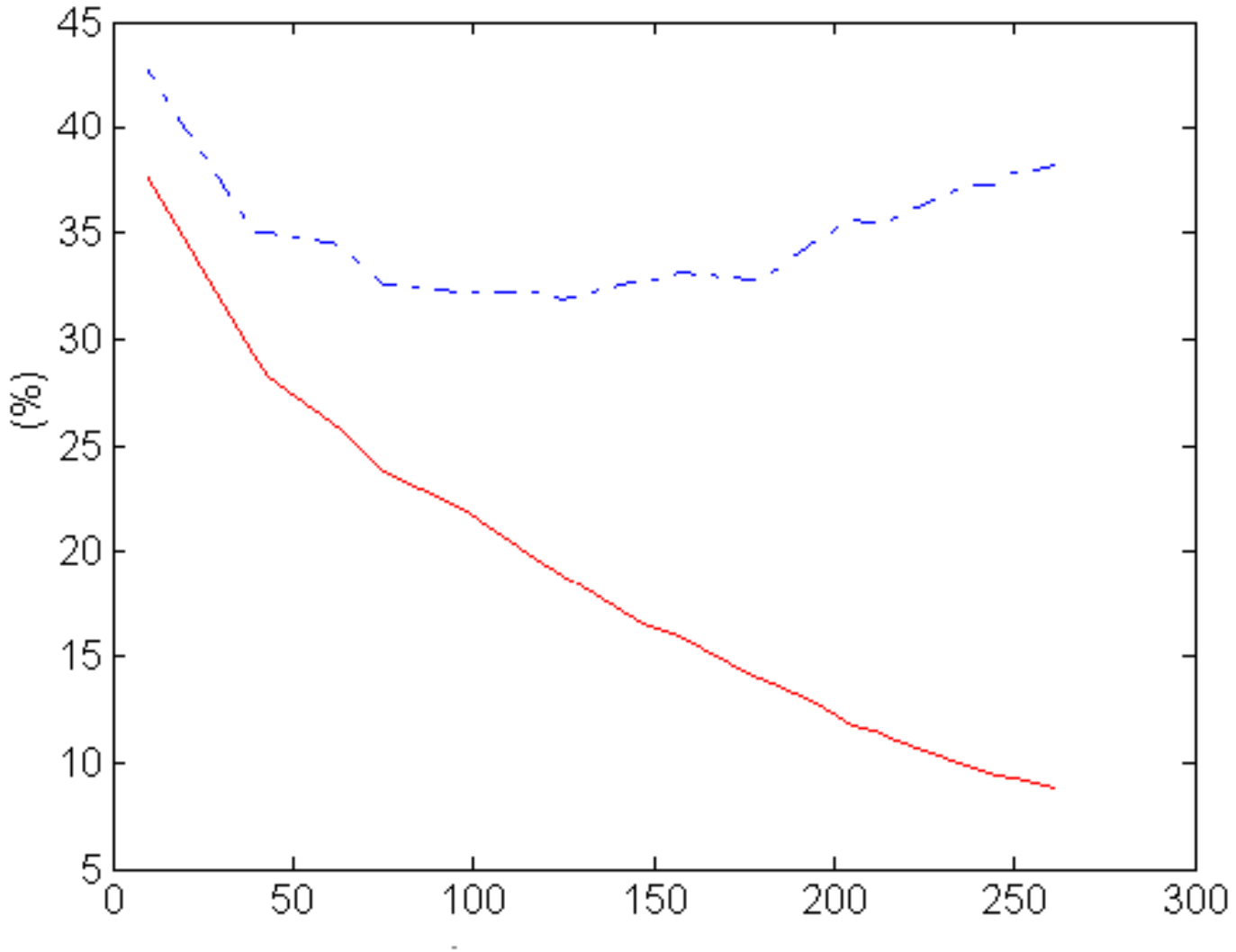
outlook	temp	humidity	windy	play
rainy	mild	high	FALSE	yes
rai	Podzial wg.	mal	FALSE	yes
rai	windy	mal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no

outlook	temp	humidity	windy	play
s		igh	FALSE	no
s	Podział wg.	igh	TRUE	no
s	humidity	igh	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Przeuczenie klasyfikatora

- Zbyt silne dopasowanie do danych
  - Niepoprawne/sprzeczne przykłady uczące
  - Za mało przykładów uczących
  - Niereprezentacyjność przykładów uczących
- Klasyfikator nauczył się na pamięć a nie utworzył generalizacji danych
- Formalnie, o przeuczeniu klasyfikatora mówimy wtedy, jeżeli możliwe jest stworzenie innego klasyfikatora o większym błędzie treningowym a mniejszym testowym

Przeuczenie klasyfikatora



Przycinanie drzewa

- Nadmierne dopasowanie objawia się gałęziami w drzewie dostosowanymi do anomalii i punktów osobliwych
- Przycinanie drzew decyzyjnych – usuwanie mało wiarygodnych gałęzi
  - poprawia efektywność klasyfikacji
  - poprawia zdolność klasyfikatora do klasyfikacji nowych przypadków

Przycinanie drzewa

- Wstępne przycinanie
  - Zatrzymanie dalszego podziału, jeżeli będzie on tworzyć zbyt specyficzne gałęzie
- Przycinanie po zakończeniu konstrukcji
  - Budowane jest całe drzewo, a potem usuwane są niektóre gałęzie jeżeli poprawi to jakość klasyfikacji

Klasyfikator Bayesa

- Szacuje prawdopodobieństwo, że dany obiekt należy do danej klasy decyzyjnej
- Przydziela obiektowi najbardziej prawdopodobną klasę
- Oparty o twierdzenie Bayesa, które wiąże ze sobą prawdopodobieństwa *apriori* i *aposteriori*.
- Trudny do zastosowania w praktyce – stosuje się jego uproszczenie: *naiwny klasyfikator Bayesa*

Definicje

- Niech X oznacza przykład - każdy przykład jest reprezentowany w postaci n-wymiarowego wektora, X=(x1, x2, ..., xn)
- Prawdopodobieństwo *apriori*
  - $P(X|C=C_i)$ – prawdopodobieństwo, że obiekt będzie miał wartości atrybutów x1, x2, ..., xn, jeżeli należy do klasy C_i
- Prawdopodobieństwo *aposteriori*
  - $P(C=C_i|X)$ – prawdopodobieństwo, że obiekt należy d klasy C_i jeżeli wartości jego atrybutów wynoszą x1, x2, ..., xn.

Twierdzenie Bayesa

- Wzór

$$P(C = C_i \mid X) = \frac{P(X \mid C = C_i)\, P(C = C_i)}{P(X)}$$

- Gdzie:
  - $P(X)$ to prawdopodobieństwo wystąpienia obiektu X
  - $P(C = C_i)$ to prawdopodobieństwo wystąpienia biektu o klasie C_i

Klasyfikator Bayesa

- Znajdź najbardziej prawdopodobną klasę dla obiektu X wykorzystując twierdzenie Bayesa
- $P(X)$ jest takie samo dla każdej klasy, więc można pominąć ze wzoru twierdzenia.
- Klasyfikator zatem szuka klasy, dla której wartość wyrażenia:

$$F(C_i) = P(X \mid C = C_i)\, P(C = C_i)$$

- Przyjmuje największą wartość

„Naiwność” klasyfikatora

- Dokładnie oszacowanie prawdopodobieństw

$$P(X \mid C = C_i)$$

  wymaga nierealistycznie dużych zbiorów danych

- W celu ominięcia tego problemu zakłada się **warunkową niezależność atrybutów**, stąd dla X=(x1, x2, ..., xn) otrzymujemy

$$P(X \mid C = C_i) = \prod_{j=1}^{n} P\left(A_j = x_j \mid C = \mathrm{C_i}\right)$$

Naiwny klasyfikator Bayesa

- Niech będzie dany przykład X=(x1, x2, ..., xn)
- Oblicz wartość wyrażenia dla każdej klasy

$$F(C_i) = P(C = C_i) \prod_{j=1}^{n} P\left(A_j = x_j \middle| C = C_i\right)$$

- Wybierz klasę dla której $F(C_i)$ przyjmuje największą wartość.

