Metody eksploracji danych
5. Klasyfikacja (kontynuacja)

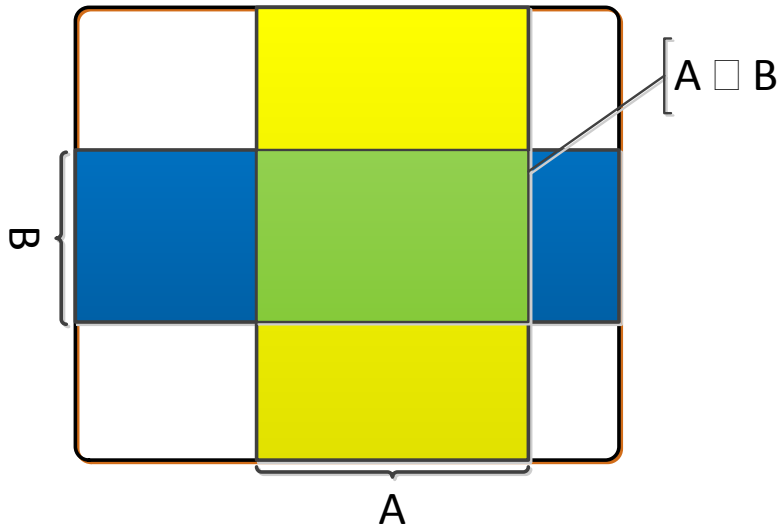
Piotr Szwed
Katedra Informatyki Stosowanej AGH
2016

Naiwny model Bayesa
Drzewa decyzyjne

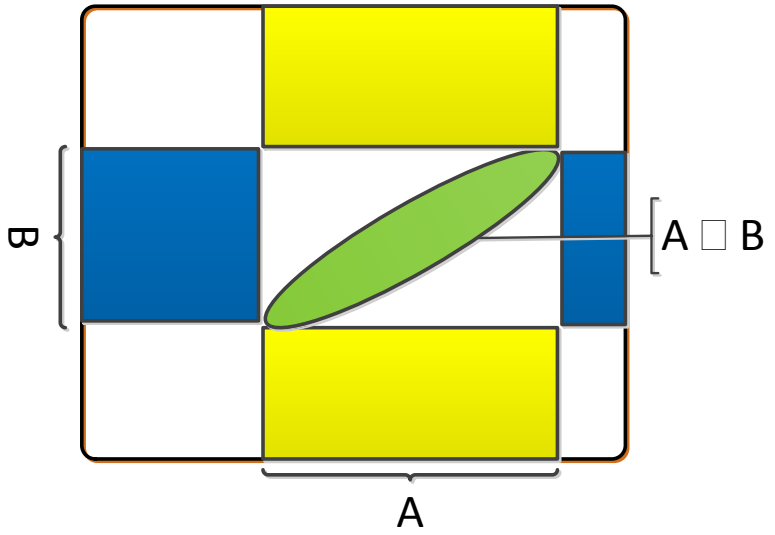
Naiwny model Bayesa

Prawdopodobieństwo warunkowe

- $P(A \wedge B)$ - prawdopodobieństwo wspólne wystąpienia A i B
- $P(A|B) = \frac{P(A \wedge B)}{P(B)}$ - prawdopodobieństwo warunkowe: jeśli zaszło B, to prawdopodobieństwo, że również zaszło A wynosi $P(A|B)$
- Jeżeli A i B są niezależne, to $P(A|B) = P(A)$ oraz $P(B|A) = P(B)$
- **Jeżeli A i B są niezależne, to $P(A \wedge B) = P(A) \cdot P(B)$**



A i B niezależne, $P(A) = 1/2$, $P(B) = 1/3$, $P(A \wedge \mathrm{B}) = P(A) \cdot P(B) = 1/6$ (zielony kwadrat). $P(A|B) = 1/2$



A i B zależne, $P(A|B) = elipsa/(2\ niebieskie\ prostokąty + elipsa)$

Wnioskowanie bayesowskie

- **Twierdzenie Bayesa**: Z definicji: $P(A \wedge B) = P(A|B)P(B) = P(B|A)P(A)$, stąd:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

- Zamiast A i B użyjemy zmiennych losowych X, Y. Zakładamy, że Y przybiera wartości dyskretne: $y_1, \ldots, y_K$

$$P(Y = y_k|X = x) = \frac{P(X = x|Y = y_k)P(Y = y_k)}{P(X = x)}$$

  - $P(Y = y_k)$ – prawdopodobieństwo **a priori** (ang. priors)
  - $P(X = x|Y = y_k)$ - **prawdopodobieństwo wystąpienia**, wiarygodność (ang. likelihood)
  - $P(Y = y_k|X = x)$ - prawdopodobieństwo **a posteriori** (ang. posteriors)

$$posterior\ probabilty \propto likelihood \times priors$$

- Jeżeli znane są $P(X = x|Y = y_k)$ i $P(Y = y_k)$, mianownik $P(X = x)$ może zostać obliczony jako:

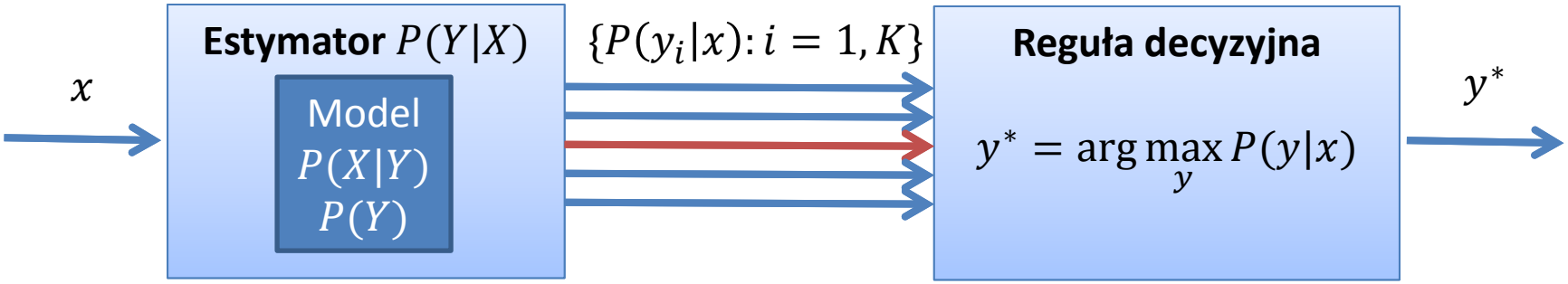
$$P(X = x) = (X = x|Y = y_1)P(Y = y_1) + P(X = x|Y = y_2)P(Y = y_2) + \cdots + P(X = x|Y = y_K)P(Y = y_K)$$

$$P(X = x) = \sum_{i=1}^{K} P(X = x|Y = y_i)P(Y = y_i)$$

Naiwny klasyfikator Bayesa

Jest klasyfikatorem generatywnym:

- Model obejmuje $P(X|Y)$ – prawdopodobieństwo obserwacji dla różnych etykiet klas oraz $P(Y)$ - prawdopodobieństwa a priori klas
- Wykorzystywana jest reguła Bayesa do wyznaczenia $P(Y|X)$ – prawdopodobieństwa warunkowego klasy dla danej obserwacji
- Wybierana jest ta etykieta, dla której prawdopodobieństwo $P(Y = y^*|x)$ jest największe



Ostateczny wynik (y^*)zależy od porównania prawdopodobieństw warunkowych.

$$P(Y = y_k|X = x) = \frac{P(X = x|Y = y_k)P(Y = y_k)}{P(X = x)}$$

Dla przeprowadzenia porównania obliczenie $P(X = x)$, który jest jedynie czynnikiem skalującym nie jest konieczne.

Naiwny klasyfikator Bayesa

- Model nazywany jest **naiwnym**, ponieważ zakłada bardzo mocne uproszczenie: dla ustalonej etykiety klasy: wszystkie są cechy są niezależne od siebie.
- Formalnie, jeżeli X_i oraz X_j są cechami (zmiennymi), to:
$$P(X_i|X_j, Y) = P(X_i|Y)$$
- Obliczmy $P(X_1, X_2, Y)$:
$$P(X_1, X_2, Y) = P(X_1|X_2, Y)P(X_2, Y) = P(X_1|X_2, Y)P(X_2|Y)P(Y)$$
- Stąd, stosując założenie naiwnego modelu Bayesa
$$P(X_1, X_2, Y) = P(X_1|Y)P(X_2|Y)P(Y)$$
$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y)$$

Uogólniając:

$$P(X_1, X_2, \dots, X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

- Założenie bardzo upraszcza złożoność modelu. Dla n cech binarnych (0/1):
  - pełny model: $(2^n - 1) \cdot K$ parametrów
    (każdej z 2^n wariacji przypisujemy prawdopodobieństwo dla K klas, jedno możemy pominąć)
  - Naiwny model Bayesa: $n \cdot K$ parametrów
    (przypisujemy prawdopodobieństwo n zmiennym dla K klas)

Przykład: weather

$X_1: outlook \in \{sunny, overcast, rainy\}$
$X_2: temp \in \{hot, mild, cool\}$
$X_3: humidity \in \{high, normal\}$
$X_4: windy \in \{TRUE, FALSE\}$
$Y: play \in \{yes, no\}$

W sumie X obejmuje 3x3x2x2=36 wariacji.

	outlook	temp	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Załóżmy, że mamy ocenić $P(y|x^j)$ dla danych wejściowych x^j
Bezpośrednie wyznaczenie

$$P(y|X_1 = x_1^j, X_2 = x_2^j, X_3 = x_3^j, X_4 = x_4^j)$$

na podstawie danych tabeli byłoby błędem:

- nadmierne dopasowanie (**overfitting**) do istniejących danych
- jak wybrać prawdopodobieństwa dla danych nie ujętych w tabeli? Przyjąć prawdopodobieństwa klas, czyli: $P(yes) = \#yes/14$ oraz $P(no) = \#no/14$?

Obliczenia

	outlook	temp	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Obliczamy prawdopodobieństwo wystąpienia klasy $P(c_k)$ dla wszystkich klas c_k

- $P(no) = \#no/(\#no + \#yes)$
- $P(yes) = \#yes/(\#no + \#yes)$

#no	5	P(no)	0.36
#yes	9	P(yes)	0.64

Dla każdej wartości atrybutu v_{ij} zmiennej X_i i każdej klasy c_k obliczamy

$$P(v_{ij}|c_k) = \frac{\#(X_i = v_{ij} \ \wedge \ Y = c_k)}{\#c_k}$$

Na przykład dla atrybutu **outlook**:

$$P(sunny|no) = \frac{\#(sunny \wedge no)}{\#no}$$

$$P(overcast|yes) = \frac{\#(overcast \wedge yes)}{\#yes}$$

#sunny ^ no	3	P(sunny|no)	0.60
#overcast ^ no	0	P(overcast|no)	0.00
#rainy ^ no	2	P(overcast|no)	0.40
#sunny ^ yes	2	P(sunny|yes)	0.22
#overcast ^ yes	4	P(overcast|yes)	0.44
#rainy ^ yes	3	P(overcast|yes)	0.33

Estymacja prawdopodobieństwa

#no	5	P(no)	0.36	class priors
#yes	9	P(yes)	0.64	
#sunny ^ no	3	P(sunny|no)	0.60	outlook
#overcast ^ no	0	P(overcast|no)	0.00	
#rainy ^ no	2	P(rainy|no)	0.40	
#sunny ^ yes	2	P(sunny|yes)	0.22	
#overcast ^ yes	4	P(overcast|yes)	0.44	
#rainy ^ yes	3	P(rainy|yes)	0.33	
#hot ^ no	2	P(hot|no)	0.40	temp
#mild ^ no	2	P(mild|no)	0.40	
#cool ^ no	1	P(cool|no)	0.20	
#hot ^ yes	2	P(hot|yes)	0.22	
#mild ^ yes	4	P(overcast|yes)	0.44	
#cool ^ yes	3	P(cool|yes)	0.33	
#high ^ no	4	P(high|no)	0.80	humid
#normal ^ no	1	P(normal|no)	0.20	
#high ^ yes	3	P(high|yes)	0.33	
#normal ^ yes	6	P(normal|yes)	0.67	
#TRUE ^ no	3	P(TRUE|no)	0.60	windy
#FALSE^ no	2	P(FALSE|no)	0.40	
#TRUE ^ yes	3	P(TRUE|yes)	0.33	
#FALSE^ yes	6	P(FALSE|yes)	0.67	

$$P(no|sunny, cool, normal, FALSE)$$
$$= P(no) \cdot P(sunny|no)P(cool|no)$$
$$\cdot P(normal)|no)P(FALSE|no)$$
$$= 0.36 \cdot 0.60 \cdot 0.20 \cdot 0.20 \cdot 0.40$$
$$= 0.0034$$
$$P(yes|sunny, cool, normal, FALSE)$$
$$= 0.0141$$

	outlook	temp	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Estymacja prawdopodobieństwa

