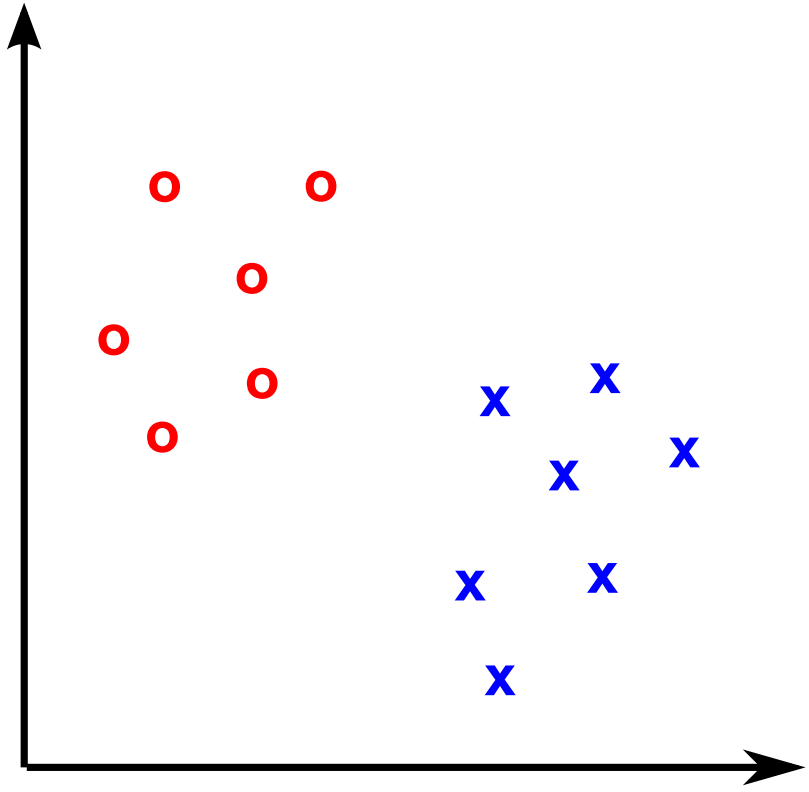
Metody klasyfikacji danych - część 1

Jerzy Dembski

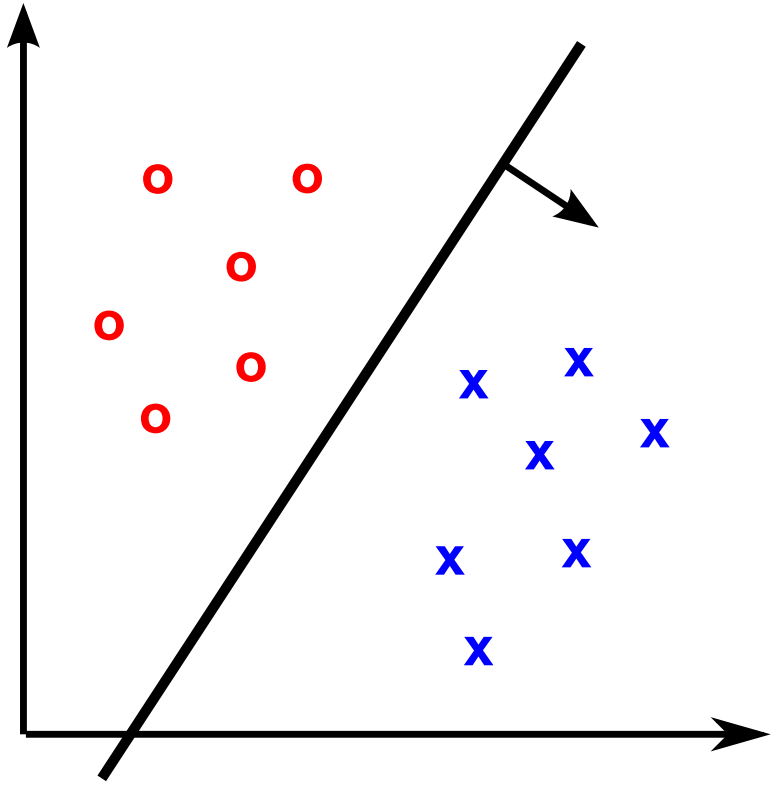
Plan wykładu

- Zadanie klasyfikacji danych
- Przegląd problemów klasyfikacji
- Przegląd metod klasyfikacji
- Metody uczenia klasyfikatorów
- Uczenie a uogólnianie
- Selekcja cech obiektów