Trening klasyfikatora

- Oszacowanie prawdopodobieństw:
  $P(C = C_i)$ i $P(A_j = x_j | C = C_i)$
- Oznaczenia:
  - $|S|$ oznacza liczbę wszystkich przykładów
  - s_i oznacza liczbę przykładów należących do C_i
  - s_{ij} oznacza liczbę przykładów dla których $A_j = x_j$ i należą do klasy C_i
- $P(C = C_i) \approx s_i / |S|$
- $P(A_j = x_j | C = C_i) \approx s_{ij} / s_i$

Przykład

outlook	temp	humidity	windy	play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

$|S| = 14 \quad s_{yes} = 9 \quad s_{no} = 5$

$P(C = yes) = 9/14 = 0.64$

$P(C = no) = 5/14 = 0.36$

Outlook

$s_{overcast,yes} = 5 \quad P(overcast|yes) = 5/9$

$s_{overcast,no} = 0 \quad P(overcast|no) = 0/5 = 0$

$s_{rainy,yes} = 3 \quad P(rainy|yes) = 3/9$

$s_{rainy,no} = 2 \quad P(rainy|no) = 2/5$

$s_{sunny,yes} = 2 \quad P(sunny|yes) = 2/9$

$s_{sunny,no} = 3 \quad P(sunny|no) = 3/5$

Przykład

outlook	temp	humidity	windy	play
overcast	hot	high	FALSE	yes
overcast	cool	normal	TRUE	yes
overcast	mild	high	TRUE	yes
overcast	hot	normal	FALSE	yes
rainy	mild	high	FALSE	yes
rainy	cool	normal	FALSE	yes
rainy	cool	normal	TRUE	no
rainy	mild	normal	FALSE	yes
rainy	mild	high	TRUE	no
sunny	hot	high	FALSE	no
sunny	hot	high	TRUE	no
sunny	mild	high	FALSE	no
sunny	cool	normal	FALSE	yes
sunny	mild	normal	TRUE	yes

Temperature

$s_{hot,yes} = 2$ $P(hot|yes) = 2/9$
$s_{hot,no} = 2$ $P(hot|no) = 2/5$

$s_{cool,yes} = 3$ $P(cool|yes) = 3/9$
$s_{cool,no} = 1$ $P(cool|no) = 1/5$

$s_{mild,yes} = 3$ $P(mild|yes) = 3/9$
$s_{mild,no} = 2$ $P(mild|no) = 2/5$

Humidity

$s_{high,yes} = 2$ $P(high|yes) = 2/9$
$s_{high,no} = 2$ $P(high|no) = 2/5$

$s_{normal,yes} = 3$ $P(normal|yes) = 3/9$
$s_{normal,no} = 1$ $P(normal|no) = 1/5$

Windy

$s_{true,yes} = 3$ $P(true|yes) = 3/9$
$s_{true,no} = 3$ $P(true|no) = 3/5$

$s_{false,yes} = 6$ $P(false|yes) = 6/9$
$s_{false,no} = 2$ $P(false|no) = 2/5$

Przykład

- Zaklasyfikujmy obiekt:

outlook	temp	humidity	windy
rainy	cool	high	true

- Potrzebne wartości:

$P(C = yes) = 9/14$
$P(C = no) = 5/14$

$P(rainy|yes) = 3/9$
$P(rainy|no) = 2/5$

$P(cool|yes) = 3/9$
$P(cool|no) = 1/5$

$P(high|yes) = 2/9$
$P(high|no) = 2/5$

$P(true|yes) = 3/9$
$P(true|no) = 3/5$

- $F(yes) = \frac{9}{14} * \frac{3}{9} * \frac{3}{9} * \frac{2}{9} * \frac{3}{9} = \frac{6}{1134} = 0.005$
- $F(no) = \frac{5}{14} * \frac{2}{5} * \frac{1}{5} * \frac{2}{5} * \frac{3}{5} = \frac{12}{1750} = 0.006$

Przykład

- Zaklasyfikujmy obiekt:

outlook	temp	humidity	windy
sunny	hot	normal	false

- Potrzebne wartości:

$P(C = yes) = 9/14$ $P(sunny|yes) = 2/9$ $P(hot|yes) = 2/9$
$P(C = no) = 5/14$ $P(sunny|no) = 3/5$ $P(hot|no) = 2/5$

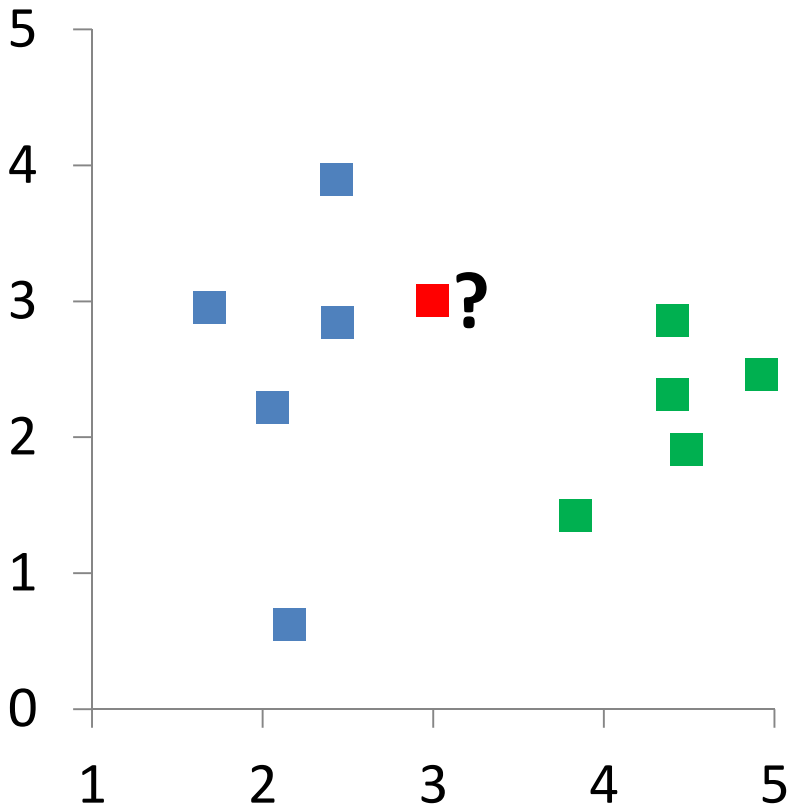
$P(normal|yes) = 3/9$ $P(false|yes) = 6/9$
$P(normal|no) = 1/5$ $P(false|no) = 2/5$

- $F(yes) = \frac{9}{14} * \frac{2}{9} * \frac{2}{9} * \frac{3}{9} * \frac{6}{9} = \frac{8}{1134} = 0.007$
- $F(no) = \frac{5}{14} * \frac{3}{5} * \frac{2}{5} * \frac{1}{5} * \frac{2}{5} = \frac{12}{1750} = 0.006$

Problem „częstości 0”

- W przykładzie, obliczono, że: $P(overcast|no) = 0$
- W praktyce wcale tak być nie musi!
- Problem: jeżeli mamy obiekt X dla którego outlook ma wartość overcast, to wartość $F(no)$ będzie zawsze równa 0.
- W praktyce, jeżeli prawdopodobieństwo apriori wynosi 0, zastępuje się je niewielką, wybraną liczbą.