#no	5	P(no)	0.36	
#yes				
#sunny ^ no				
#overcast ^ no				
#rainy ^ no				
#sunny ^ yes				
#overcast ^ yes				
#rainy ^ yes				
#hot ^ no				
#mild ^ no				
#cool ^ no				
#hot ^ yes	2	P(hot|yes)	0.22	te
#mild ^ yes	4	P(overcast|yes)	0.44	
#cool ^ yes	3	P(cool|yes)	0.33	
#high ^ no	4	P(high|no)	0.80	humid
#normal ^ no	1	P(normal|no)	0.20	
#high ^ yes	3	P(high|yes)	0.33	
#normal ^ yes	6	P(normal|yes)	0.67	
#TRUE ^ no	3	P(TRUE|no)	0.60	windy
#FALSE^ no	2	P(FALSE|no)	0.40	
#TRUE ^ yes	3	P(TRUE|yes)	0.33	
#FALSE^ yes	6	P(FALSE|yes)	0.67	

$P(no|sunny, cool, normal, FALSE)$

Możliwa jest estymacja prawdopodobieństwa dla nieznanych danych:

$P(yes|sunny, cool, normal, FALSE) = 0.0105$

$P(no|rainy, mild, high, TRUE) = 0.0068$

	outlook	temp	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Estymacja prawdopodobieństwa

#no	5	P(no)	0.36	class priors
#yes	9	P(yes)	0.64	
#sunny ^ no	3	P(sunny|no)	0.60	outlook
#overcast ^ no	0	P(overcast|no)	0.00	
#rainy ^ no	2	P(rainy|no)	0.40	
#sunny ^ yes	2	P(sunny|yes)	0.22	
#overcast ^ yes	4	P(overcast|yes)	0.44	
#rainy ^ yes	3	P(rainy|yes)	0.33	
#hot ^ no	2	P(hot|no)	0.40	temp
#mild ^ no	2	P(mild|no)	0.40	
#cool ^ no	1	P(cool|no)	0.20	
#hot ^ yes	2	P(hot|yes)	0.22	
#mild ^ yes	4	P(overcast|yes)	0.44	
#cool ^ yes	3	P(cool|yes)	0.33	
#high ^ no	4	P(high|no)	0.80	humid
#normal ^ no	1	P(normal|no)	0.20	
#high ^ yes	3	P(high|yes)	0.33	
#normal ^ yes	6	P(normal|yes)	0.67	
#TRUE ^ no	3	P(TRUE|no)	0.60	windy
#FALSE^ no	2	P(FALSE|no)	0.40	
#TRUE ^ yes	3	P(TRUE|yes)	0.33	
#FALSE^ yes	6	P(FALSE|yes)	0.67	

$$P(no|sunny, cool, normal, FALSE)$$
$$= P(no) \cdot P(sunny|no)P(cool|no)$$
$$\cdot P(normal)|no)P(FALSE|no)$$
$$= 0.36 \cdot 0.60 \cdot 0.20 \cdot 0.20 \cdot 0.40$$
$$= 0.0034$$
$$P(yes|sunny, cool, normal, FALSE)$$
$$= 0.0141$$

	outlook	temp	humidity	windy	play
1	sunny	hot	high	FALSE	no
2	sunny	hot	high	TRUE	no
3	overcast	hot	high	FALSE	yes
4	rainy	mild	high	FALSE	yes
5	rainy	cool	normal	FALSE	yes
6	rainy	cool	normal	TRUE	no
7	overcast	cool	normal	TRUE	yes
8	sunny	mild	high	FALSE	no
9	sunny	cool	normal	FALSE	yes
10	rainy	mild	normal	FALSE	yes
11	sunny	mild	normal	TRUE	yes
12	overcast	mild	high	TRUE	yes
13	overcast	hot	normal	FALSE	yes
14	rainy	mild	high	TRUE	no

Estymacja prawdopodobieństwa

#no	5	P(no)	0.36	class priors
#yes	9	P(yes)	0.64	
#sunny ^ no	3	P(sunny|no)	0.60	outlook
#overcast ^ no	0	P(overcast|no)	0.00	
#rainy ^ no	2	P(rainy|no)	0.40	
#sunny ^ yes	2	P(sunny|yes)	0.22	
#overcast ^ yes	4	P(overcast|yes)	0.44	
#rainy ^ yes	3	P(rainy|yes)	0.33	
#hot ^ no	2	P(hot|no)	0.40	

$$P(no|sunny, cool, normal, FALSE)$$
$$= P(no) \cdot P(sunny|no)P(cool|no)$$
$$\cdot P(normal)|no)P(FALSE|no)$$
$$= 0.36 \cdot 0.60 \cdot 0.20 \cdot 0.20 \cdot 0.40$$
$$= 0.0034$$
$$P(yes|sunny, cool, normal, FALSE)$$

Problem

Jeśli kombinacja wartości atrybutu v_{ij} i zmiennej wyjściowej c_k nigdy nie występuje w zbiorze uczącym, zawsze przypisywane jest zerowe prawdopodobieństwo:

$$P(Y = c_k| \ldots, X_i = v_{ij}, \ldots) = 0$$

Na przykład, jeżeli zachodzi:

$$P(overcast|no) = 0$$

wówczas zawsze:

$$P(no|overcast,*,*,*) = P(no)P(overcast|no) \cdot \ldots = 0$$

Potencjalnie zwiększa się błąd generalizacji

emp	humidity	windy	play
ot	high	FALSE	no
ot	high	TRUE	no
ot	high	FALSE	yes
ild	high	FALSE	yes
ool	normal	FALSE	yes
ool	normal	TRUE	no
ool	normal	TRUE	yes
ild	high	FALSE	no
ool	normal	FALSE	yes
ild	normal	FALSE	yes
ild	normal	TRUE	yes
ild	high	TRUE	yes
ot	normal	FALSE	yes
ild	high	TRUE	no

Wygładzanie

- Idea wygładzania polega na zastąpieniu zera w liczniku wartością α (powiedzmy $\alpha = 1$)
- Wówczas: $P(overcast|no) = \frac{\#(overcast \wedge no)+1}{\#no}$
- Taka zmiana dla jednego atrybutu byłaby jednak niesprawiedliwa. Analogicznie powinniśmy zmienić:
  - $P(sunny|no) = \frac{\#(sunny \wedge no)+1}{\#no}$
  - $P(rainy|no) = \frac{\#(rainy \wedge no)+1}{\#no}$
- Prawdopodobieństwa $P(sunny|no) + P(overcast|no) + P(rainy|no)$ nie sumują się do 1, ale do $\frac{\#no+3}{\#no}$. Aby to skorygować, należy dodać 3 do mianownika:
  - $P(overcast|no) = \frac{\#(overcast \wedge no)+1}{\#no+3}$
  - $P(sunny|no) = \frac{\#(sunny \wedge no)+1}{\#no+3}$
  - $P(rainy|no) = \frac{\#(rainy \wedge no)+1}{\#no+3}$

Wygładzanie

- Prawdopodobieństwa $P(Y = c_k | X_j = v_{ij})$ mają wstępnie przypisane wartości $1/K$, gdzie K jest liczbą klas.
- Następnie podczas uczenia są one korygowane:

$$P\left(Y = c_k \middle| X_i = v_{ij}\right) = \frac{\#(X_j = v_{ij} \wedge Y = c_k) + \alpha}{\#(Y = c_k) + K\alpha}$$

- Dzięki temu prawdopodobieństwo $P\left(Y = c_k \middle| X_j = v_{ij}\right)$ nigdy nie będzie zerowe, minimalna wartość to $\frac{\alpha}{\#(Y=c_k)+K\alpha}$

#no	5	**6**	P(no)	0.38	class priors
#yes	9	**10**	P(yes)	0.63	
#sunny ^ no	3	**4**	P(sunny|no)	0.50	outlook
#overcast^ no	0	**1**	P(overcast|no)	0.13	
#rainy^ no	2	**3**	P(rainy|no)	0.38	
#sunny ^ yes	2	**3**	P(sunny|yes)	0.25	
#overcast^ yes	4	**5**	P(overcast|yes)	0.42	
#rainy^ yes	3	**4**	P(rainy|yes)	0.33	

$$\frac{1}{5+3} = 0.125$$

$$P(no|overcast, cool, normal, false) = 0.0014$$
$$P(yes|overcast, cool, normal, false) = 0.0281$$

Kategoryzacja tekstów

- Zaliczenie artykułu prasowego do grupy tematycznej:
  $Y = \{wiadomości, biznes, sport, styl\ życia, wiedza, motoryzacja \ldots\}$
- Klasyfikacja poczty e-mail: $Y = \{spam, nospam\}$
- Analiza sentymentu, np. w ocenach usług, towarów:
  $Y = \{pozytywny, neutralny, negatywny\}$

Y $\quad$ $\mathrm{X} = [x_1, \ldots, x_N]$

Kategoria	Tekst	…	grupa	inwestycyjne	ligowe	nakłady	negocjacje	rozgrywki	rozpocząć	…
biznes	Wczoraj grupa EDF poinformowała o wybraniu IFM Investors do wyłącznych negocjacji. Rozpoczęcie negocjacji nastąpi…		1+α	α	α	α	2+α	α	1+α	
biznes	Po trzech kw. 2016 r. nakłady inwestycyjne grupy wyniosły 152 mln zł w porównaniu do 166 mln zł w analogicznym okresie 2015 r.		1+α	1+α	α	1+α	α	α	α	
sport	Drugoligowa Polonia Bytom ligowe rozgrywki rozpoczęła z bagażem czterech minusowych punktów		α	α	1+α	α	α	1+α	1+α	

N kolumn
x_i - liczba wystąpień słowa w dokumencie

N – to liczba słów w słowniku

Kategoryzacja tekstów - uczenie

- Aby wyznaczyć $P(y_k)$ – należy policzyć dokumenty w kategorii y_k

$$P(y_k) = \frac{\#dokumentów\ w\ kategorii\ y_k}{\#dokumentów}$$

- Aby wyznaczyć $P(x_i|y_k)$ – należy policzyć, ile razy słowo w_i występuje w dokumentach zaliczonych do kategorii y_k

$$P(x_i|y_k) = \frac{\#wystąpień\ słowa\ w_i\ w\ dokumentach\ kategorii\ y_k}{\#słów\ w\ dokumentach\ kategorii\ y_k}$$

Kategoria	Tekst	...	grupa	inwestycyjne	ligowe	nakłady	negocjacje	rozgrywki	rozpocząć	...
biznes	Wczoraj grupa EDF poinformowała o wybraniu IFM Investors do wyłącznych negocjacji. Rozpoczęcie negocjacji nastąpi...		1+α	α	α	α	2+α	α	1+α	
biznes	Po trzech kw. 2016 r. nakłady inwestycyjne grupy wyniosły 152 mln zł w porównaniu do 166 mln zł w analogicznym okresie 2015 r.		1+α	1+α	α	1+α	α	α	α	
sport	Drugoligowa Polonia Bytom ligowe rozgrywki rozpoczęła z bagażem czterech minusowych punktów		α	α	1+α	α	α	1+α	1+α	

$$P(grupa|biznes) = \frac{2+2\alpha}{Suma(biznes)}$$

Przyjętą nazwą tego modelu jest Multinomial Naive Bayes.

Jeśli $\alpha = 1$, jest to tzw. wygładzania Laplace'a. Jeśli $\alpha < 1$ wygładzanie Lidstone'a

Kategoryzacja tekstów - klasyfikacja

1. Dla wszystkich kategorii y_k wyznacz
   $P(y_k|x) = P(y_k)\prod_{i=1}^{N} P(x_i|y_k)$
2. Wybierz $y^* = \arg\max_{y_k} P(y_k)\prod_{i=1}^{N} P(x_i|y_k)$

- Problemy obliczeniowe:
  - N – liczba wyrazów w słowniku, $\mathrm{N} \approx 10^4$
  - W dokumencie występuje m słów, $\mathrm{m} \approx 10^3$
  - Jeśli i-te słowo występuje w dokumentach kategorii y_k,
    $p(x_i|y_k) \approx \frac{1}{10^3} = 10^{-3}$
  - Jeśli i-te słowo nie występuje w dokumentach kategorii y_k,
    $p(x_i|y_k) \approx \frac{1}{10^4} = 10^{-4}$
  - W sumie $P(y_k|x) \approx 10^{-3000} \cdot 10^{-36000} = 10^{-39000}$
  - Tego nie da się obliczyć z wymaganą dokładnością! Mnożenie wprowadza znacznie więcej błędów numerycznych niż dodawanie.

Kategoryzacja tekstów

- Przy określaniu klasy $y^* = \arg\max_{y_k} P(y_k) \prod_{i=1}^{N} P(x_i|y_k)$ przeprowadzamy porównanie prawdopodobieństw:

$$P(y_1|x) > P(y_2|x) \text{ ???}$$

- Zamiast porównywać prawdopodobieństwa można porównać ich logarytmy:

$$\ln\left(P(y_1|x)\right) > \ln(P(y_2|x)) \text{ ???}$$

- Obliczana jest suma, która jest stabilna numerycznie:

$$\ln\left(P(y_1|x)\right) = \ln\left(P(y_k)\right) + \sum_{i=1}^{n} \ln(P(x_i|y_k))$$

Kategoryzacja tekstów

- Ze względu na prostotę założeń, zastosowanie klasyfikatora opartego na naiwnym modelu Bayesa daje dobre rezultaty kiedy N (liczba cech) jest duża, nawet rzędu 10^4.
- Dla takich wymiarów inne modele często zawodzą.
- Dla problemu kategoryzacji tekstów osiąga się trafność (accuracy) rzędu 85%-90%.