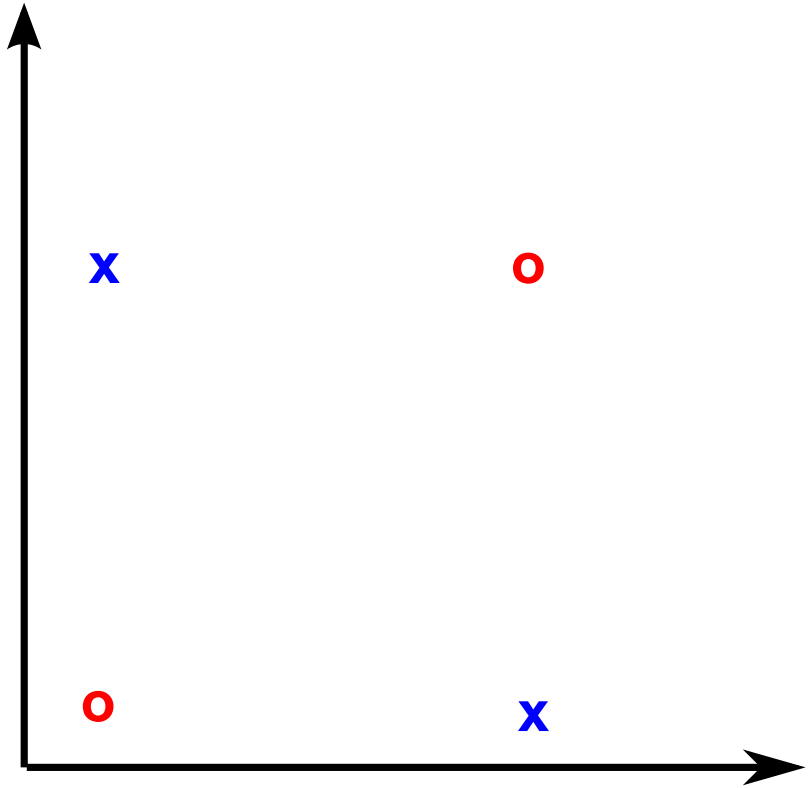
Klasyfikacja punktów na płaszczyźnie



Klasyfikacja punktów na płaszczyźnie - granica decyzji

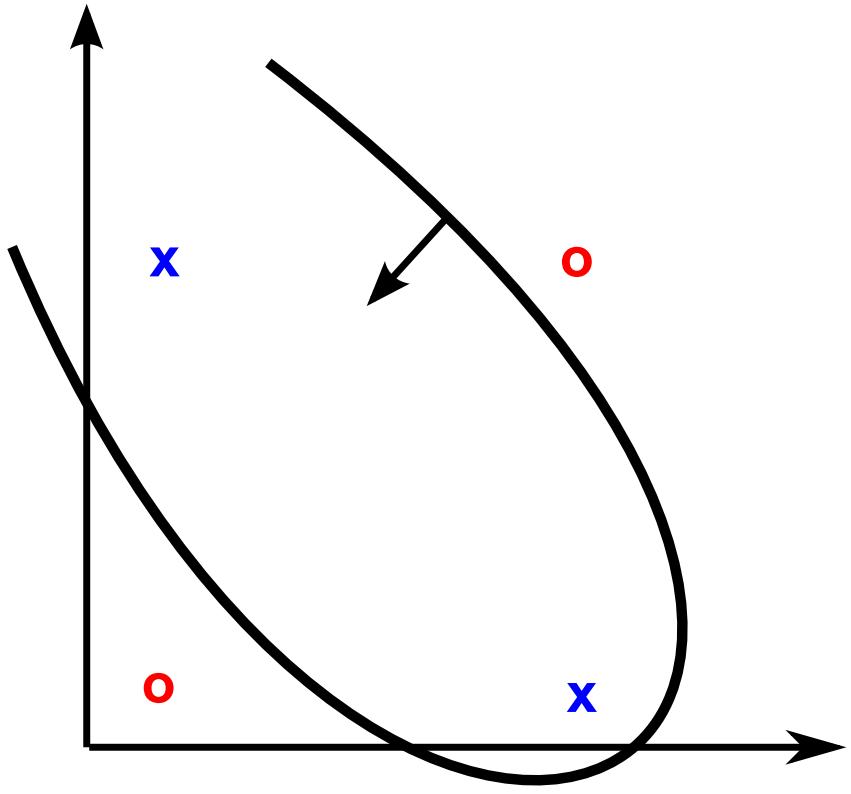


Klasyfikacja punktów na płaszczyźnie - problem XOR



x_1	x_2	klasa
0	0	0
0	1	1
1	0	1
1	1	0

Klasyfikacja punktów na płaszczyźnie - problem XOR



Przegląd problemów klasyfikacji

- Klasyfikacja punktów na płaszczyźnie dwuwymiarowej
- Klasyfikacja rekordów bazodanowych o wartościach dyskretnych (przypisywanie soczewek kontaktowych (*lenses*), klasyfikacja samochodów (*car*)
- Klasyfikacja tekstów (treść, opinia, autorstwo)
- Detekcja i rozpoznawanie obrazów graficznych (twarze, znaki pisane, znaki drogowe, itd.)
- Klasyfikacja dźwięków (rozpoznawanie utworów muzycznych, rozpoznawanie mowy)

Podstawowe pojęcia

Funkcja klasyfikacji $f(\mathbf{x})$ przypisuje numer klasy każdemu wektorowi obserwacji $\mathbf{x}$. Funkcję $f(\mathbf{x})$ można uzyskać drogą heurystyczną, wykorzystując wiedzę eksperta lub metodą uczenia klasyfikatora na podstawie zbioru przykładów uczących.