Klasyfikator najbliższego sąsiedztwa



- Każdy przykład to punkt w wielowymiarowej przestrzeni
- Nauka - zapamiętanie przykładów
- Klasyfikacja – znajdź K najbliższych przykładów i wybierz najczęstszą klasę

Problemy z kNN

- Co to znaczy „najbliższy przykład”?
- Jak dobrać k?
  - Różna liczba przedstawicieli klas
  - Wrażliwość na szumy i punkty odstające
- Jak przetransformować przykład do punktu w przestrzeni wzorców?
- Klątwa wymiarowości
- Wydajność

Miary odległości

- Atrybuty liczbowe
  - Metryka Minkowskiego
  - Metryka Manhattan
  - Metryka Euklidesowa
  - Metryka Chebysheva
- Atrybuty nominalne
  - Współczynnik Jaccarda
  - Odległość Hamminga

Metryka Minkowskiego

$$d(X,Y) = \left(\sum_{i=1}^{n} |x_i - y_i|^h \right)^{1/h}$$

Odległość ze względu na i-ty atrybut $d(x_i, y_i)$

- Specjalne przypadki
  - Manhatan: $h = 1$
  - Euklidesowa: $h = 2$
  - Chebysheva: $h = \infty$

$$\text{cheb}(X,Y) = \max|x_i - y_i|$$

Współczynnik Jaccarda

- Służy do obliczenia podobieństwa pomiędzy obiektami o atrybutach binarnych
- Niech będą dane dwa obiekty X i Y, wówczas
  - $M_{0,0}$ - liczba atrybutów, gdzie X i Y są równe 0
  - $M_{1,1}$ - liczba atrybutów, gdzie X i Y są równe 1
  - $M_{0,1}$ - liczba atrybutów, gdzie X=0 i Y=1
  - $M_{1,0}$ - liczba atrybutów, gdzie X=1 i Y=0
- Wsp. Jaccarda: $\mathrm{J}(\mathrm{X},\mathrm{Y}) = \frac{M_{11}}{M_{0,1}+M_{1,0}+M_{1,1}}$
- Odl. Jaccarda: $d(X,Y) = 1 - J(X,Y)$

Odległość Hamminga

- Służy do obliczenia podobieństwa pomiędzy obiektami o atrybutach nominalnych
- Niech będą dane dwa obiekty X i Y
  - Odległość Hamminga ze względu na i-ty atrybut:

$$d(x_i, y_i) = \begin{cases} 1 & x_i = y_i \\ 0 & x_i \neq y_i \end{cases}$$

  - Odległość Hamminga to liczba atrybutów, które posiadają różne wartości:

$$d(X, Y) = \sum_{i=1}^{n} d(x_i, y_i)$$

Atrybuty porządkowe

- Niech atrybut posiada M wartości
- Każdej wartości przydziel liczbę od 0 do M-1 zgodnie z porządkiem
- Niech będą dane dwa obiekty X i Y
  - Odległość ze względu na i-ty atrybut:

$$d(x_i, y_i) = \frac{|x_i - y_i|}{M - 1}$$

Transformacja do przest. wzorców

- Problem: różne atrybuty mogą posiadać różną skalę, różne jednostki i różne przedziały zmienności
- Bezpośrednie zastosowanie funkcji odległości może spowodować dominację pewnych atrybutów nad pozostałymi (np. atrybuty: temperatura ciała i dochody_roczne) i zafałszowanie wyniku
- Rozwiązanie: nadanie wag atrybutom, normalizacja lub standaryzacja wartości atrybutów

Preprocessing atrybutów liczbowych

- Normalizacja wartości atrybutu A (atrybut ma wartość a):

$$a' = \frac{a - min}{max - min}$$

- Standaryzacja wartości atrybutu A (atrybut ma wartość a):

$$a' = \frac{a - \mu(A)}{\sigma(A)}$$

Obiekty o różnych typach atrybutów

$$d(X,Y) = \left(\sum_{i=1}^{N} \left(d(x_i, y_i) \right)^h \right)^{1/h}$$

- Oblicz odległość $d(x_i, y_i)$ używając odpowiedniego wzoru w zależności od typu atrybutu.
- Podstaw do jakiejś metryki Minkowskiego (np. Euklidesowej)

Przykład

Oid	Miasto	Ocena końcowa	Średnia	Preferencja
1	Klagenfurt	A	4.8	Math
2	Vienna	B	4.2	Math
3	Klagenfurt	D	3.1	Sociology

- Miasto – atrybut nominalny
- Ocena końcowa – atrybut porządkowy, 5 wartości
- Średnia – zmienna ciągła
- Preferencja – atrybut nominalny

Przykład, wstępna transformacja

Oid	Miasto	Ocena końcowa	Średnia	Preferencja
1	Klagenfurt	0	0.93	Math
2	Vienna	1	0.73	Math
3	Klagenfurt	3	0.37	Sociology

- Ocena końcowa (M=5):
  - Mapowanie : A-0,B-1,C-2,D-3,E-4
- Średnia – normalizacja (max=5, min=2)
  - $(4.8 - 2)/(5 - 3) = 2.8/3 \approx 0{,}93$
  - $(4.2 - 2)/(5 - 3) = 2.2/3 \approx 0{,}73$
  - $(3.1 - 2)/(5 - 3) = 1.1/3 \approx 0{,}37$

Przykład, porównanie 1 i 2

- Atrybut City (odl. Hamminga):
  - $d(x_{city}, y_{city}) = 1$, bo Klagenfurt<>Vienna
- Atrybut Final Grade (M=5):
  - $d(x_{fg}, y_{fg}) = \frac{|0-1|}{5-1} = 0.25$
- Atrybut GPA
  - $d(x_{gpa}, y_{gpa}) = |0.93 - 0.73| = 0.2$
- Atrybut Preference
  - $d(x_{pref}, y_{pref}) = 0$, bo Math=Math
- Ostatecznie (odl. Euklidesowa: h=2):
  - $d(X, Y) = \sqrt{1^2 + 0.25^2 + 0.2^2 + 0^2} = \sqrt{1{,}1025} = 1{,}05$