True\predicted	c_1	c_2	c_3
c_1	5	0	2
c_2	1	7	2
c_3	1	4	3

nieprawidłowo sklasyfikowane

prawidłowo sklasyfikowane

$$accuracy = \frac{\sum_{i=1}^{k} e[i,i]}{\sum_{i=1}^{k}\sum_{j=1}^{k} e[i,k]}$$

Zmienne numeryczne

- W przypadku zmiennych numerycznych zakłada się, że dla każdej z cech X_i rozkład $P(x_i|y_k)$ jest rozkładem normalnym

$$p(x_i, y_k) = N(\mu_{ik}, \sigma_{ik}^2) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{1}{2\sigma_{ik}^2}(x_i-\mu_{ik})^2}$$

- Średnia μ_{ik} i wariancja σ_{ik}^2 są najbardziej prawdopodobnymi estymatami (ang. MLE, maximum likelihood estimates) liczonymi dla poszczególnych zmiennych X_i i wartości klas y_k :

$$\mu_{ik} = \frac{1}{\sum_{j=1}^{m} \mathbf{1}(y^j = y_k)} \sum_{j=1}^{m} x_i^j \mathbf{1}(\mathrm{y^j} = y_k)$$

$$\sigma_{ik}^2 = \frac{1}{\sum_{j=1}^{m} \mathbf{1}(y^j = y_k)} \sum_{j=1}^{m} (x_i^j - \mu_{ik})^2 \mathbf{1}(\mathrm{y^j} = y_k)$$

- $\sum_{j=1}^{m} \ldots$ - iteracja po danych uczących
- Indykator: $\mathbf{1}(p) = 1$, jeżeli $p = true$; 0, jeśli $p = false$
- $\sum_{j=1}^{m} \mathbf{1}(y^j = y_k)$ - liczba obserwacji należących do klasy y_k

Przykład

$y = 0$	
x_1	x_2
-3.00	2.00
-2.00	0.80
-1.30	0.60
0.00	-1.00
0.80	2.70
2.00	3.80
-2.80	-3.10
$\mu_{10} = -0.90$	$\mu_{20} = 0.83$
$\sigma_{10}^2 = 3.07$	$\sigma_{20}^2 = 4.65$

$y = 1$	
x_1	x_2
-2.20	-3.00
-1.10	-0.50
-0.70	-0.20
0.20	-2.10
1.40	2.00
1.10	-2.20
1.90	0.00
$\mu_{11} = 0.09$	$\mu_{21} = -0.86$
$\sigma_{11}^2 = 1.90$	$\sigma_{21}^2 = 2.49$

- Obserwacje $(x_1^j, x_2^j, y^j), j = 1, m$ zostały podzielone na dwie części: po lewej dla $y^j = 0$, po prawej $y^j = 1$
- Dla każdej z kolumn obliczana jest wartość średnią μ_{ik} i wariancję σ_{ik}^2

Przykład

Oznaczmy:

$$gauss(\mu, \sigma^2, x) = \frac{1}{\sqrt{2\pi\sigma_{ik}^2}} e^{-\frac{1}{2\sigma_{ik}^2}(x_i - \mu_{ik})^2}$$

- Dla danego wektora $x = [x_1, x_2]$ prawdopodobieństwa aposteriori obliczane są ze wzoru Bayesa (pomijając mianownik):

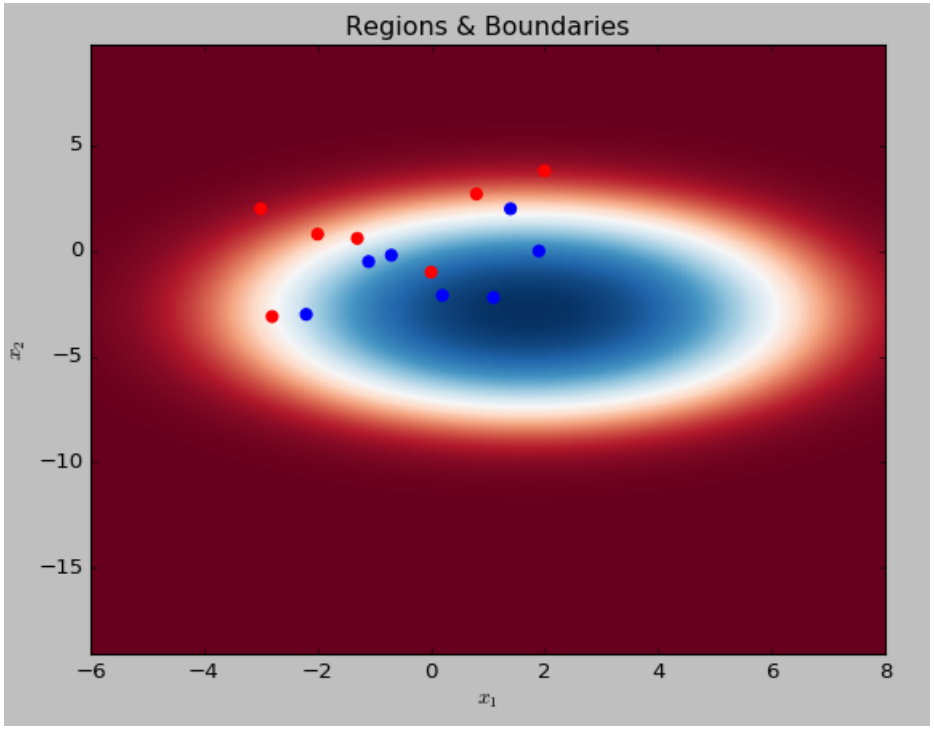
$$p_0 = \mathrm{p}(0) \cdot gauss(\mu_{10}, \sigma_{10}^2, x_1) \cdot gauss(\mu_{20}, \sigma_{20}^2, x_2)$$
$$p_1 = \mathrm{p}(1) \cdot gauss(\mu_{11}, \sigma_{11}^2, x_1) \cdot gauss(\mu_{21}, \sigma_{21}^2, x_2)$$

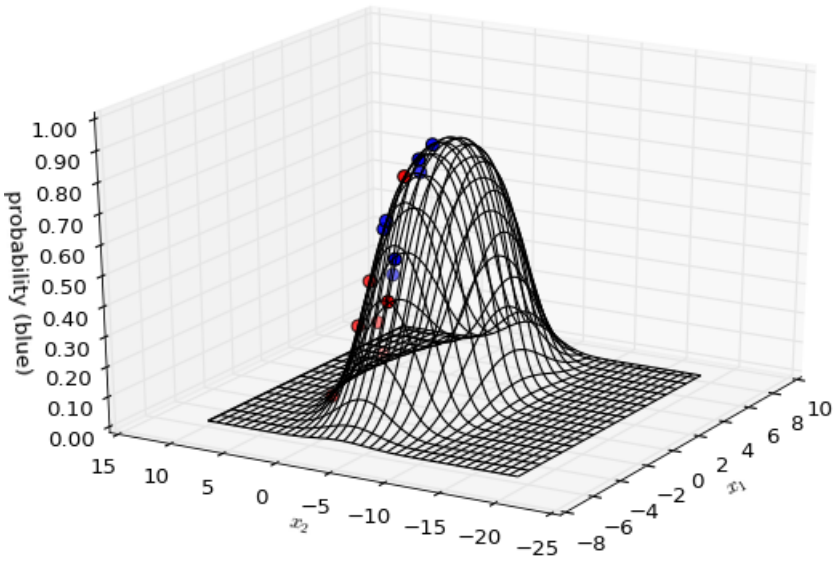
Ostatecznie:

$$p(0|x) = \frac{p_0}{p_0 + p_1} \text{ oraz } p(1|x) = \frac{p_1}{p_0 + p_1}$$

Przykład

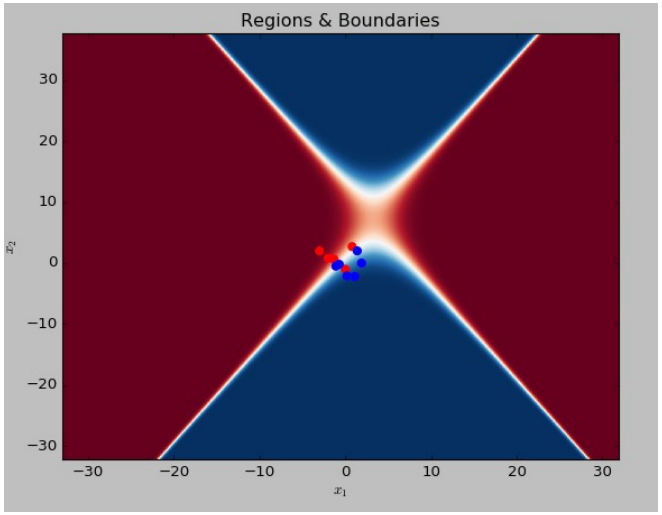
- Po lewej wyznaczone regiony decyzyjne
- Po prawej gęstość prawdopodobieństwa
- Najczęściej regiony mają postać elipsoid o osiach wzdłuż kierunków X_i

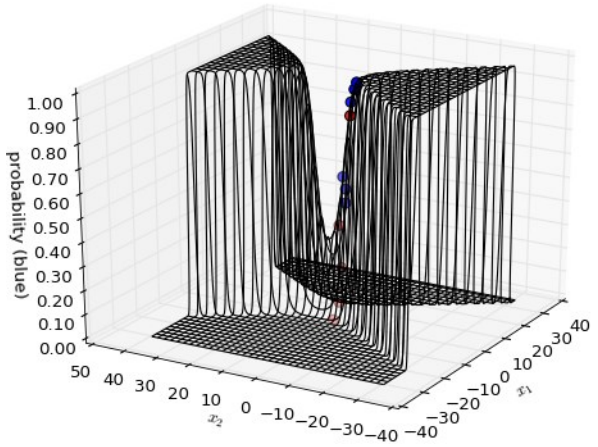


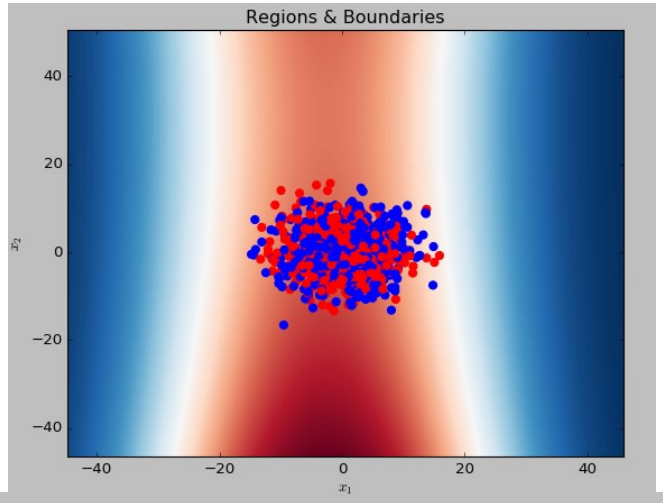


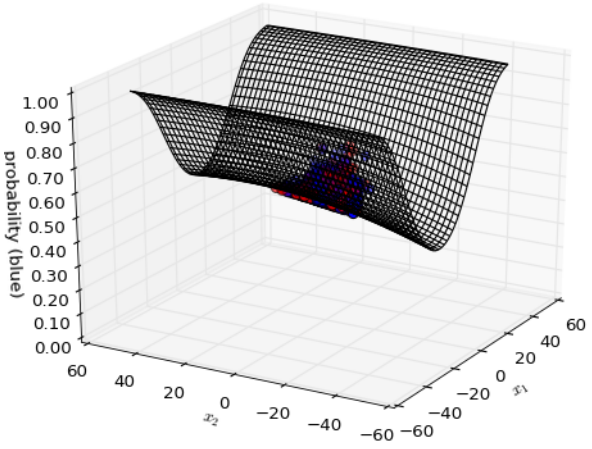
Wykresy regionów decyzyjnych

W przypadku 2D+klasyfikacja binarna występują dwie elipsy i ich kombinacje mogą dać różne kształty.