Klasyfikatorem jest algorytm przypisujący numer klasy każdemu wektorowi obserwacji $\mathbf{x}$, reprezentujący funkcję $f(\mathbf{x})$.

Zadanie klasyfikacyjne polega na szukaniu funkcji klasyfikacji na podstawie wektorów obserwacji. Elementami zbioru przykładów $\mathbf{z}$, są pary złożone z wektora obserwacji charakteryzującego obiekt (obrazu, wektora cech, wektora atrybutów) $\mathbf{x}_i$ oraz klasy (kategorii) do której dany obiekt jest przypisany d_i:

$$\mathbf{z} = \{(\mathbf{x}_1, d_1), (\mathbf{x}_2, d_2) \ldots (\mathbf{x}_K, d_K)\}, \quad \mathbf{x}_i \in X \subset \mathbb{R}^N, \quad d_i \in \{C_1, C_2, \ldots, C_{L_C}\},$$

gdzie $C_1, C_2, \ldots, C_{L_C}$ są etykietami klas, L_C jest liczbą klas obiektów.

Poprawność klasyfikacji przykładów testowych dla danego klasyfikatora jest ilorazem liczby obrazów testowych poprawnie sklasyfikowanych przez dany klasyfikator i liczby wszystkich obrazów testowych.

Metody testowania klasyfikatorów

- „pół na pół” - połowa przykładów służy do uczenia klasyfikatora, połowa do testu ($K_u = K/2$),
- *cross validation* – najpierw pierwsza połowa do uczenia, druga do testu, a następnie na odwrót, wyniki uśrednione dla obu prób,
- *10 fold validation* – cały zbiór przykładów dzielony jest na 10 mniej więcej rónolicznych podzbiorów. W każdej próbie jeden z podzbiorów służy do testu a pozostałych 9 do uczenia ($K_u = 0,9K$). Wyniki uśredniane są po dziesięciu próbach,
- *leave-one-out* – uczenie klasyfikatora odbywa się na podstawie $K_u = K - 1$ przykładów natomiast testowanie na jednym przykładzie. Wyniki uśredniane są po K epizodach (dla każdego przykładu do testu).

Przegląd metod klasyfikacji

- Klasyfikator większościowy (trywialny)
- Klasyfikator najbliższego sąsiada (NS)
- Klasyfikator najbliższego centrum (NC)
- Klasyfikator bayesowski (KB)
- Wektory dyskryminacyjne Fishera (WDF)
- Drzewa decyzyjne (DD)
- AdaBoost (AB)
- Metoda wektorów wspierających (SVM)
- Sztuczne sieci neuronowe (SNN)
- Głębokie uczenie - *Deep Learning* (DL)

Drzewa decyzyjne - problem doboru soczewek

		indeks wartości zmiennej		
	opis zmiennej/klasy	1	2	3
x_1	wiek pacjenta	młody	pre-presbyopia	presbyopia
x_2	rodzaj wady	bliskowzroczność	dalekowzroczność	$\times$
x_3	astygmatyzm	nie	tak	$\times$
x_4	produkcja łez	zredukowana	w normie	$\times$
klasa	typ soczewek	brak	miękkie	twarde

Drzewa decyzyjne - problem doboru soczewek

Lp	x_1	x_2	x_3	x_4	klasa
1	1	1	2	1	1
2	1	1	2	2	3
3	1	2	1	1	1
4	1	2	1	2	2
5	2	1	1	1	1
6	2	1	2	1	1
7	2	2	1	1	1
8	2	2	1	2	2
9	2	2	2	2	1
10	3	1	1	1	1
11	3	2	1	1	1
12	3	2	2	1	1

Drzewa decyzyjne - algorytm tworzący węzeł

function twórz_węzeł(*lista_zmiennych*, *zbiór_przykładów*)
if wszystkie przykłady ze zbioru są tej samej klasy C **then**
return *liść_klasy_*C**;**
else
znajdź zmienną x_m o maksymalnym zysku informacyjnym**;**
for $i = 1$ **to** *liczba_wartości_zmiennej_*x_m
podzbiór_przykl $\leftarrow$ wybierz przykłady, dla których $x_m = i$**;**
lista_zmiennych' $\leftarrow$ *lista_zmiennych* - $\{x_m\}$**;**
drzewo.gałąź_i $\leftarrow$ twórz_węzeł(*lista_zmiennych'*, *podzbiór_przykl*)**;**
endfor
return *drzewo*;
endif