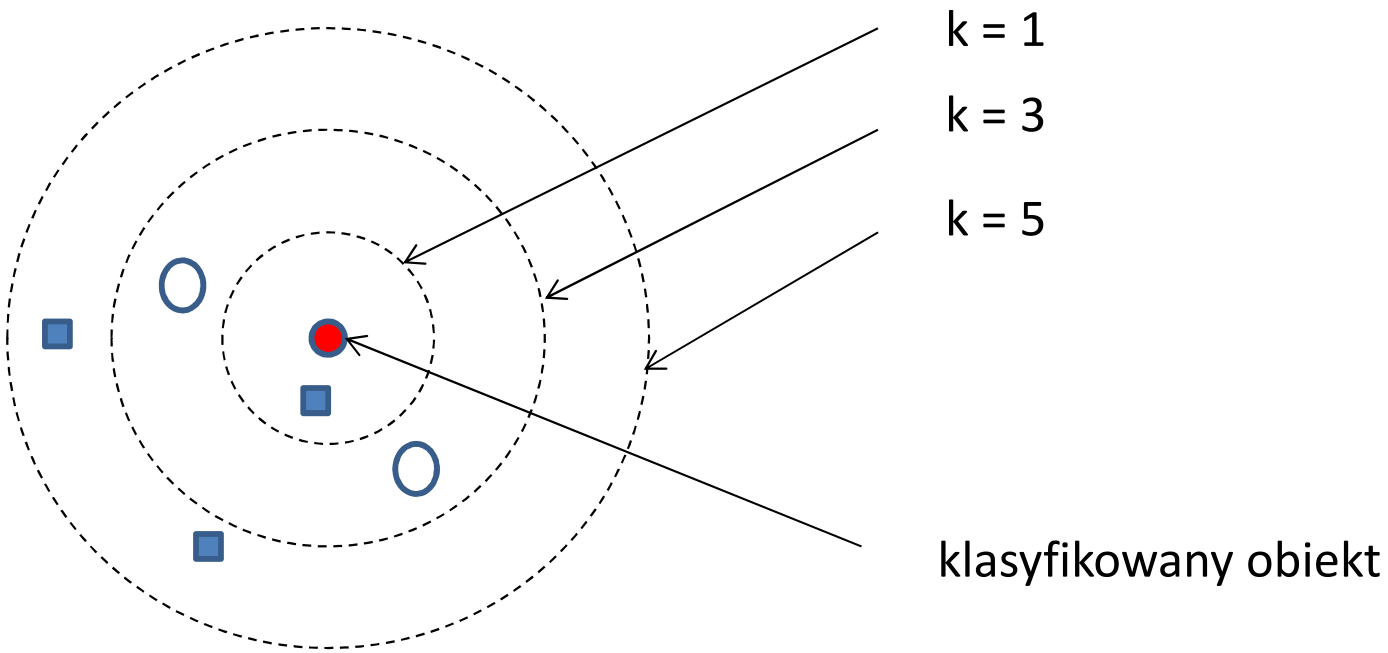
Przykład, porównanie 1 i 3

- Atrybut City (odl. Hamminga):
  - $d(x_{city}, y_{city}) = 0$, bo Klagenfurt=Klagenfurt
- Atrybut Final Grade (M=5):
  - Mapowanie: A-0,B-1,C-2,D-3,E-4
  - $d(x_{fg}, y_{fg}) = d(A, D) = \frac{|0-3|}{5-1} = 0.75$
- Atrybut GPA
  - $d(x_{gpa}, y_{gpa}) = |0.93 - 0.37| = 0.56$
- Atrybut Preference
  - $d(x_{pref}, y_{pref}) = 1$, bo Math<>Sociology
- Ostatecznie (odl. Euklidesowa: h=2):
  - $d(X, Y) = \sqrt{0^2 + 0.75^2 + 0.56^2 + 1^2} = \sqrt{1{,}8761} \approx 1{,}37$

Przykład, porównanie 2 i 3

- Atrybut City (odl. Hamminga):
  - $d(x_{city}, y_{city}) = 1$, bo Vienna<>Klagenfurt
- Atrybut Final Grade (M=5):
  - Mapowanie: A-0,B-1,C-2,D-3,E-4
  - $d(x_{fg}, y_{fg}) = d(B, D) = \frac{|1-3|}{5-1} = 0.5$
- Atrybut GPA
  - $d(x_{gpa}, y_{gpa}) = |0.73 - 0.37| = 0.36$
- Atrybut Preference
  - $d(x_{pref}, y_{pref}) = 1$, bo Math<>Sociology
- Ostatecznie (odl. Euklidesowa: h=2):
  - $d(X, Y) = \sqrt{1^2 + 0.5^2 + 0.36^2 + 1^2} = \sqrt{2{,}3796} \approx 1{,}54$

Wrażliwość k-NN, dobór k



Dobór k

- Określenie klasy decyzyjnej klasyfikowanego obiektu na podstawie listy najbliższych k sąsiadów
  - głosowanie większościowe
  - waga głosu proporcjonalna do odległości od klasyfikowanego obiektu (waga głosu w, $w = 1/d^2$)
- Adaptacyjny dobór wielkości parametru k na podstawie wyników analizy trafności klasyfikacji (lub błędu klasyfikacji)

Klątwa wymiarowości

- Problem danych wysoko wymiarowych (problem tak zwanej klątwy wymiarowości) – obiekty są rzadko rozmieszczone w przestrzeni wzorców
- Wraz z rosnąca liczbą wymiarów odległość obiektu do najbliższego sąsiada zbliża się do odległości do najdalszego sąsiada – stąd dla wysokowymiarowej przestrzeni, o dużej liczbie wymiarów, wszystkie obiekty znajdują się w podobnej odległości
- W tym przypadku poszukiwanie k-najbliższego sąsiedztwa przestaje mieć sens
- Dotyczy to szczególnie miar odległości, dla których h > 1 (np. odległość euklidesowa)

Wydajność

- Klasyfikacja każdego przykładu wymaga znalezienia k najbliższych sąsiadów
- Im więcej wymiarów i więcej przykładów uczących tym dłużej może to trwać.
- Optymalizacje
  - Redukcja wymiarowości
  - Filtrowanie przykładów
  - Indeksy wielowymiarowe

Klasyfikatory regułowe

- Reprezentują wiedzę w postaci zbioru reguł:

  IF <warunek> THEN <konkluzja>

- albo

  <warunek> → <konkluzja>

- Gdzie:
  - konkluzja (nastepnik) określa klas decyzyjną
  - warunek (poprzednik) jest koniunkcją testów zdefiniowanych na atrybutach opisowych
  - Warunek ma postać koniunkcji testów:

(A1 op v1) and (A2 op v2) and … (Ak op vk)

Przykładowe dane

RID	age	status	income	children	risk
1	<=30	bachelor	low	0	high
2	<=30	married	low	1	high
3	<=30	bachelor	high	0	low
4	31..40	bachelor	low	0	high
5	31..40	married	medium	1	low
6	31..40	divorced	high	2	low
7	31..40	divorced	low	2	high
8	31..40	divorced	high	0	low
9	> 40	married	medium	1	low
10	> 40	divorced	medium	4	high
11	> 40	married	medium	2	low
12	> 40	married	medium	1	high
13	> 40	married	high	2	high
14	> 40	divorced	high	3	low

Pokrycie rekordów

- Reguła r pokrywa rekord x (obiekt x) jeżeli wartości atrybutów opisowych rekordu x spełniają poprzednik reguły r
- Przykład:
  - Reguła:

    income=low and children=0 → risk=low

  - Pokrywa rekord:

RID	age	status	income	children	risk
1	<=30	bachelor	low	0	high

  - Ale nie pokrywa rekordu:

RID	age	status	income	children	risk
3	<=30	bachelor	high	0	low

Podstawowe pojęcia

- Zbiór reguł klasyfikacyjnych R ma najczęściej normalną postać dysjunkcyjną:

$$R = (r_1 \vee r_2 \vee \cdots \vee r_k)$$

- (każda reguła jest analizowana przez klasyfikator niezależnie od innych reguł)
- Podstawowe miary oceny jakości reguł klasyfikacyjnych:
  - pokrycie (ang. coverage)
  - trafność (ang. accuracy)

Pokrycie

- Stosunek liczby rekordów zbioru D, które są pokrywane przez regułę r do łącznej liczby rekordów zbioru D
- Dla reguły r: cond → y i zbioru rekordów D

$$pokrycie(r) = |D_{cond}|/|D|$$

- Gdzie
  - D_{cond} - to zbiór rekordów które spełniają poprzednik reguły
  - D – to zbiór wszystkich rekordów

Trafność

- Stosunek liczby rekordów zbioru D, które spełniają zarówno poprzednik, jak i następnik reguły r do liczby rekordów, które spełniają poprzednik reguły r:

$$trafność(r) = \left|D_{cond \cap y}\right| / |D_{cond}|$$

- Gdzie:
  - $D_{cond \cap y}$ to zbiór rekordów, które spełniają poprzednik i następnik reguły
  - D_{cond} to zbiór rekordów które spełniają poprzednik reguły

Jak działa klasyfikator regułowy?

- Przykładowe reguły:

r1: **IF** income = high **and** children = 0 **and** status = bachelor **THEN** risk = low

r2: **IF** income = medium **and** children = 1..2 **THEN** risk = low

r3: **IF** income = low **THEN** risk = high

r4: **IF** income = medium **and** status = divorced **THEN** risk = high

- Jak zostaną zaklasyfikowane rekordy:

RID	age	status	income	children	risk
1	<=30	bachelor	high	0	?
2	31..40	divorced	medium	1	?
3	> 40	bachelor	medium	0	?

  - Rek. 1 pokryty przez regułę r1, czyli risk=low
  - Rek. 2 pokryty przez przez reguły r1 i r4
  - Rek. 3 nie jest pokryty przez żadną regułę

Pożądane własności

- Wzajemne wykluczanie:
  - Żadne dwie reguły nie pokrywają tego samego przykładu.
- Wyczerpujące pokrycie:
  - Każdy rekord jest pokryty przez co najmniej jedną regułę.
- Jeżeli zbiór reguł spełnia te własności, to każdy przykład jest pokryty przez dokładnie jedną regułę.

Problemy

- Co zrobić, jeżeli zbiór reguł nie ma własności wyczerpującego pokrycia?
  - Wprowadzić regułę „domyślną”
  - Reguła domyślna ma pusty poprzednik – jest aplikowana w sytuacji gdy żadna inna reguła nie pokrywa klasyfikowanego przykładu
  - Następnik reguły domyślnej reprezentuje klasę większościową.