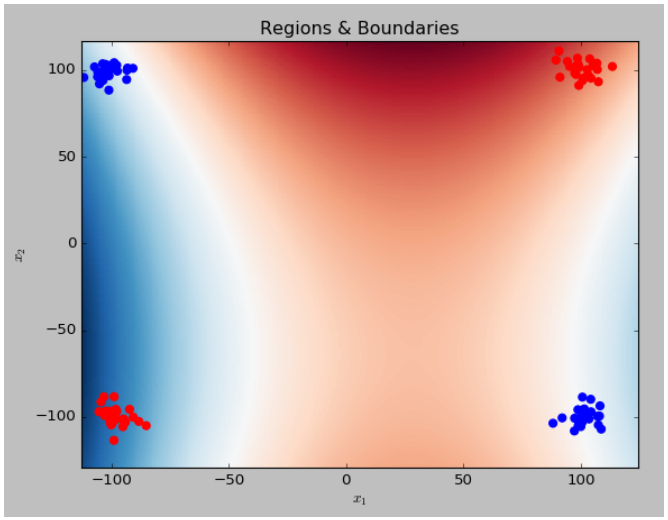


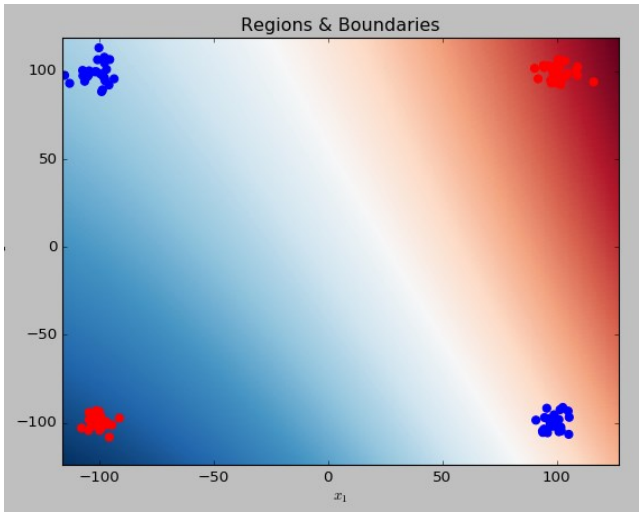


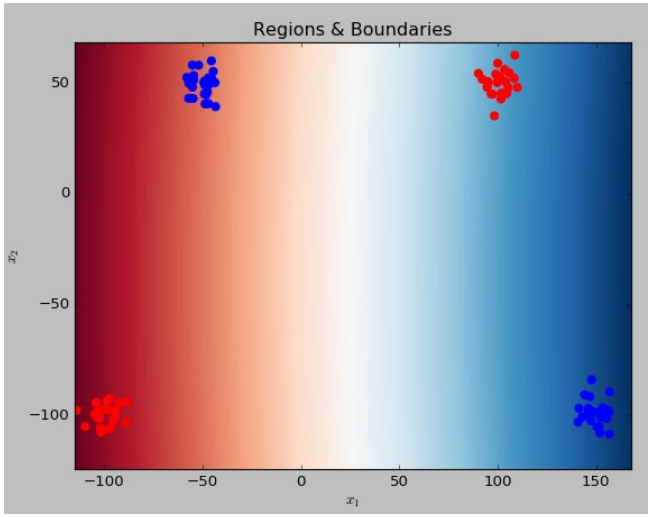




Wykresy regionów decyzyjnych







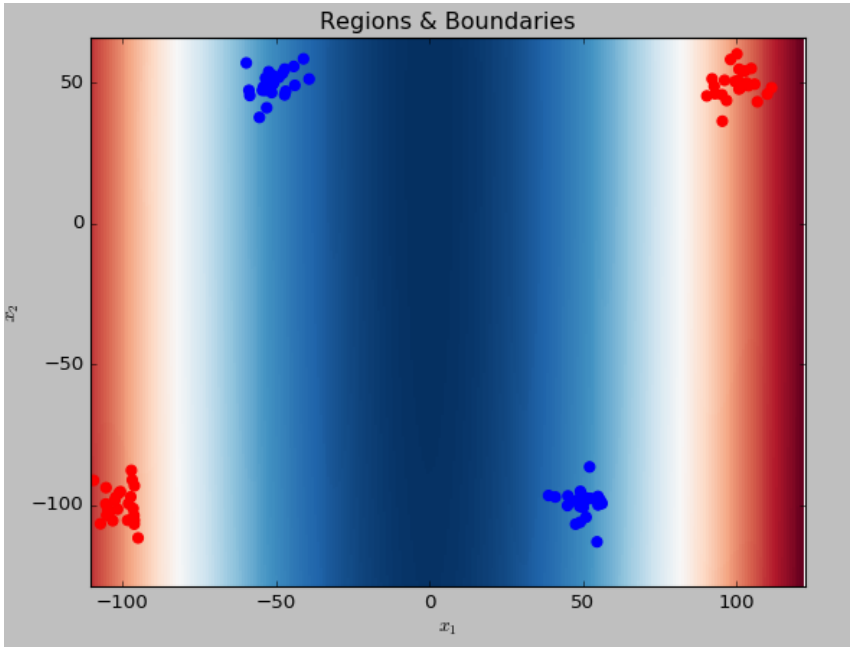
- Klasyfikator nie jest w stanie odseparować danych układających się w schemat XOR

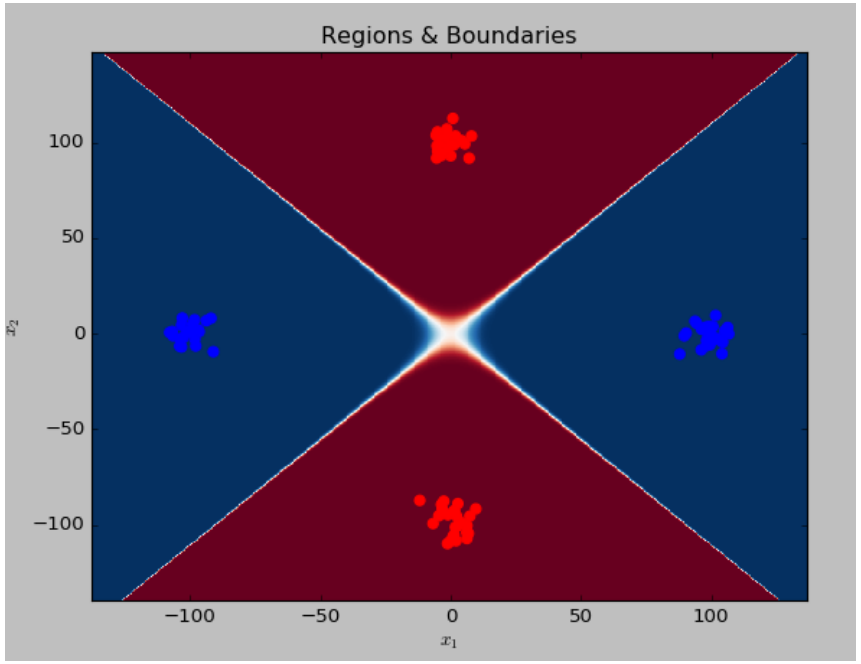
1	0
0	1

- Nie może obrócić elips, ponieważ musiałby uwzględnić kowariancje – niezgodne z założeniem o niezależności zmiennych

Wykresy regionów decyzyjnych

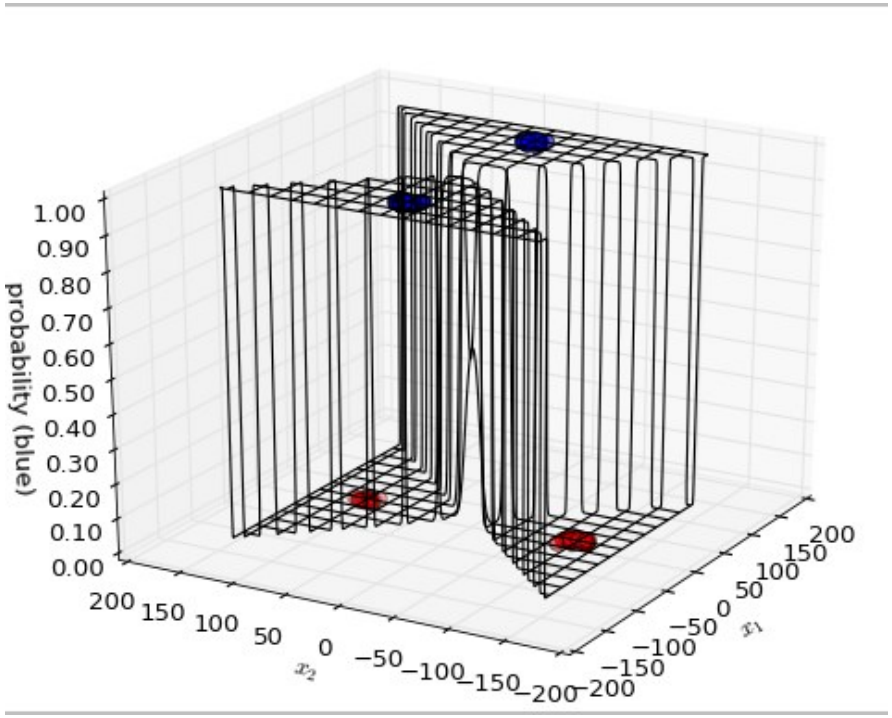
- W szczególnych przypadkach separacja jest możliwa:
  - Po lewej udało się dopasować wąską elipsę o dłuższej osi wzdłuż x_2
  - Po prawej suma dwóch elips dających ostre granice pomiędzy regionami

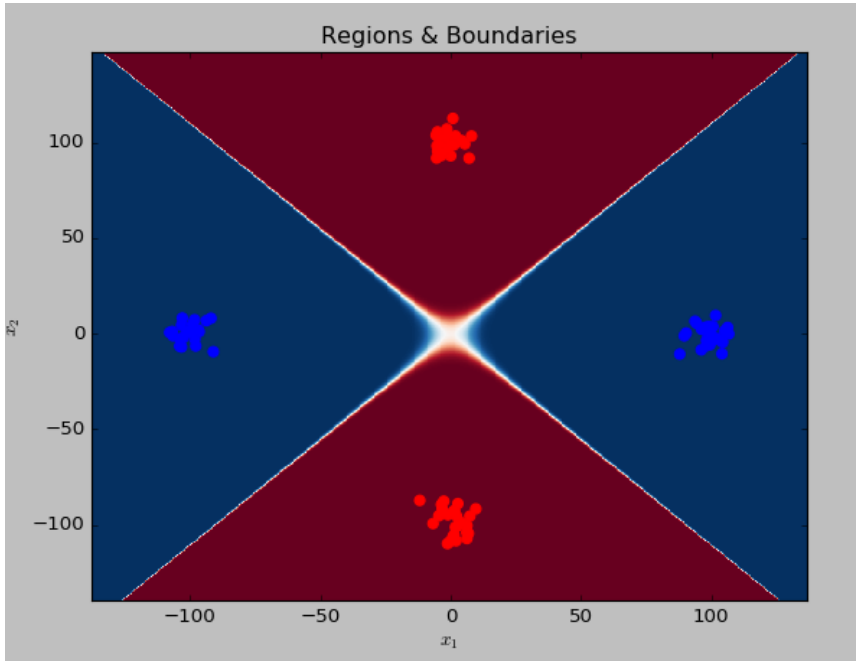




Wykresy regionów decyzyjnych

- W szczególnych przypadkach separacja jest możliwa:
  - Po lewej udało się dopasować wąską elipsę o dłuższej osi wzdłuż x_2
  - Po prawej suma dwóch elips dających ostre granice pomiędzy regionami

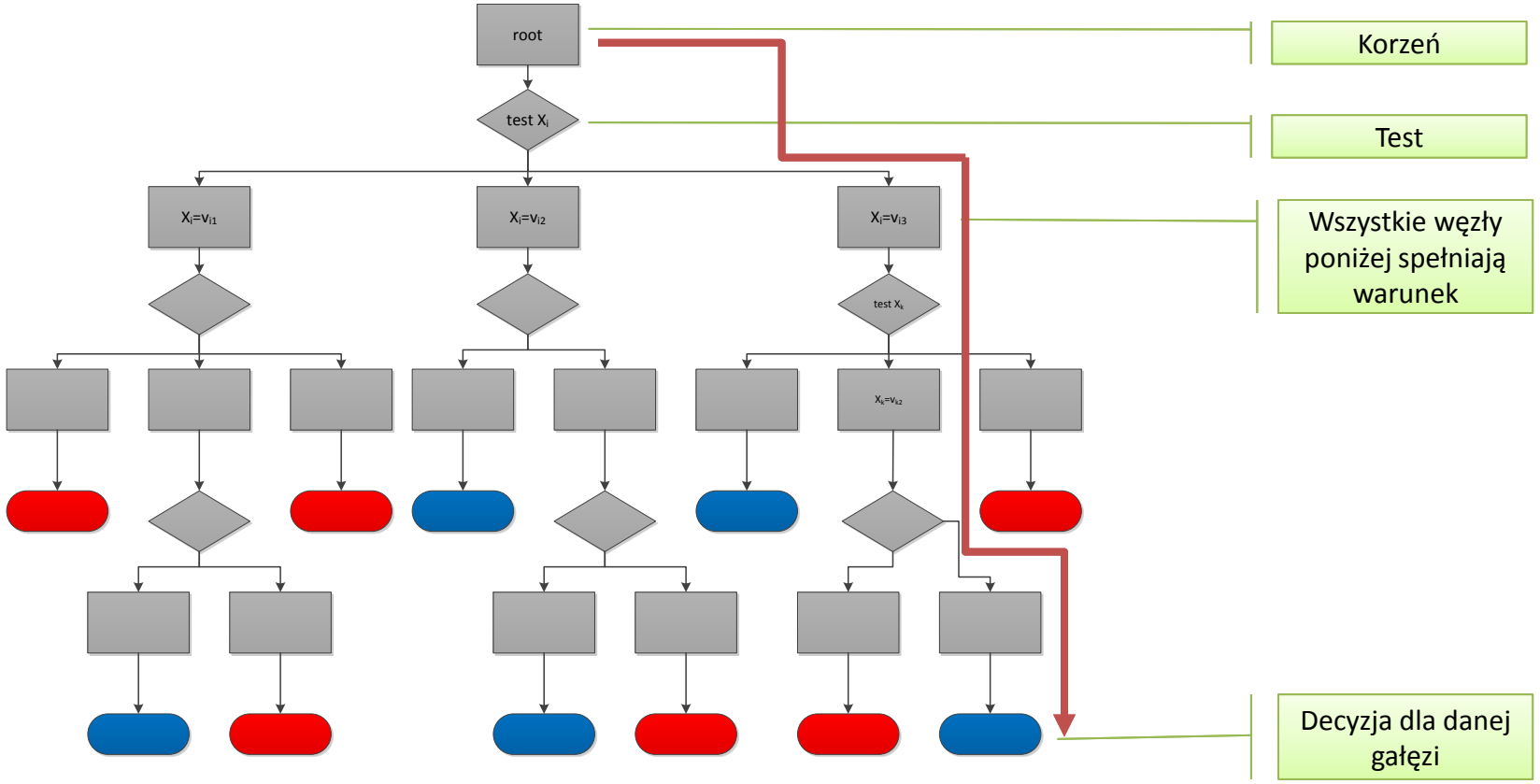




Drzewa decyzyjne

Reprezentacja

- Reprezentacja klasyfikatora w postaci drzewa
- Sztuczny węzeł root
- Przy przejściu pomiędzy węzłami: test wartości atrybutu
- Liście – decyzja klasyfikatora