Drzewa decyzyjne - początkowa entropia

Początkowa entropia zbioru przykładów $\mathbf{z}$ jest obliczana jako entropia rozkładu klas obiektów:

$$E(\mathbf{z}) = -\sum_{i=1}^{L_C} P(d = C_i) \log_2 P(d = C_i),$$

gdzie d jest numerem klasy przykładu. W przypadku nieznajomości prawdopodobieństw przynależności obrazu do każdej z klas – $P(d = C_i)$, gdzie C_i – numer i-tej klasy, prawdopodobieństwa można zastąpić częstościami:

$$E(\mathbf{z}) = -\sum_{i=1}^{L_C} \left(\frac{K_i}{K}\right) \log_2 \left(\frac{K_i}{K}\right),$$

Dla danych do klasyfikacji soczewek entropia początkowa wynosi:

$$E(\mathbf{z}) = -\frac{9}{12} \log_2 \left(\frac{9}{12}\right) - \frac{2}{12} \log_2 \left(\frac{2}{12}\right) - \frac{1}{12} \log_2 \left(\frac{1}{12}\right) = 1,0409.$$

Drzewa decyzyjne - zysk informacyjny

Różnica pomiędzy entropią rozkładu klas a średnią entropią rozkładu klas dla podzbiorów przykładów ważoną względem prawdopodobieństwa wystąpienia poszczególnych wartości zmiennej x_i (liczności podzbiorów) stanowi tzw. zysk informacyjny związany z wyborem zmiennej x_i do podziału zbioru przykładów:

$$Zysk(\mathbf{z}, x_i) = E(\mathbf{z}) - \sum_{v} \frac{K(x_i = v)}{K} E(\mathbf{z}|x_i = v),$$

gdzie $K(x_i = v)$ jest liczbą przykładów, dla których zmienna x_i przyjmuje wartość v, natomiast $E(\mathbf{z}|x_i = v)$ jest entropią rozkładu klas dla podzbioru przykładów ze zbioru $\mathbf{z}$, dla których wartość zmiennej x_i wynosi v.

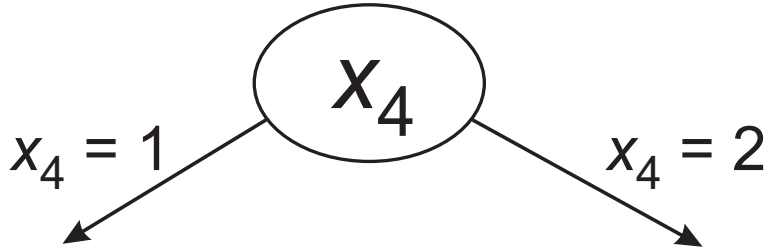
Drzewa decyzyjne - zysk informacyjny

Dla danych do klasyfikacji soczewek zysk informacyjny dla zmiennej x_1 oblicza się w następujący sposób:

$$
\begin{aligned}
Zysk(\mathbf{z}, x_1) &= E(\mathbf{z}) - \sum_v \frac{K(x_1 = v)}{K} E(\mathbf{z}|x_1 = v) = \\
&= E(\mathbf{z}) - \frac{K(x_1 = 1)}{K} E(\mathbf{z}|x_1 = 1) - \frac{K(x_1 = 2)}{K} E(\mathbf{z}|x_1 = 2) - \\
&\quad \frac{K(x_1 = 3)}{K} E(\mathbf{z}|x_1 = 3) = \\
&= 1,0409 - \frac{4}{12}\left[-\frac{2}{4}\log_2\left(\frac{2}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right) - \frac{1}{4}\log_2\left(\frac{1}{4}\right)\right] - \\
&\quad \frac{5}{12}\left[-\frac{4}{5}\log_2\left(\frac{4}{5}\right) - \frac{1}{5}\log_2\left(\frac{1}{5}\right)\right] - \frac{3}{12}\left[-\frac{3}{3}\log_2\left(\frac{3}{3}\right)\right] = \\
&= 0,2401.
\end{aligned}
$$

Dla zmiennych x_2, x_3 i x_4 wynosi odpowiednio $0,2366$; $0,2366$ oraz $0,5409$.

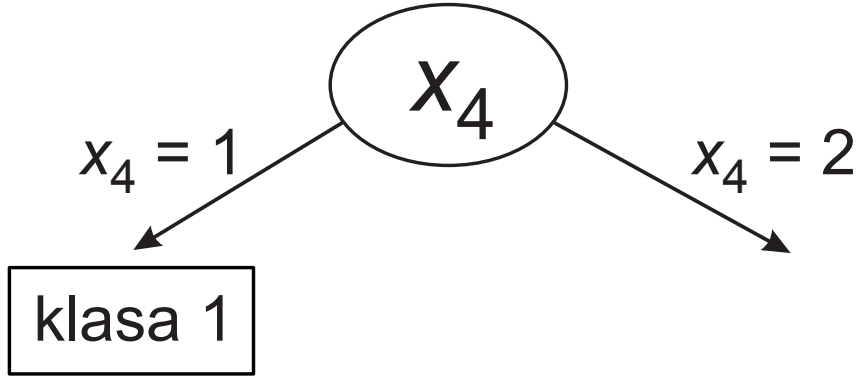
Drzewa decyzyjne - podział zbioru przykładów