Problemy

- Co zrobić, jeżeli zbiór reguł nie ma własności wzajemnego wykluczania?
  - Uporządkowany zbiór reguł
    - Reguły są uporządkowane (priorytet)
    - Pierwsza w kolejności reguła decyduje o klasie
    - Porządek: pokrycie, trafność, kolejność tworzenia
  - Nieuporządkowany zbiór reguł
    - Głosowanie
    - Każda reguła może mieć inną wagę

Algorytmy indukcji reguł decyzyjnych

- Metody pośrednie:
  - ekstrakcja reguł z innych modeli klasyfikacyjnych (np. drzewa decyzyjne, sieci neuronowe, etc).
  - np.: C4.5rules
- Metody bezpośrednie:
  - ekstrakcja reguł bezpośrednio z danych
  - np.: RIPPER, CN2

Pojęcia podstawowe

- Przykład pozytywny
  - Przykład który „potwierdza” regułę – jest pokryty przez jej poprzednik i wartość atrybutu decyzyjnego jest zgodna z następnikiem
- Przykład negatywny
  - Przykład, który nie jest pozytywny

Algorytm sekwencyjnego pokrycia

- Wejście:
  - zbiór przykładów uczących D,
  - zbiór klas C={C_1,C_2,…,C_k} uporządkowany wg. pokrycia

1. $R = \{\}$ //zbiór reguł
2. **for** $C_i \in C \backslash \{C_k\}$ **do**
3. **while not** *warunek stopu* **do**
4. wygeneruj regułę r, o następniku C_i
5. usuń rekordy pokryte przez regułę r
6. dodaj regułę r na końcu zbioru R
6. **end while**
7. **end for**
8. dodaj regułę domyślną () →C_k

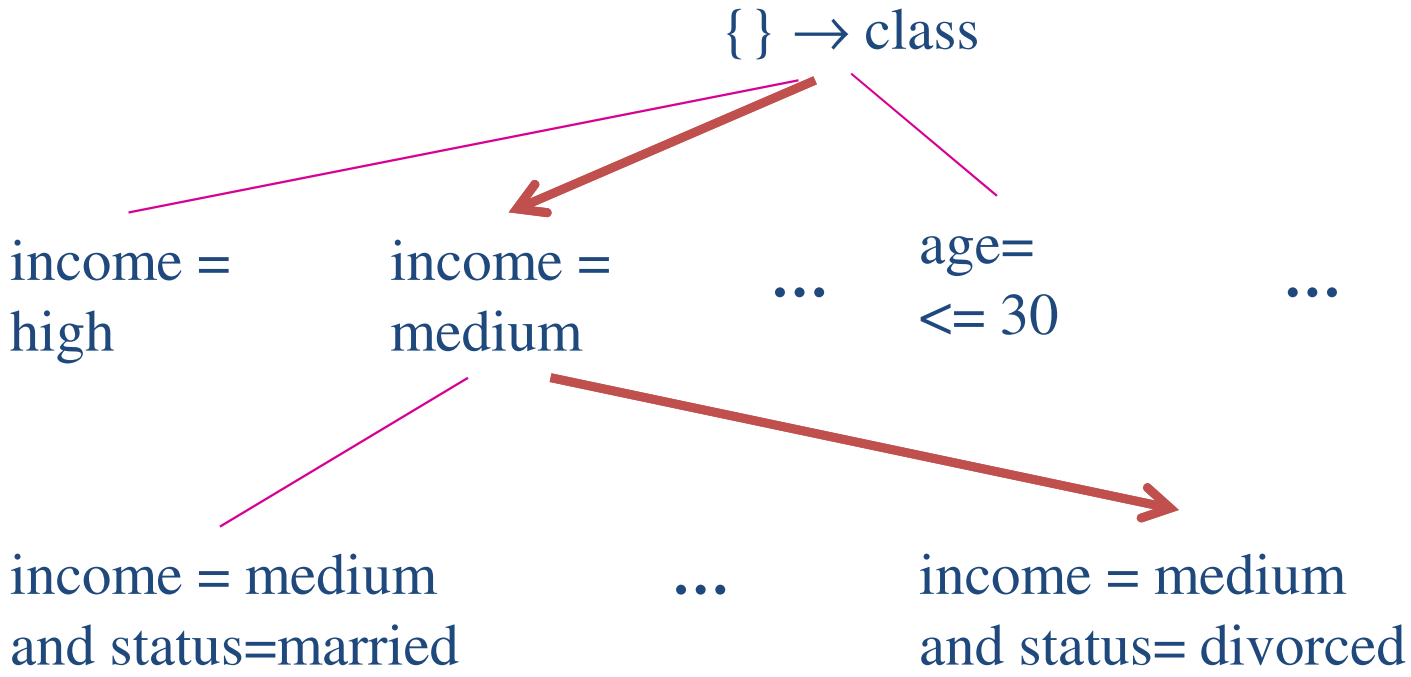
Algorytm sekwencyjnego pokrycia

- Budowany jest uporządkowany zbiór reguł
- Sednem algorytmu jest wiersz 5 – „wygeneruj regułę r”
  - Reguła powinna pokrywać jak najwięcej przykładów o klasie C_i i jak najmniej przykładów z innych klas
  - Znalezienie optymalnej reguły jest kosztowne
  - Stosuje się algorytmy zachłanne
- Reguły danej klasy są generowane dopóki są przykłady z tej klasy bądź trafność generowanych reguł jest powyżej progu.

Konstrukcja reguły

- Od ogółu do szczegółu
  - Zacznij od reguły () → C_i (pokrywa cały zbiór D)
  - Dodawaj nowe testy do tej reguły tak długo jak poprawia to jej trafność, a potem zwróć regułę.
- Od szczegółu do ogółu
  - Utwórz regułę pokrywającą losowy wybrany przykład pozytywny (z klasy C_i)
  - Usuwaj z niej testy tak długo jak poprawia to jej trafność i nie pokrywa ona przykładów negatywnych (nie z klasy C_i).

Od ogółu do szczegółu



Od szczegółu do ogółu

age <= 30 and status = bachelor and
income = low and children = 0 → high

status = bachelor and
income = low and
children = 0 → high

• • •

age <= 30 and
status = bachelor and
income = low → high

Miary oceny reguł

- Zysk informacyjny
- Statystyczny wskaźnik wiarygodności
- Laplace
- m-estimate

Zysk informacyjny

- Obliczany trochę inaczej niż dla ID3
- Dana jest reguła r: $A \rightarrow C_i$
  - Reguła pokrywa *p* przykładów pozytywnych i *n* negatywnych
- Dana jest reguła r*: $A \wedge B \rightarrow C_i$ rozszerzona o test B
  - Reguła pokrywa *p** przykładów pozytywnych i *n** negatywnych
- Zysk informacyjny reguły r* oblicza się:

$$gain(r^*) = p^* \cdot \left(log_2 \left(\frac{p^*}{p^* + n^*} \right) - log_2 \left(\frac{p}{p+n} \right) \right)$$

Statystyczny wskaźnik wiarygodności

$$lrs(r) = 2\sum_{i=1}^{k} f_i log_2\left(\frac{f_i}{e_i}\right)$$

- Gdzie:
  - k – liczba klas decyzyjnych
  - f_i – częstość występowania klasy C_i w przykładach pokrytych przez regułę r
  - e_i – częstość występowania klasy C_i gdy reguła r dokonuje losowej predykcji klasy decyzyjnej.
    - a – liczba przykładów pokrytych przez regułę
    - $e_i = a * \frac{liczba\ przykładów\ klasy\ C_i}{liczba\ wszystkich\ przykładów}$

Przykładowe dane

RID	age	status	income	children	risk
1	<=30	bachelor	low	0	high
2	<=30	married	low	1	high
3	<=30	bachelor	high	0	low
4	31..40	bachelor	low	0	high
5	31..40	married	medium	1	low
6	31..40	divorced	high	2	low
7	31..40	divorced	low	2	high
8	31..40	divorced	high	0	low
9	> 40	married	medium	1	low
10	> 40	divorced	medium	4	high
11	> 40	married	medium	2	low
12	> 40	married	medium	1	high
13	> 40	married	high	2	high
14	> 40	divorced	high	3	low