Przykład

- Zbiór Auto MPG https://archive.ics.uci.edu/ml/datasets/Auto+MPG (398)
- Wszystkie atrybuty zdyskretyzowane:
  - mpg {bad, good} – zmienna decyzyjna
  - cylinders {8,6,5,4,3}
  - model year: 70-73,74-77,78-79,80-82}
  - origin {america,europe,asia}
  - pozostałe: {low, medium, high}

No	mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin
1	bad	8	medium	medium	medium	low	70-73	america
2	bad	8	high	medium	medium	low	70-73	america
21	bad	4	low	low	low	medium	70-73	europe
84	good	4	low	low	low	medium	70-73	america
144	bad	4	low	low	low	medium	74-77	europe
246	good	4	low	low	low	medium	78-79	america
304	good	4	low	low	low	medium	78-79	asia
326	good	4	low	low	low	high	80-82	europe
390	bad	6	medium	medium	medium	medium	80-82	america

Uczenie drzew decyzyjnych

- Budowa drzewa $T(D)$ na podstawie zbioru uczącego D
- Wskaźnik jakości:

$$J(T) = \frac{\#liczba\ błędnie\ przypisanych\ klas}{\#obserwacji\ w\ zbiorze\ uczącym}$$

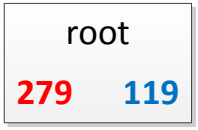
- Cel: znalezienie drzewa T^* minimalizującego $J(T)$:

$$T^* = \arg\min_{\mathrm{T}} J(T)$$

- Problem NP:
  - na poziomie korzenia wybór jednego z n atrybutów do przeprowadzenia testu
  - na poziomie $root-k$: wybór pomiędzy $n-k$ atrybutami (dla zmiennych kategorycznych)
- Stosowany jest algorytm zachłanny

Algorytm zachłanny

1. Tworzony jest korzeń drzewa.



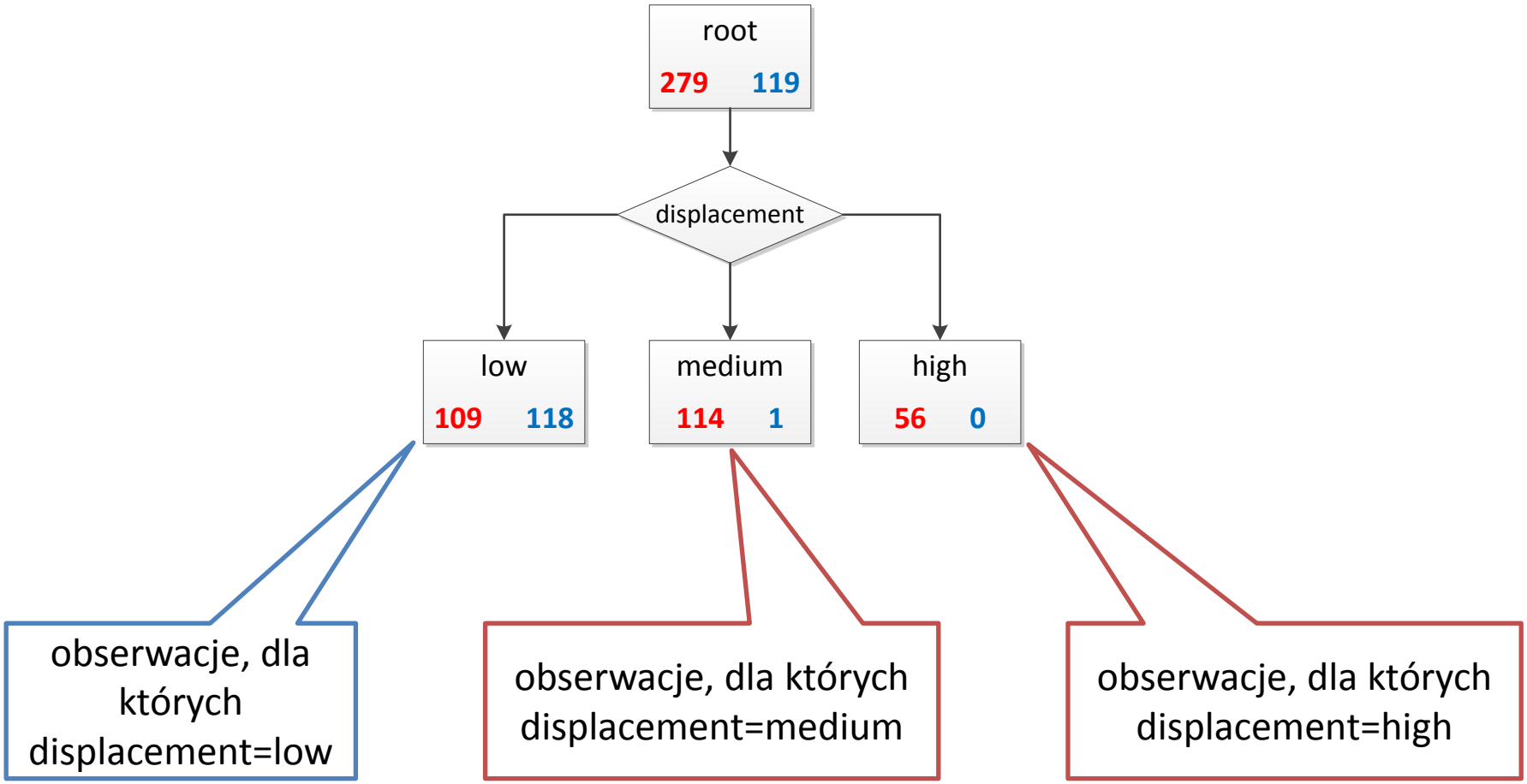
- Każdemu węzłowi przypisany jest podzbiór obserwacji zbioru uczącego: $D(node) \subseteq D$
- Korzeniowi przypisany jest cały zbiór uczący: $D(node) = D$.
- Liczby **279** i **119** odnoszą się obserwacji w $D(node)$, które mają etykiety **bad** i **good**
- Na podstawie tych liczb można przypisać węzłowi decyzję: w tym przypadku klasa większościowa (**bad**)
- Można również wyznaczyć błąd

$$J(node) = \frac{\#obserwacji\ klasy\ mniejszościowej}{|D(node)|} = \frac{119}{119+279} = 0.30$$

Węzeł root jest już klasyfikatorem. Tak skonstruowany klasyfikator jest dostępny w Weka. Nazywa się **ZeroR** (Maksymalny błąd systematyczny bias, zero wariancji.)

Algorytm zachłanny

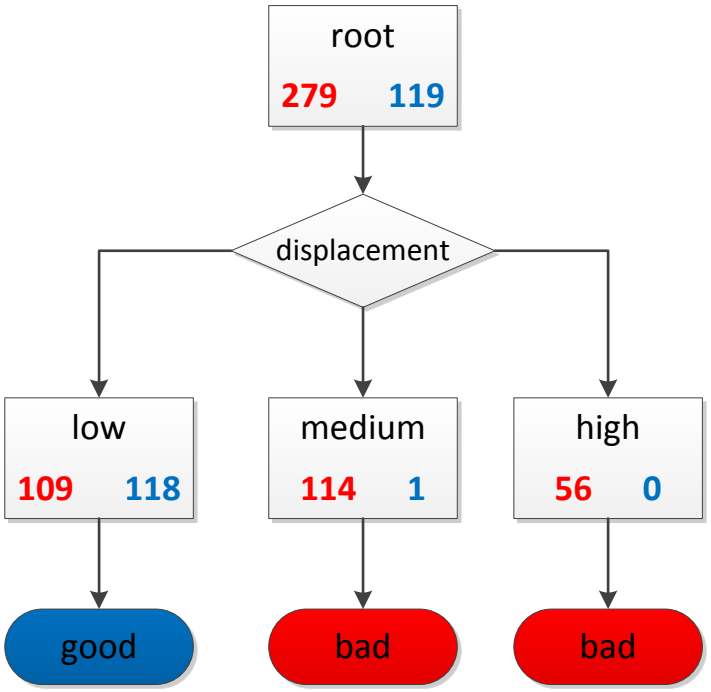
2. Wybierany jest atrybut określający podział drzewa



Decision stump

- Jednopoziomowe drzewo: węzeł + węzły potomne to tzw. decision stump („pień decyzyjny”)
- Dla danego drzewa można obliczyć błąd klasyfikacji:

$$J(T) = \frac{109 + 1}{398} = 0.28$$



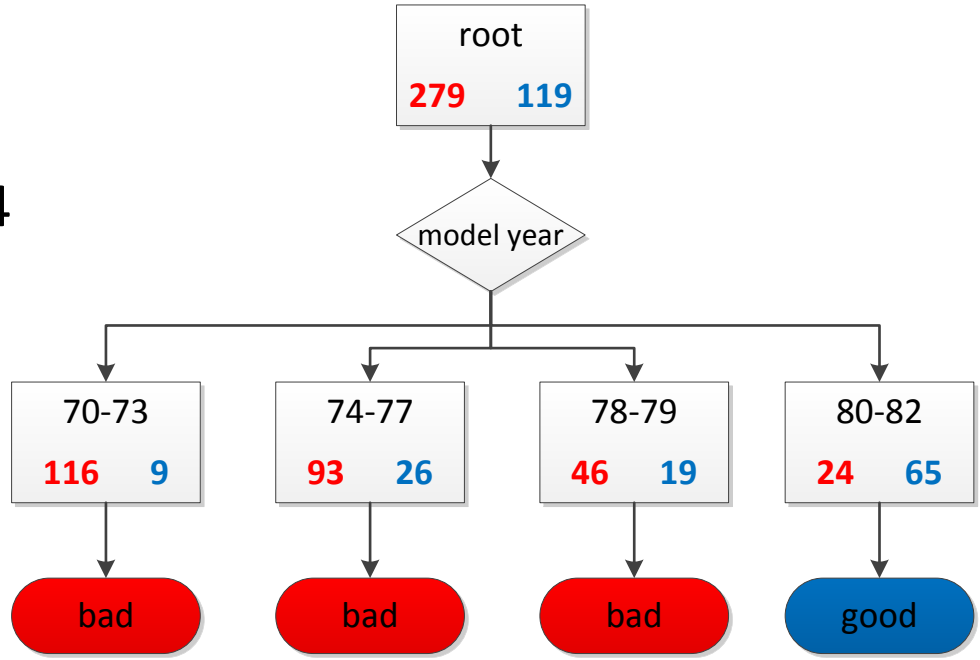
Klasyfikator OneR

- Klasyfikator w postaci prostego jednopoziomowego drzewa nazywa się **OneR** (dostępny w Weka)
- OneR ma zaskakująco dobre własności generalizacyjne:
  [Ian H. Witten, Eibe Frank, Mark A. Hall: Data Mining Practical Machine Learning Tools and Techniques Third Edition]
- W przypadku OneR wybierany jest do podziału atrybut minimalizujący błąd

Błąd dla OneR:

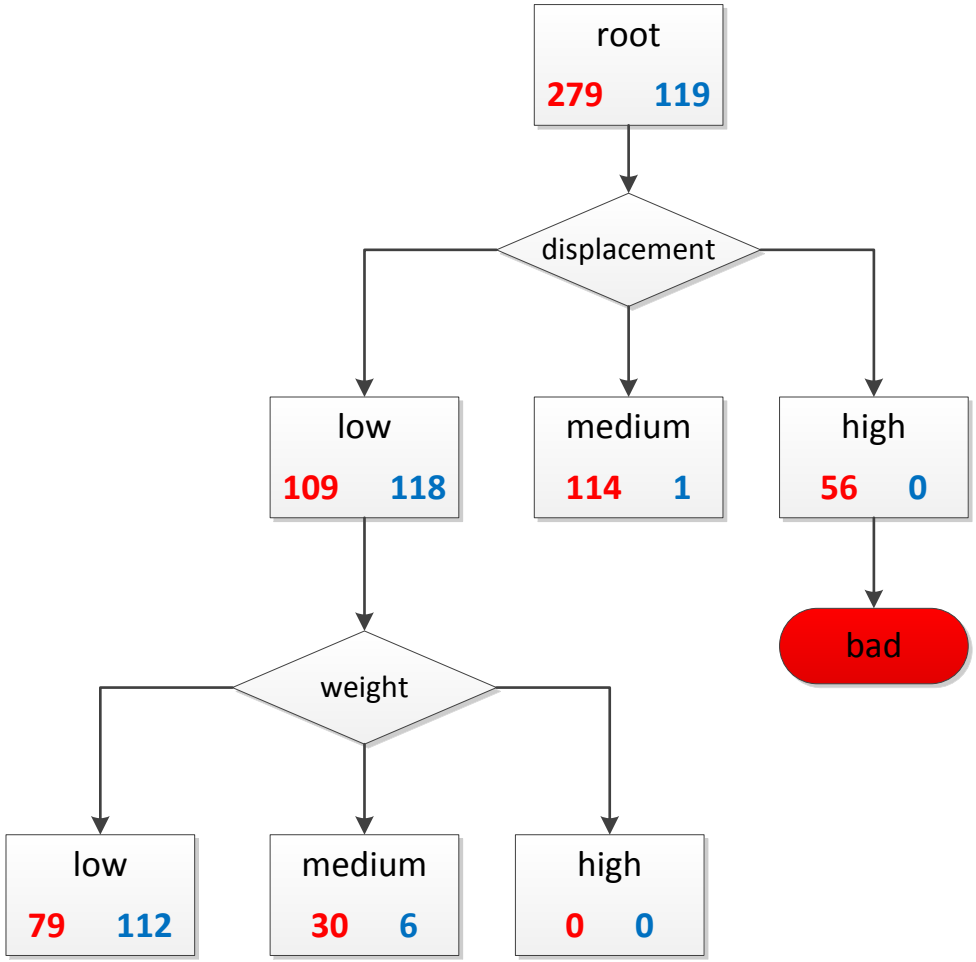
$$J(T) = \frac{9+26+19+24}{398} = 0.14$$