Lp	x_1	x_2	x_3	klasa
1	1	1	2	1
3	1	2	1	1
5	2	1	1	1
6	2	1	2	1
7	2	2	1	1
10	3	1	1	1
11	3	2	1	1
12	3	2	2	1

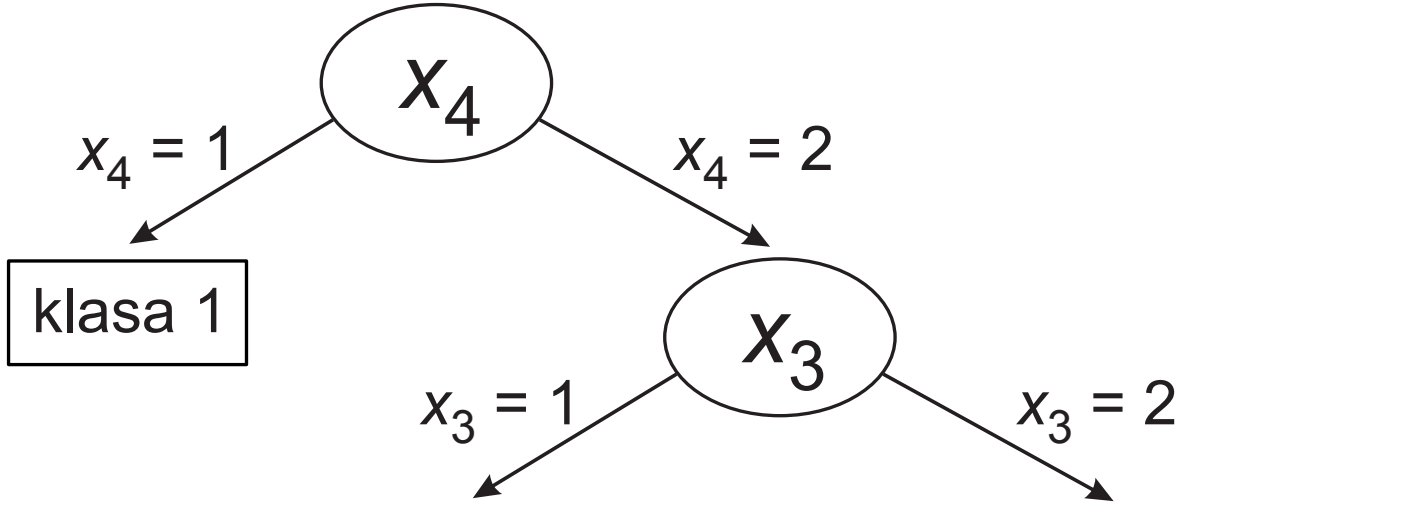
Lp	x_1	x_2	x_3	klasa
2	1	1	2	3
4	1	2	1	2
8	2	2	1	2
9	2	2	2	1

Drzewa decyzyjne - podział zbioru przykładów



Lp	x_1	x_2	x_3	klasa
2	1	1	2	3
4	1	2	1	2
8	2	2	1	2
9	2	2	2	1

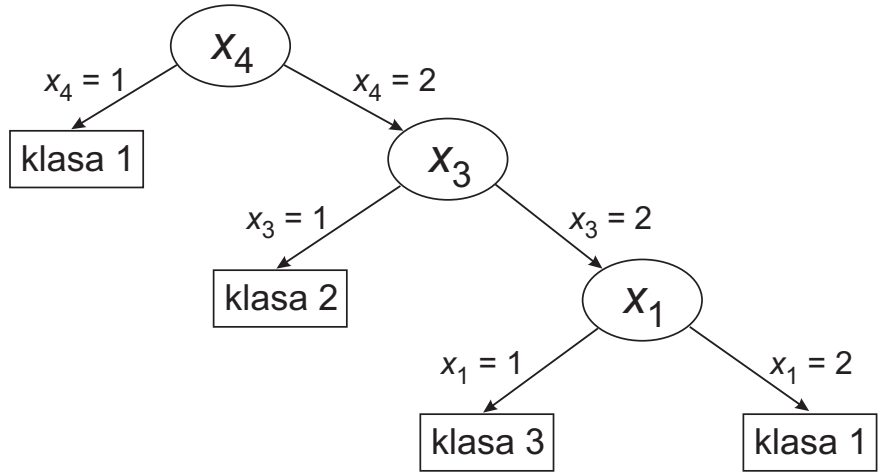
Drzewa decyzyjne - podział zbioru przykładów