Przykład

- Dana jest reguła:

  if age=31..40 then risk=low

- Reguła pokrywa p=3 przykłady pozytywne i n=2 przykładów negatywnych.
- O jaki test rozszerzyć tą regułę?
  - Sprawdzimy kilka przykładów, ale nie wyczerpuje to wszystkich możliwości

Przykład (zysk informacyjny)

- Rozszerzenie o test: children=0
  - $p^* = 1, n^* = 1$
  - $gain(r^*) = 1\left(log_2\left(\frac{1}{1+1}\right) - log_2\left(\frac{3}{2+3}\right)\right) \approx -0.26$
- Rozszerzenie o test: status=bachelor
  - $p^* = 0, n^* = 1$
  - $gain(r^*) = 0\left(log_2\left(\frac{0}{0+1}\right) - log_2\left(\frac{3}{2+3}\right)\right) = 0$
- Rozszerzenie o test: income<>low
  - $p^* = 3, n^* = 0$
  - $gain(r^*) = 3\left(log_2\left(\frac{3}{3+0}\right) - log_2\left(\frac{3}{2+3}\right)\right) \approx 2{,}21$

Przykład (stat. wsk. wiarygodności)

- Wartość wskaźnika dla reguły

  if age=31..40 then risk=low

  - $k = 2$
  - $f_{low} = \frac{3}{5}, f_{high} = \frac{2}{5}$
  - $e_{low} = 5 \cdot \frac{7}{14} = 2.5, e_{high} = 5 \cdot \frac{7}{14} = 2.5$
  - $lrs(r) = 2\left(\frac{3}{5} log_2\left(\frac{3/5}{2.5}\right) + \frac{2}{5} log_2\left(\frac{2/5}{2.5}\right)\right) \approx -4.59$

Przykład (stat. wsk. wiarygodności)

- Wartość wskaźnika dla reguły

if age=31..40 and income <>low then risk=low

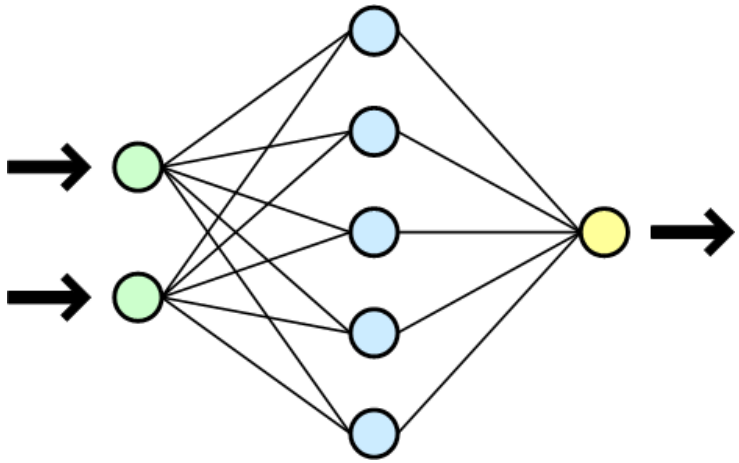
- $k = 2$
- $f_{low} = \frac{3}{3}, f_{high} = \frac{0}{3}$
- $e_{low} = 3 \cdot \frac{7}{14} = 1.5, e_{high} = 3 \cdot \frac{7}{14} = 1.5$
- $lrs(r) = 2\left(\frac{3}{3} log_2\left(\frac{3/3}{1.5}\right) + \frac{0}{3} log_2\left(\frac{0/3}{1.5}\right)\right) \approx 1.17$

Multilayer Perceptron

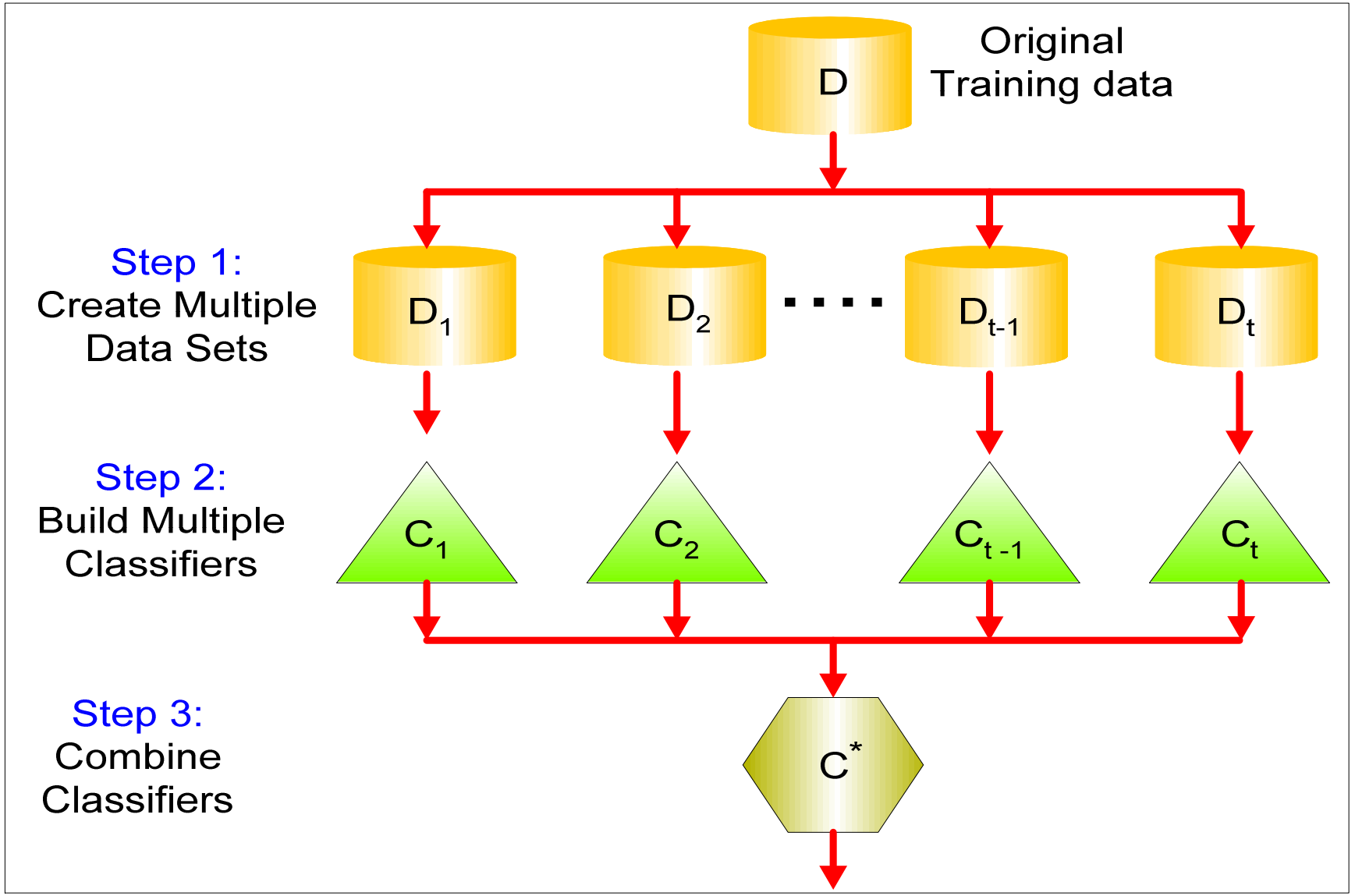
- Klasyfikator oparty o sieć neuronową.
- Składa się z wielu warstw.
  - Wejściowej: tutaj na wejścia neuronów podaje się dane klasyfikowanego przypadku
  - Wyjściowej: wyjścia neuronów tej wartwy zwracają wynik klasyfikacji
  - Warstw pośrednich przekazujących wyniki pośrednie obliczeń pomiędzy sąsiednimi warstwami.