W przypadku drzew decyzyjnych, które rozwijane są głębiej, stosowane są inne kryteria wyboru atrybutu do podziału



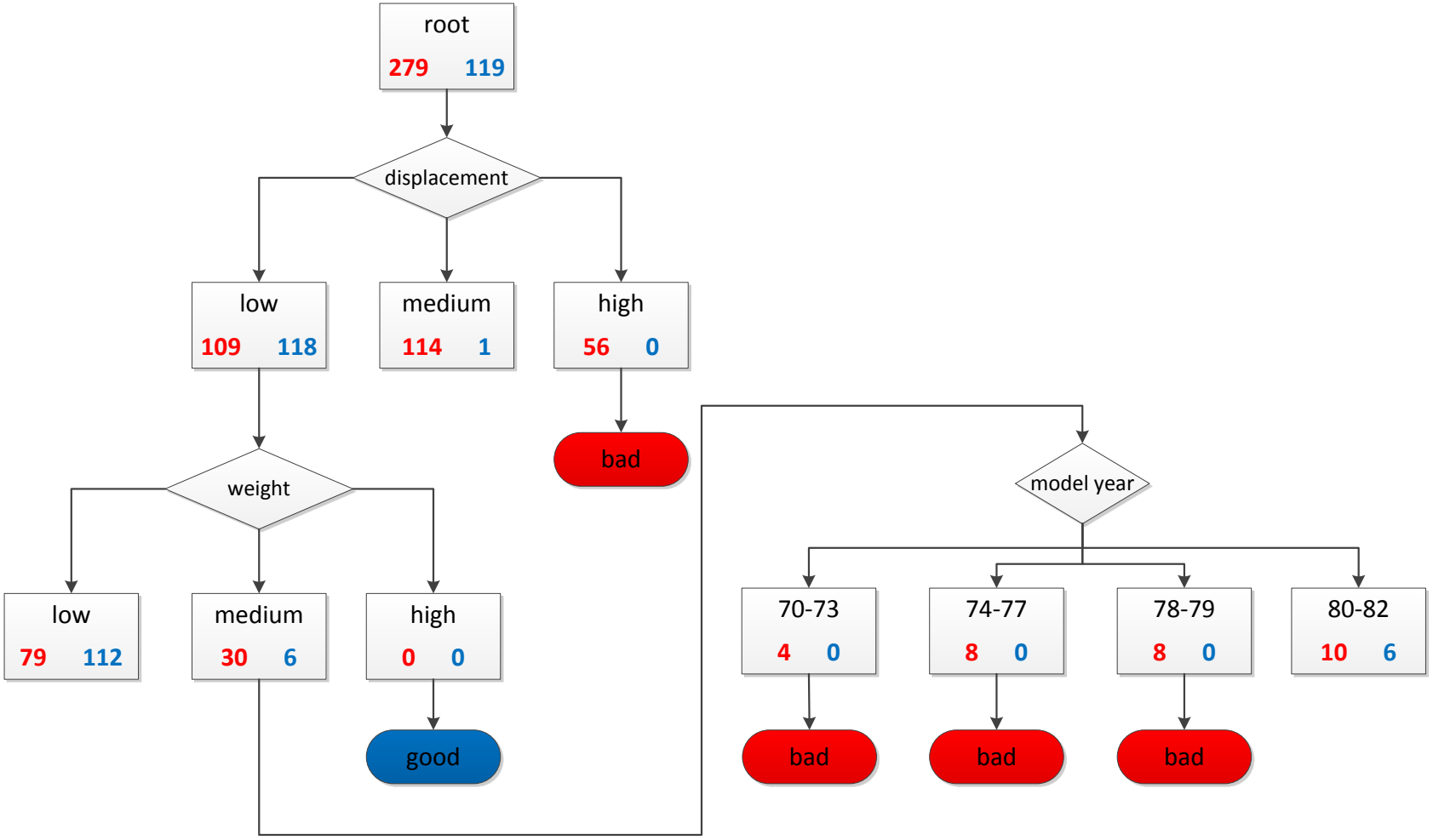
Algorytm zachłanny

3. Wybierany jest węzeł w drzewie, dla którego klasyfikacja jest niejednoznaczna i rozwijana jest kolejna gałąź.



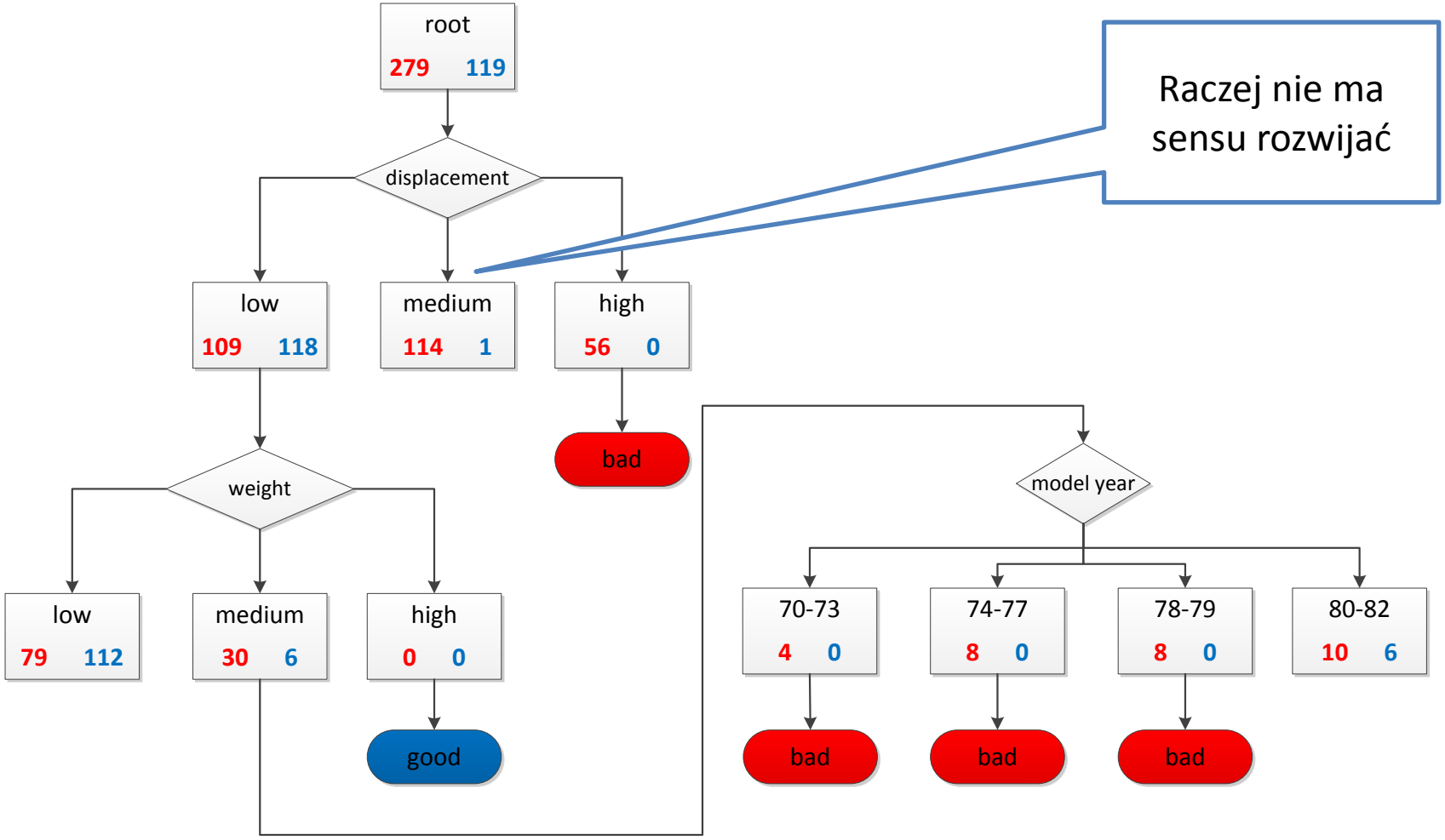
Algorytm zachłanny

Dalej rozwijana jest kolejna…



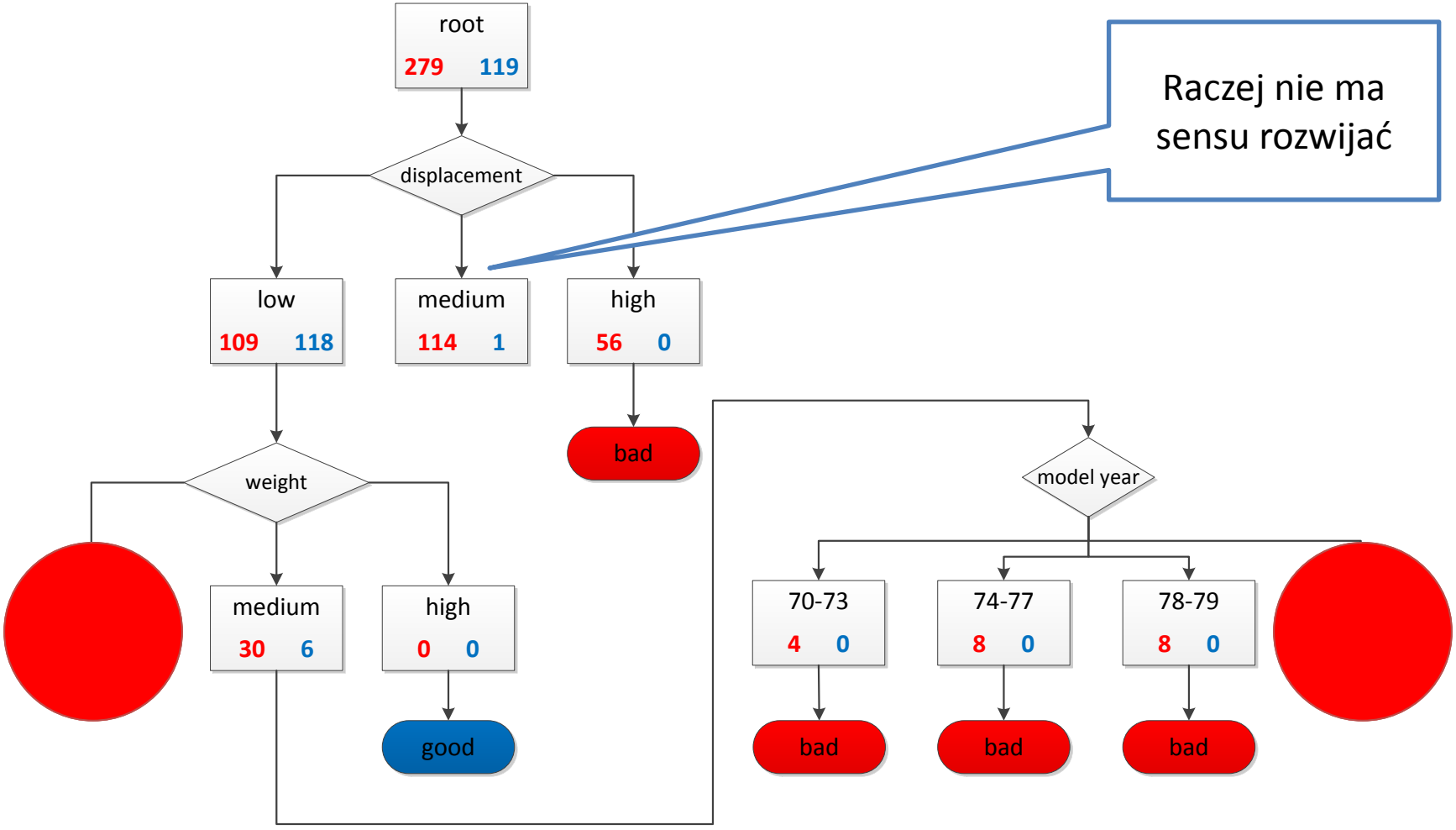
Algorytm zachłanny

Dalej rozwijana jest kolejna…



Algorytm zachłanny

Dalej rozwijana jest kolejna…



Algorytm zachłanny

1. Utwórz węzeł początkowy n_0 i przypisz mu cały zbiór uczący D: $D(n_0) \leftarrow D$
2. Dla każdego węzła n, który nie ma przypisanej decyzji:
   a) Jeśli $stop(n)$ - dalszy podział drzewa w węźle n nie ma sensu, przypisz mu decyzję na podstawie klasy większościowej w $D(n)$
   b) Wybierz atrybut (cechę) do podziału dla węzła n: $X_i \leftarrow select(n)$
   c) Dla każdej wartości v_{ij} ($j = 1, \dots, r$) atrybutu X_i:
      i. utwórz węzeł n_{ij} i połącz go z n
      ii. przypisz: $D(n_{ij}) \leftarrow \{(x, y) \in D(n): x[i] = v_{ij}\}$