Lp	x_1	x_2	klasa
4	1	2	2
8	2	2	2

Lp	x_1	x_2	klasa
2	1	1	3
9	2	2	1

Kompletne drzewo oraz zestaw reguł



jeśli produkcja łez jest zredukowana **to** soczewki nie są polecane
jeśli produkcja łez jest w normie **i** brak astygmatyzmu
to soczewki miękkie
jeśli produkcja łez jest w normie **i** jest astygmatyzm
i wiek młody **to** soczewki twarde
jeśli produkcja łez jest w normie **i** jest astygmatyzm
i wiek pre-presbyopia **to** soczewki nie są polecane

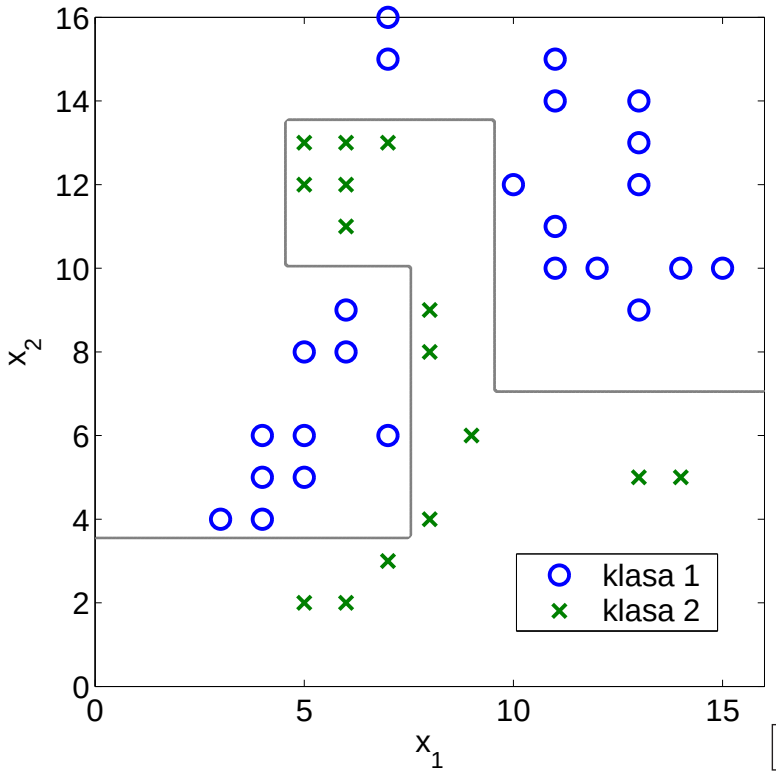
Ciągłe wartości zmiennych - metody postępowania

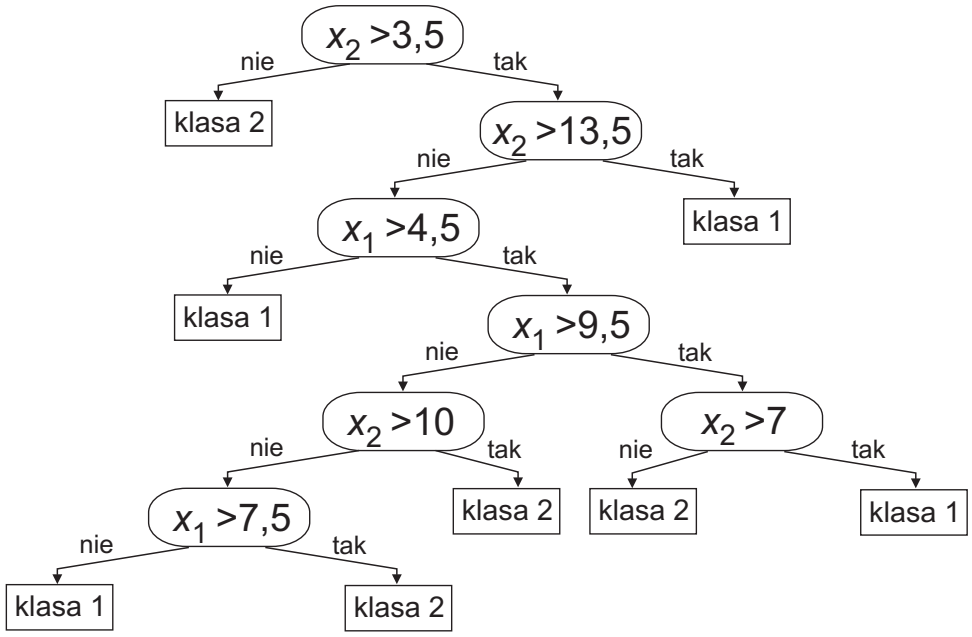
- Dyskretyzacja zmiennych – podział dziedziny wartości każdej ze zmiennych ciągłych na pewną liczbę podprzedziałów tak, by można było zastosować dowolny algorytm budowy dla wartości dyskretnych np. ID3,
- Zastosowanie specjalizowanych algorytmów budowy drzew np. CART.