Multilayer Perceptron

- Każdy neuron jako wejście pobiera sumę wyników ze wszystkich neuronów z poprzedniej warstwy przemnożonych przez współczynniki zdefiniowane w synapsach.
- Jeżeli suma ta przekracza zadany próg, to wynikiem neuronu jest 1. W przeciwnym wypadku 0.



Klasyfikatory złożone



Klasyfikacja

- Każdy z klasyfikatorów dokonuje niezależnie predykcji klasy decyzyjnej klasyfikowanego rekordu
- Agregacja predykcji poszczególnych klasyfikatorów w procesie klasyfikacji nowego rekordu i ostateczne określenie klasy decyzyjnej nowego rekordu danych.
- Klasy decyzyjna rekordu jest określana na podstawie wyników głosowania większościowego poszczególnych klasyfikatorów

Dlaczego?

- Rozważmy problem klasyfikacji binarnej:
- Załóżmy, że danych jest 21 identycznych klasyfikatorów, z których każdy charakteryzuje się identycznym błędem klasyfikacji $\varepsilon = 0,3$.
- Predykcje wszystkich klasyfikatorów są identyczne – stąd błąd klasyfikacji kombinacji klasyfikatorów jest identyczny $\varepsilon = 0,3$

Dlaczego?

- Załóżmy, że mamy do czynienia z k niezależnymi klasyfikatorami, których błędy klasyfikacji są nieskorelowane - predykcje klasyfikatorów bazowych mogą się różnić
- Predykcja oparta na głosowaniu większościowym będzie niepoprawna wtedy i tylko wtedy, gdy więcej niż połowa klasyfikatorów bazowych dokona niepoprawnej predykcji klasy danego rekordu

Dlaczego?

- W naszym przypadku (21 niezależnych klasyfikatorów, każdy popełnia błąd z prawdopodobieństwem 0.3) błąd wynosi:

$$\varepsilon = \sum_{i=11}^{21} \binom{21}{i} \varepsilon^i (1-\varepsilon)^{21-i} = 0.026$$

Wniosek

- Klasyfikator złożony może mieć znacznie mniejszy błąd niż klasyfikatory składowe
- Aby tak było muszą być spełnione 2 warunki:
  - trafność klasyfikacji klasyfikatorów bazowych musi być lepsza od trafności klasyfikatora losowego, i
  - klasyfikatory bazowe powinny być niezależne

Jak utworzyć klasyfikatory składowe?

- Modyfikacja zbioru danych treningowych (np. bagging, boosting)
- Modyfikacja zbioru atrybutów danych treningowych (metoda lasów losowych)
- Wartości klas danych treningowych (ECOC)

Ocena jakości klasyfikacji

- Zbiór danych dzielony jest na dwa podzbiory
  - Zbiór uczący
  - Zbiór testujący
- Klasyfikator budowany jest na podstawie zbioru uczącego.
- Zbudowany klasyfikator testowany jest na przykładach ze zbioru testującego.
- Wynikiem takiego testu jest macierz pomyłek.
- Na podstawie wartości z macierzy pomyłek oblicza się różne miary.

Macierz pomyłek

Faktyczna klasa:	Zaklasyfikowano jako: NIESTRAWNOŚĆ	ALERGIA	PRZECHŁODZENIE	CZARNA OSPA
NIESTRAWNOŚĆ	334	99	1	20
ALERGIA	13	234	23	100
PRZECHŁODZENIE	5	21	315	91
CZARNA OSPA	12	10	25	153

Miary jakości klasyfikacji

Alergia:

Zaklasyfikowano jako:

Faktyczna klasa:

	ALERGIA	NIE ALERGIA
ALERGIA	TP 234	FN 136
NIE ALERGIA	FP 130	TN 802

$$precision = \frac{TP}{TP+FP}$$

$$tp-rate = recall$$

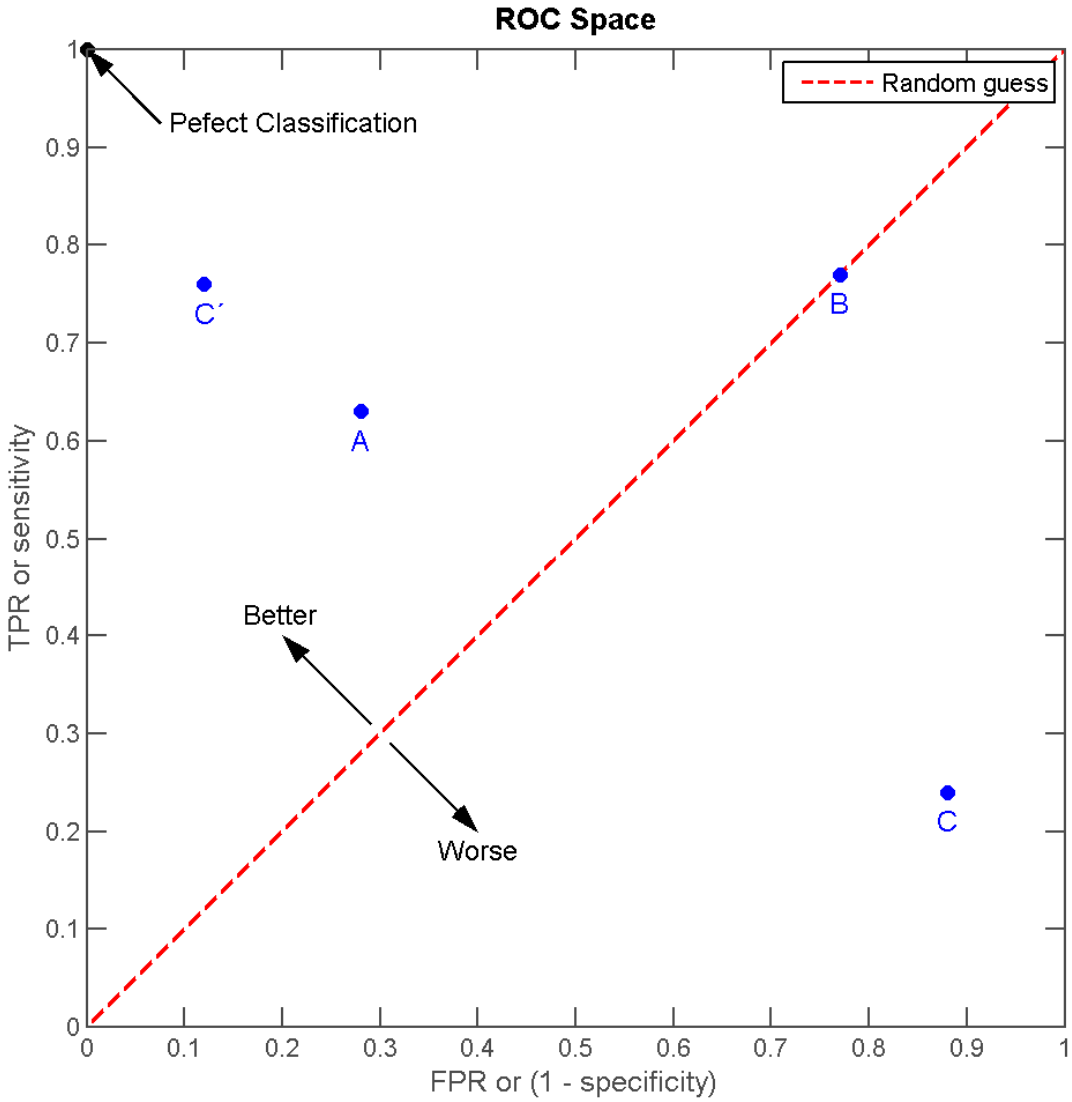
$$recall = \frac{TP}{TP+FN}$$ (sensitivity)