CART: Tworzenie węzła w przypadku ciągłych wartości zmiennych

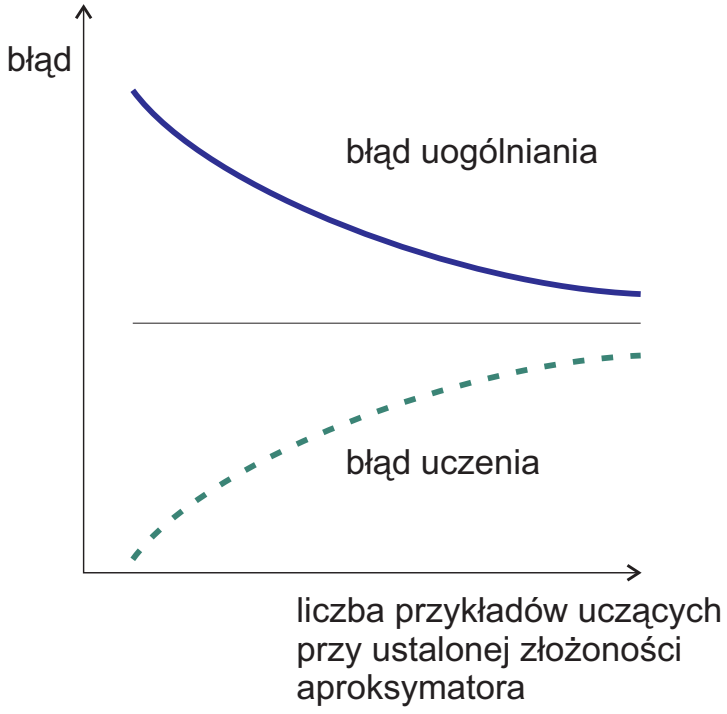
function twórz_węzeł(*zbiór_przykładów*)
  if wszystkie przykłady ze zbioru są tej samej klasy C **then**
    return *liść_klasy_*C**;**
  else
    for $i = 1$ **to** *liczba zmiennych*
      sortuj przykłady względem wartości zmiennej x_i**;**
      ustal wartości graniczne $gr_1, gr_2, \ldots, gr_{K-1}$ pomiędzy każdą parą sąsiednich przykładów na liście posortowanej, gdzie K jest liczbą przykładów wchodzących do węzła**;**
    endfor
    znajdź zmienną x_m wraz z granicą podziału gr_k o maksymalnym zysku informacyjnym**;**
    podzbiór_przykl_1 $\leftarrow$ wybierz przykłady, dla których $x_m \leq gr_k$**;**
    drzewo.lewa_gałąź $\leftarrow$ twórz_węzeł(*podzbiór_przykl_1*)**;**
    podzbiór_przykl_2 $\leftarrow$ wybierz przykłady, dla których $x_m > gr_k$**;**
    drzewo.prawa_gałąź $\leftarrow$ twórz_węzeł(*podzbiór_przykl_2*)**;**
    return *drzewo*;
  endif

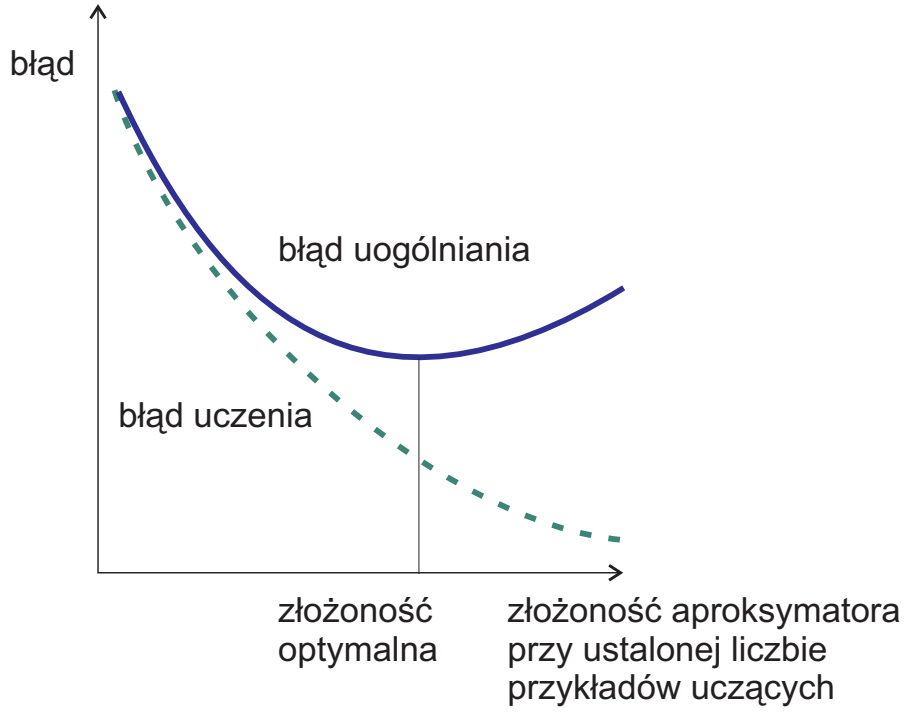
Przykładowe drzewo dla problemu o ciągłych wartościach zmiennych