$$fp-rate = \frac{FP}{FP+TN}$$

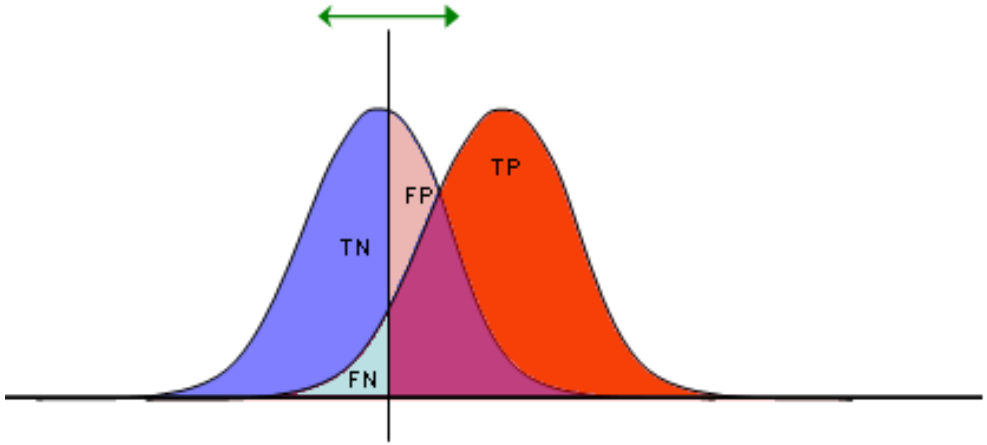
$$F-measure = \frac{2*precision*recall}{precision+recall}$$

$$specificity = \frac{TN}{TN+FP}$$

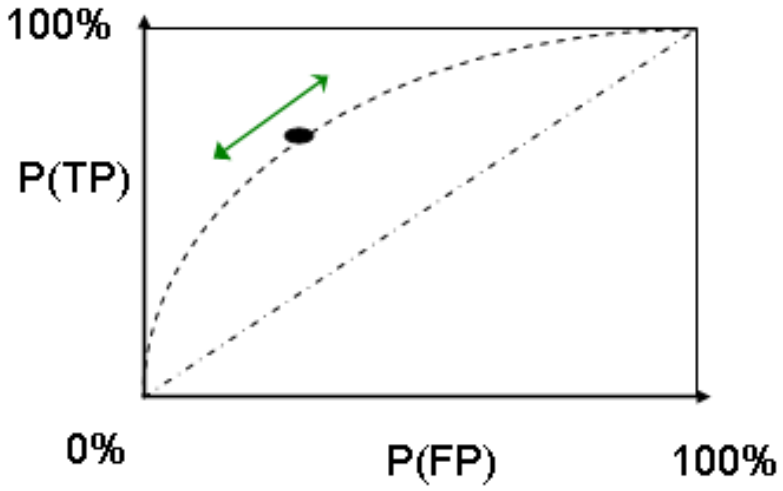
Przestrzeń ROC



Krzywa ROC



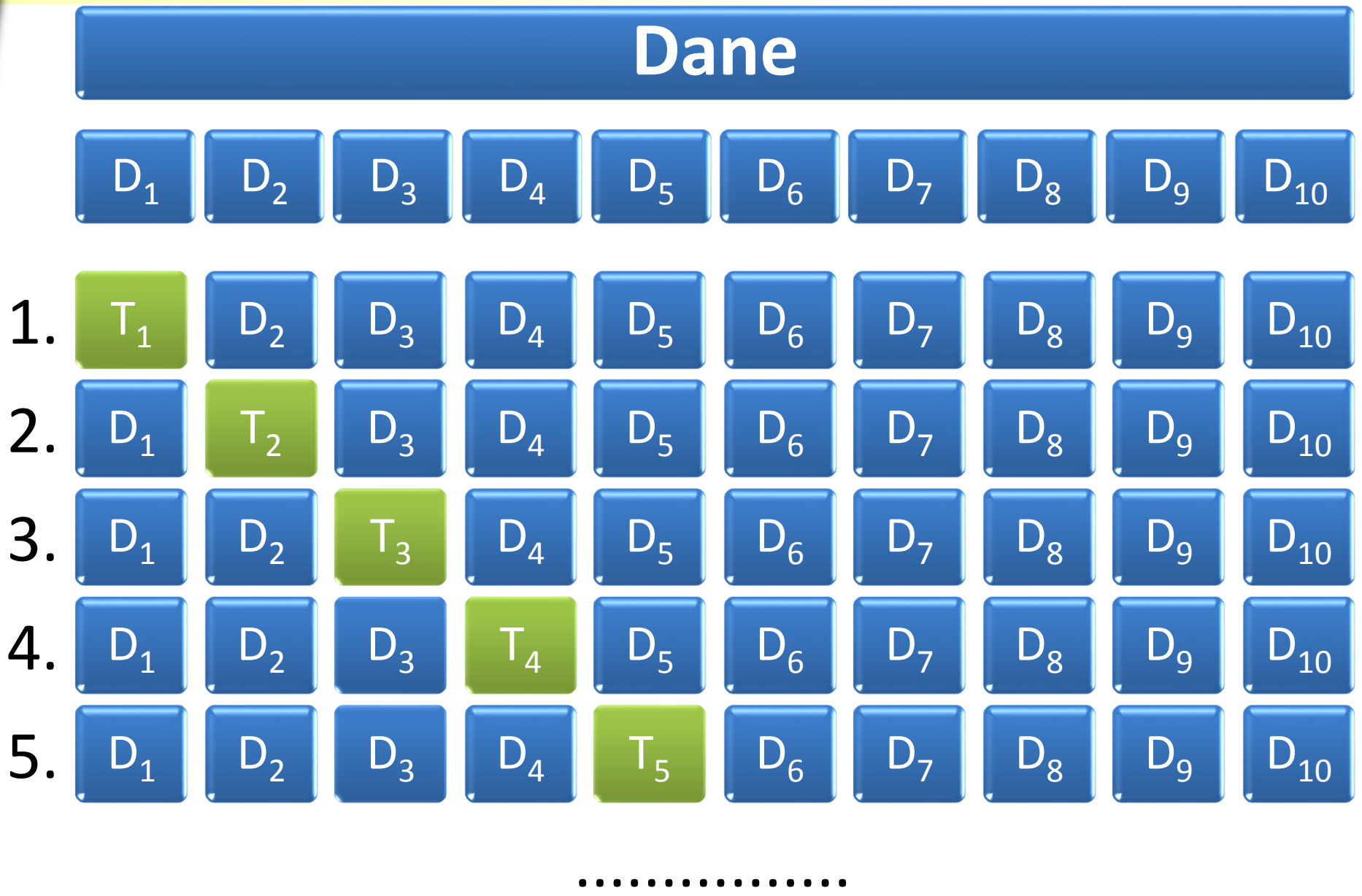
TP	FP
FN	TN
1	1



Jak podzielić zbiór danych?

- Często 2/3 zbiór uczący, 1/3 zbiór testujący
- Uwaga! Istotne rekordy pozwalające na budowę dobrego klasyfikatora mogą zostać usunięte ze zbioru uczącego.
- Aby zatem uzyskać ocenę algorytmu klasyfikacji i jego parametrów, do danej dziedziny zastosowań stosujemy tzw. walidację skrośną albo n-fold crossvalidation.

Walidacja skrośna (1)



Walidacja skrośna (2)

- Cały zbiór danych, który posiadamy dzielony jest na n części (najczęściej 10). Następnie poniższa procedura wykonywana jest n razy:
  1. Weź k-tą część i uznaj ją za zbiór testujący
  2. Weź pozostałe części i uznaj je jako zbiór uczący
  3. Zbuduj klasyfikator na podstawie zbioru uczącego
  4. Sprawdź działanie klasyfikatora na zbiorze testującym
  5. Wyniki testowania umieść w macierzy pomyłek (zwiększ odpowiednie wartości z poprzedniej iteracji).
- Po przeprowadzeniu testów wyliczane są pozostałe miary jakości klasyfikacji na podstawie macierzy pomyłek.

Leave one out

- Specjalny przypadek crosswalidacji
- Zbiór testowy składa się tylko z jednego przypadku.
- Odpowiada crosswalidacji dla k=liczba przykładow.
- Stosowany w sytuacjach, kiedy zbiór danych jest mały.