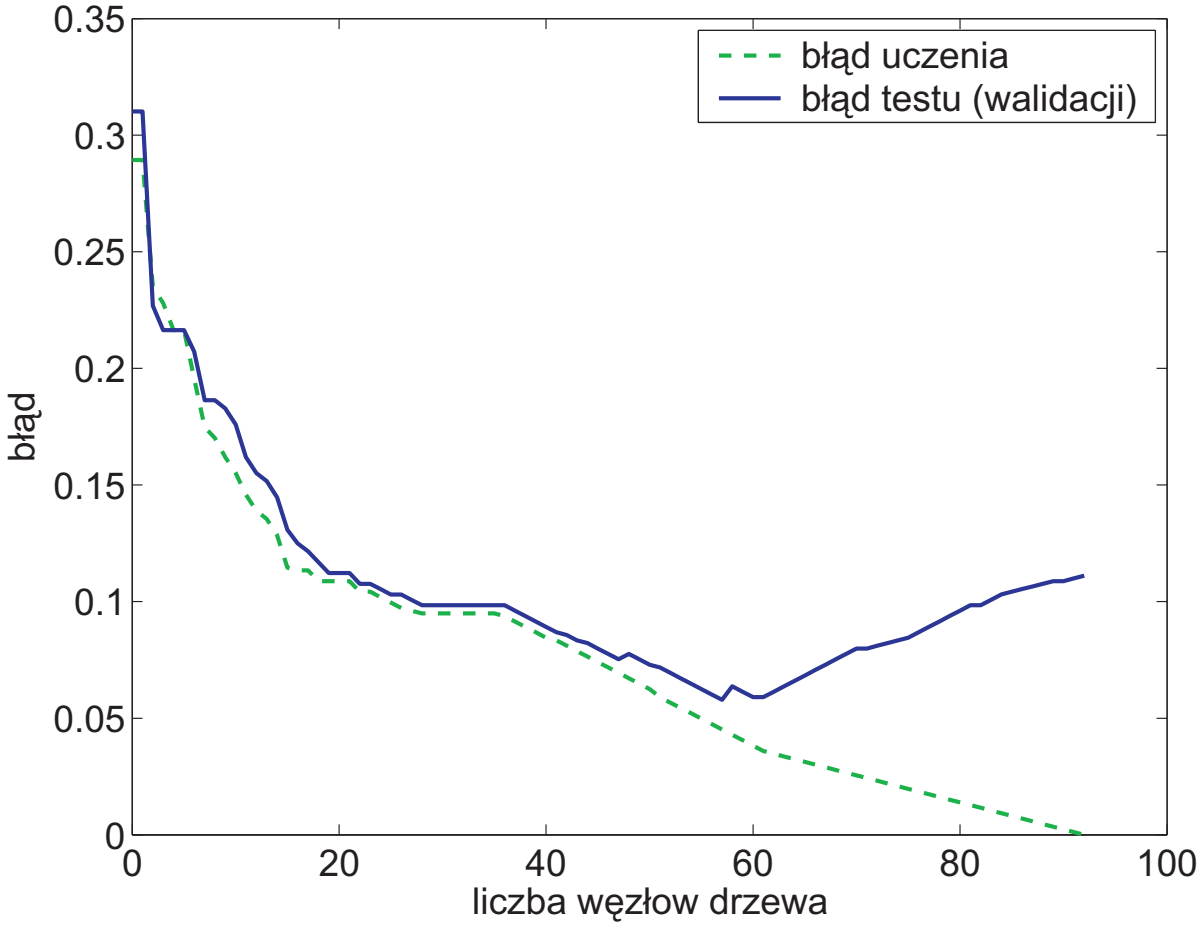


Uczenie a uogólnianie (generalizacja)





Wykres błędu klasyfikacji dla problemu *car*



Drzewa decyzyjne: metody poprawy uogólniania

- Przycinanie heurystyczne np. na stałą głębokość lub w zależności od liczby przykładów dochodzących do węzła nieterminalnego. Gdy liczba ta jest poniżej progu, węzeł jest zamieniany na liść klasy najliczniej reprezentowanej przez przykłady, które do niego dochodzą.
- Przycinanie w oparciu o zbiór przykładów do walidacji: węzeł jest zamieniany na liść lub na jeden z węzłów potomnych, gdy pozwala zmniejszyć błąd walidacji. Wadą tego rozwiązania jest konieczność zmniejszenia zbioru przykładów do budowy drzewa.
- Komitet klasyfikatorów - drzew zbudowanych na losowych podzbiorach przykładów do uczenia (*random forests*).