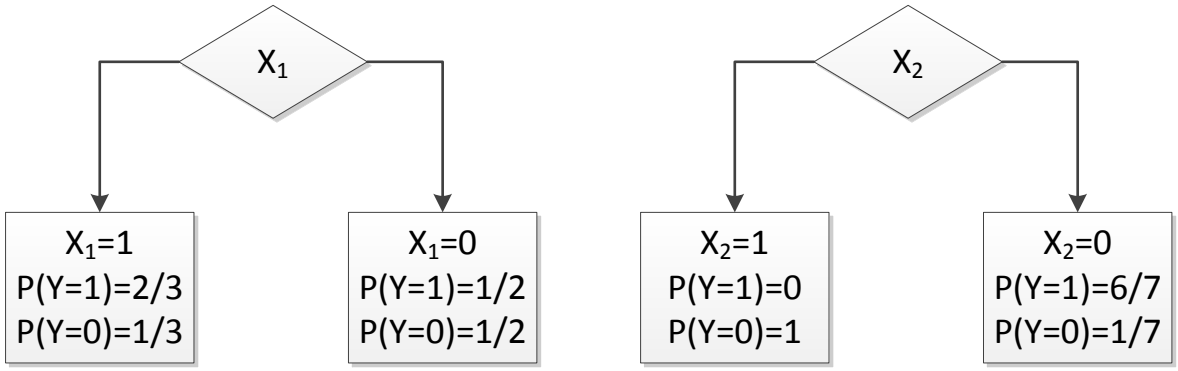
Algorytm wymaga zdefiniowania dwóch elementów:

- Określenia procedury wyboru atrybutu do podziału: $select(n)$
- Zdefiniowania kryterium stopu: $stop(n)$

Wybór atrybutu do podziału

X_1	X_2	Y
1	0	1
0	0	1
1	0	1
1	0	1
0	0	1
1	0	0
0	1	0
1	1	0
0	1	0
1	0	1

- Atrybut X_1
  - $X_1 = 1 \rightarrow Y = 4 \times 1{,}2 \times 0$
  - $X_1 = 0 \rightarrow Y = 2 \times 1{,}2 \times 0$
- Atrybut X_2
  - $X_2 = 1 \rightarrow Y = 0 \times 1{,}3 \times 0$
  - $X_2 = 0 \rightarrow Y = 6 \times 1{,}1 \times 0$



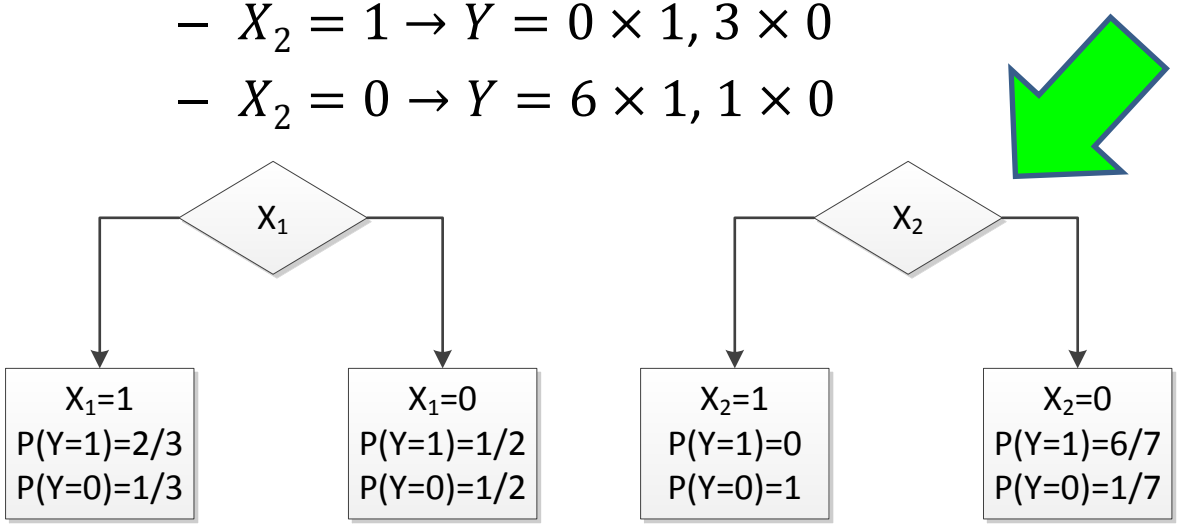
Podział, w wyniku którego wartości wyjściowe są **nierównomiernie rozłożone** jest lepszy, ponieważ zwiększa wiarygodność klasyfikacji.

- Pożądane: wszystkie Y=0 lub Y=1
- Niepożądane: równomierny rozkład wartości

Wybór atrybutu do podziału

X_1	X_2	Y
1	0	1
0	0	1
1	0	1
1	0	1
0	0	1
1	0	0
0	1	0
1	1	0
0	1	0
1	0	1

- Atrybut X_1
  - $X_1 = 1 \rightarrow Y = 4 \times 1{,}2 \times 0$
  - $X_1 = 0 \rightarrow Y = 2 \times 1{,}2 \times 0$
- Atrybut X_2
  - $X_2 = 1 \rightarrow Y = 0 \times 1{,}3 \times 0$
  - $X_2 = 0 \rightarrow Y = 6 \times 1{,}1 \times 0$



Podział, w wyniku którego wartości wyjściowe są **nierównomiernie rozłożone** jest lepszy, ponieważ zwiększa wiarygodność klasyfikacji.

- Pożądane: wszystkie Y=0 lub Y=1
- Niepożądane: równomierny rozkład wartości

Entropia

- Niech X będzie zmienną losową przyjmującą k dyskretnych wartości, każdą z nich z prawdopodobieństwem $P(X = x_k)$
- Entropia zdefiniowana jest jako:

$$H(X) = -\sum_{i=1}^{k} P(X = x_k) \log_2 P(X = x_k)$$

- Entropia (mierzona w bitach) podaje minimalną liczbę bitów niezbędną do zakodowania losowo wybranej wartości X.
- Jeśli $P = 0$, przyjmuje się, że $0 \cdot \log_2 0 \ = \ 0$

Symbol	Częstość
A	0.17
B	0.33
C	0.55

- $H = -(0.17 \cdot \log_2(0.17) + 0.33 \cdot \log_2(0.33) + 0.5 \cdot \log_2(0.5)) = 1.46$
- Proponowane kodowanie (średnio 1.66 bitu/symbol):
  A: 0
  B: 10
  C: 11

Entropia warunkowa

Entropia warunkowa obliczana jest dla dwóch zmiennych Y i X

$$H(Y|X) = -\sum_{i=1}^{k}\sum_{j=1}^{r} P\big(Y = yi, X = x_j\big) \log_2 \frac{P\big(Y = y_i, X = x_j\big)}{P\big(X = x_j\big)}$$

przekształcając:

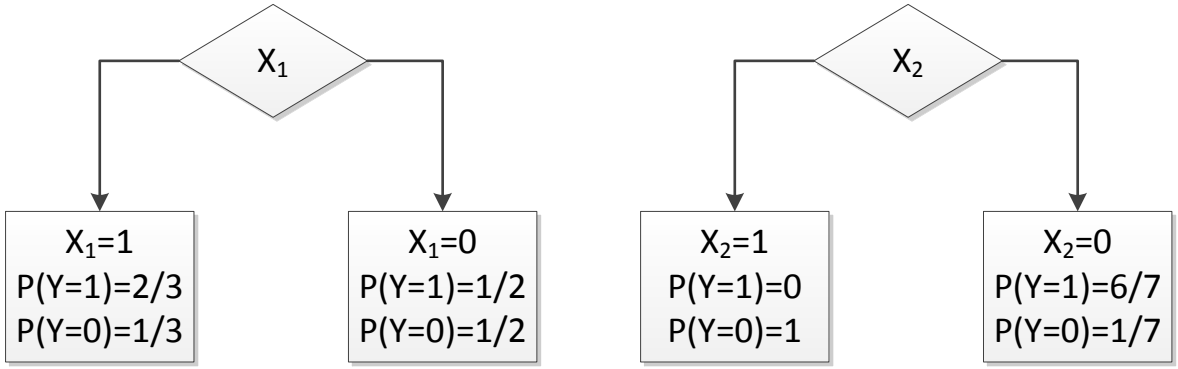
$$H(Y|X) = -\sum_{i=1}^{k}\sum_{j=1}^{r} P\big(Y = y_i|X = x_j\big)\, P(X = x_j) \log_2 P(Y = y_i|X = x_j)$$

$$H(Y|X) = -\sum_{j=1}^{r} P(X = x_j) \sum_{i=1}^{k} P\big(Y = y_i|X = x_j\big)\, \log_2 P(Y = y_i|X = x_j)$$

- Wydzielany jest podzbiór $X \times Y$, gdzie $X = x_j$
- Obliczane jest entropia w tym podzbiorze
- Obliczana jest suma z wagami $P(X = x_j)$

Przykład

X_1	X_2	Y
1	0	1
0	0	1
1	0	1
1	0	1
0	0	1
1	0	0
0	1	0
1	1	0
0	1	0
1	0	1



Entropia przed podziałem:

$$H(Y) = 0.6\log_2(0.6) + 0.4\log_2(0.4) = 0.97$$

Entropia warunkowa dla podziału X_1

$$H(Y|X_1) = 0.6 \cdot 0.92 + 0.4 \cdot 1 = 0.95$$

Zmiana entropii $IG = 0.97 - 0.95 = 0.2$

Entropia warunkowa dla podziału X_2

$$H(Y|X_2) = 0.3 \cdot 0 + 0.7 \cdot 0.41 = 0.41$$

Zmiana entropii $IG = 0.97 - 0.41 = 0.56$

Zysk informacyjny

- Zysk informacyjny podziału (ang. information gain) zdefiniowany jest jako zmiana entropii uzyskana w wyniku podziału:

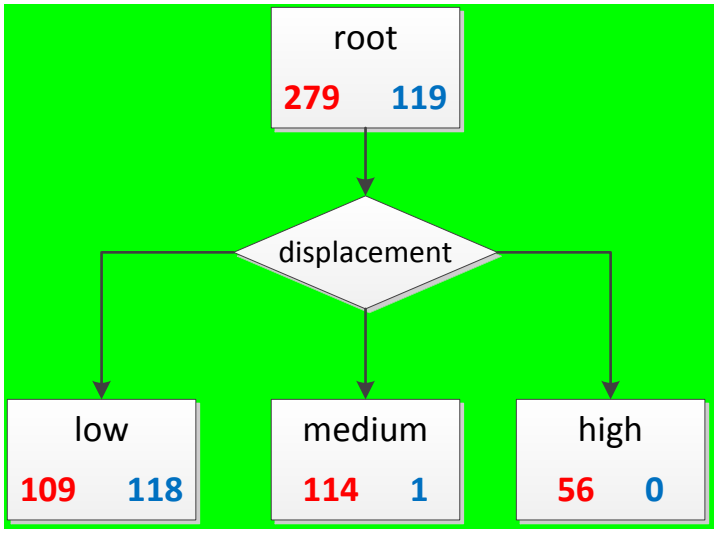
$$IG(X_i) = H(Y) - H(Y|X_i)$$

- Zawsze $IG(X_i) \geq 0$

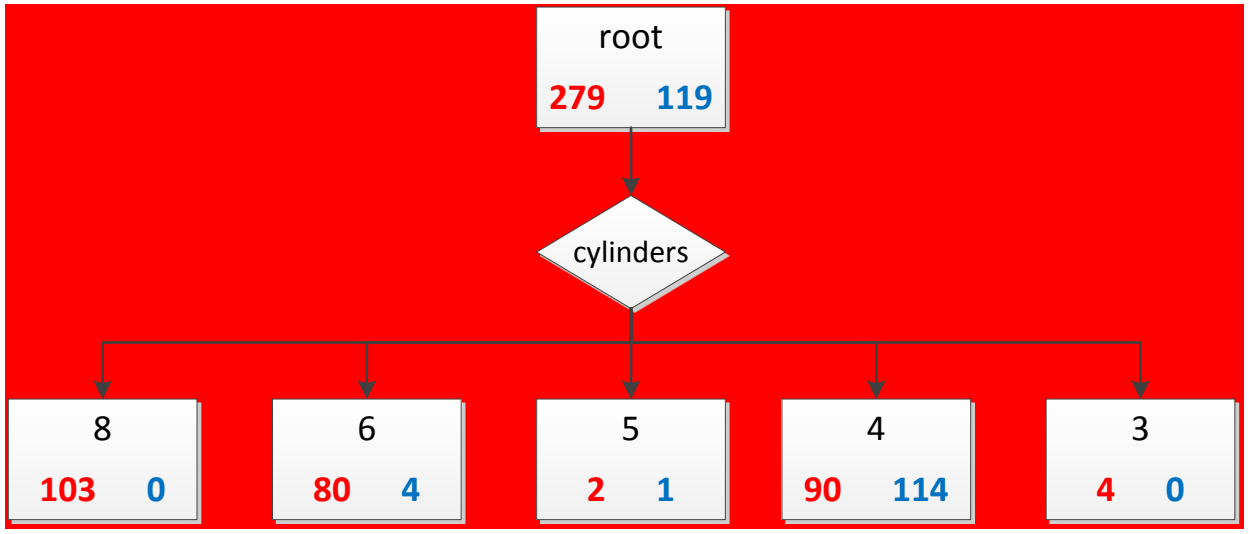
Wybór atrybutu do podziału:

$$X_i^* = \arg\max_{X_i} IG(X_i)$$

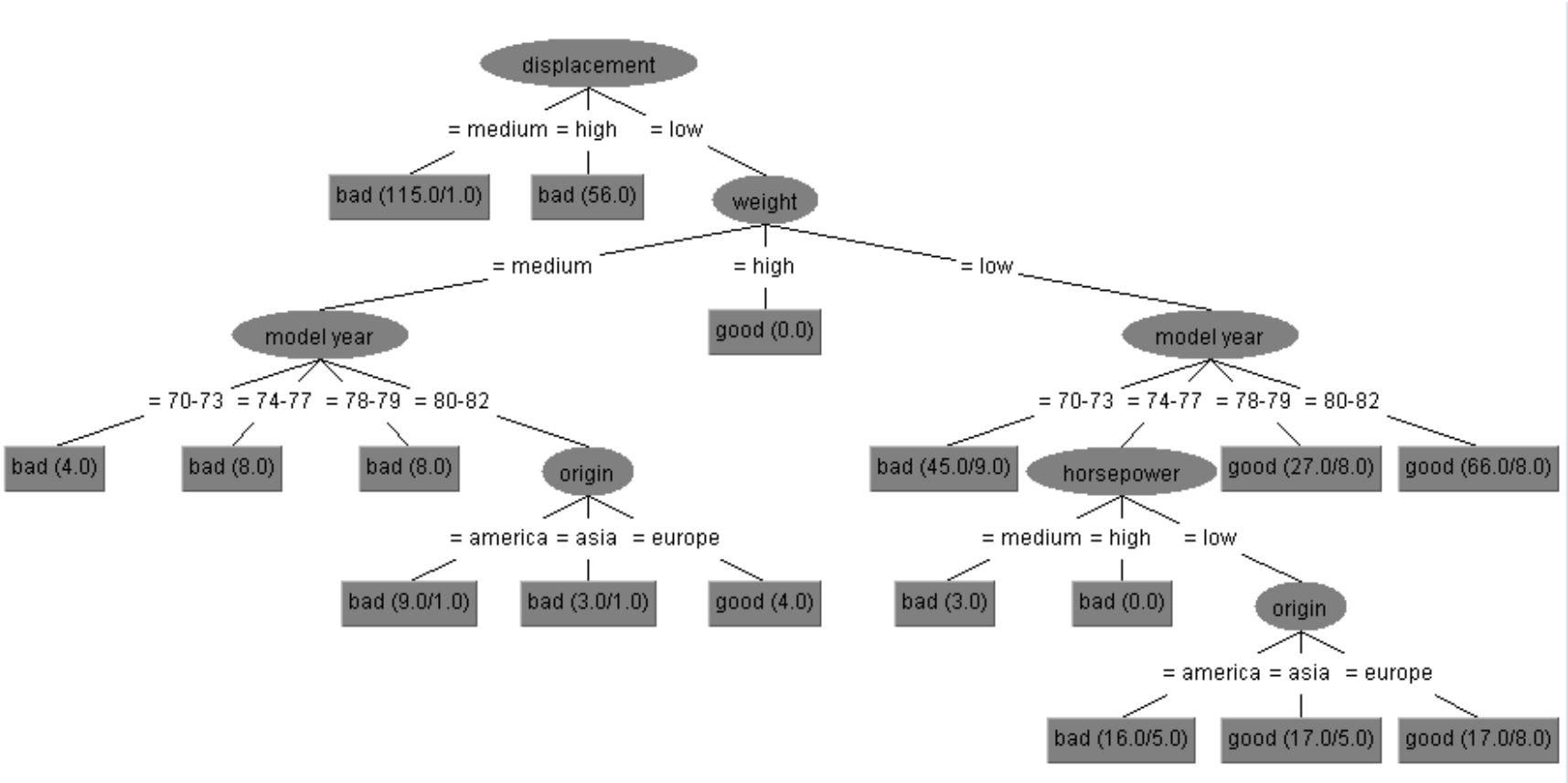
Przykład mpg



Atrybut	IG
cylinders	0.31
displacement	0.29
horsepower	0.19
weight	0.30
acceleration	0.06
model year	0.21
origin	0.15



Przykład mpg



Mimo, że kryterium IG wskazywało atrybut cylinders, algorytm J48 (C4.5) pakietu Weka wybrał podział dla displacement.

Stosowane jest nieco inne kryterium: gain ratio (stosunek zysku)

Stosunek zysku (gain ratio)

- Wskaźnik IG (information gain) preferuje atrybuty, które mają wiele wartości.
- W skrajnym przypadku najlepszym może okazać się atrybut typu ID (numer rekordu). Wówczas po podziale w każdym węźle będzie dokładnie jedna obserwacja, stąd $H(Y|ID) = 0$
- Sam podział też wnosi zawartość informacyjną. Współczynnik SI (spilit information) reprezentuje potencjalną informację wygenerowaną przez podział.
- Oznaczając przez r – liczbę wartości atrybutu X_i współczynnik SI jest zdefiniowany jako:

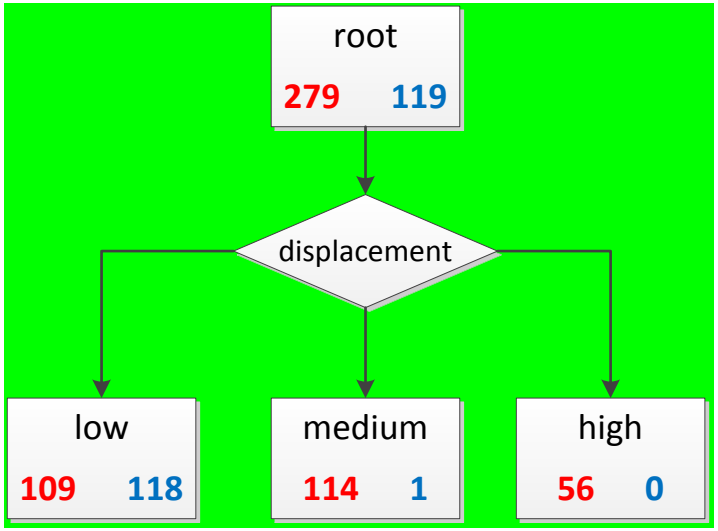
$$SI(X_i) = -\sum_{j=1}^{r} \frac{|D_j|}{|D|} \log_2 \frac{|D_j|}{|D|}$$

- Stosunek zysku (gain ratio):

$$GR(X_i) = \frac{IG(X_i)}{SI(X_i)}$$

Porównanie dla mpg

Atrybut	IG	GR
cylinders	0.31	0.19
displacement	0.29	0.21
horsepower	0.19	0.16
weight	0.30	0.20
acceleration	0.06	0.05
model year	0.21	0.11
origin	0.15	0.12



Wskaźnik Gini

- Wskaźnik Gini (Gini index) jest miarą nieczystości danych uzyskanych w wyniku podziału

$$Gini(D) = \sum_{i=1}^{k} \big(1 - p(i)\big) p(i) = 1 - \sum_{i=1}^{k} p(i)^2$$

k – liczba klas

$p(i)$ – prawdopodobieństwo, że obserwacja z zbioru D należy do i-tej klasy c_i

- Stosuje się do podziałów binarnych (na dwa zbiory D_1 i D_2):

$$Gini(D, X_i) = \frac{|D_1|}{|D|} Gini(D_1) + \frac{|D_2|}{|D|} Gini(D_2)$$

D_1, D_2 – podzbiory D wyznaczone z podziału zgodnie z atrybutem X_i

- Zmiana indeksu:

$$\Delta Gini(D, X_i) = Gini(D) - Gini(D, X_i)$$

- Wybierany jest atrybut X_i maksymalizujący zmianę $\Delta Gini(D, X_i)$

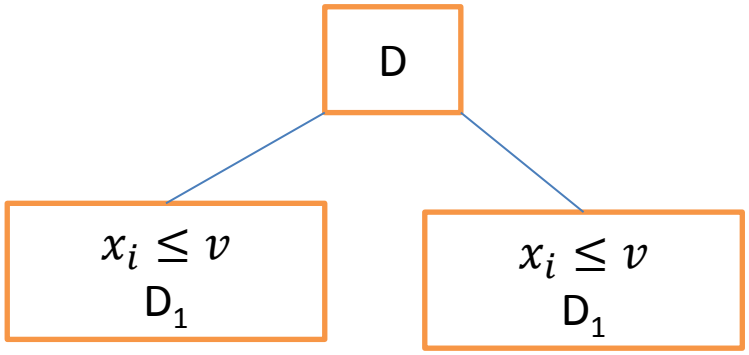
Trzy wskaźniki wyboru atrybutu do podziału

- IG (information gain) – algorytm ID3
- GR (gain ratio) – algorytm C4.5 (J48 Weka)
- Gini – algorytm CART

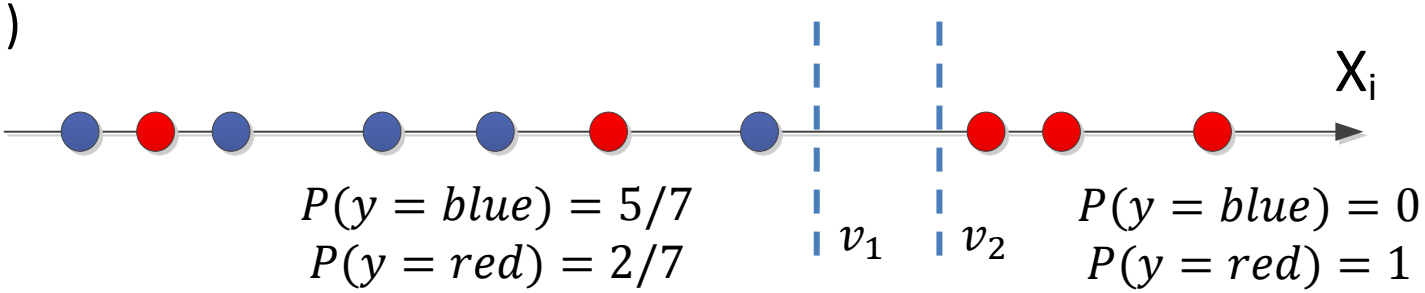
Atrybuty numeryczne

- W przypadku atrybutów numerycznych stosuje się podział binarny.
  - $D_1 = \{(x, y) \in D : x_i \leq v\}$
  - $D_2 = \{(x, y) \in D : x_i > v\}$

  i – indeks atrybutu X_i

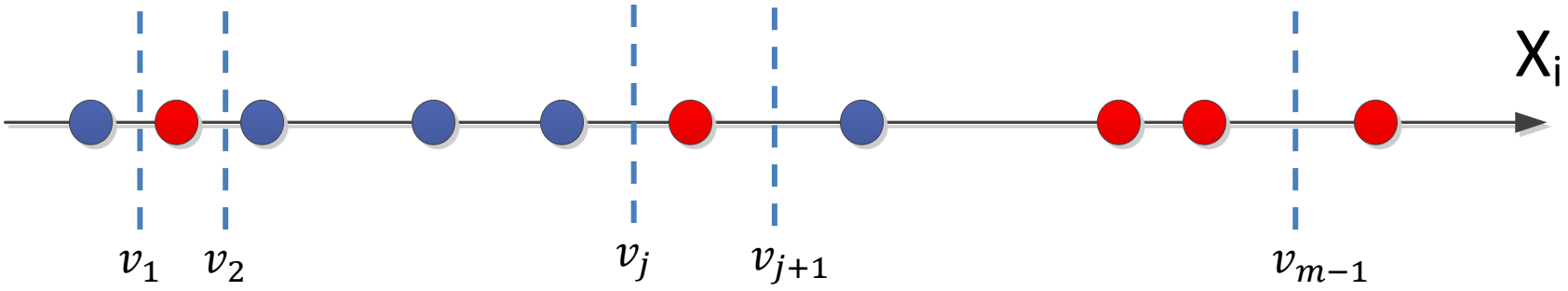


- Próg podziału v może być wybrany, na nieskończenie wiele sposobów, jednak z punktu widzenia miar jakości podziału liczy się położenie pomiędzy obserwacjami na osi X_i (posortowanymi wg. wartości atrybutu X_i)



- Dla v_1 i v_2 prawdopodobieństwa wynikające z podziału identyczne

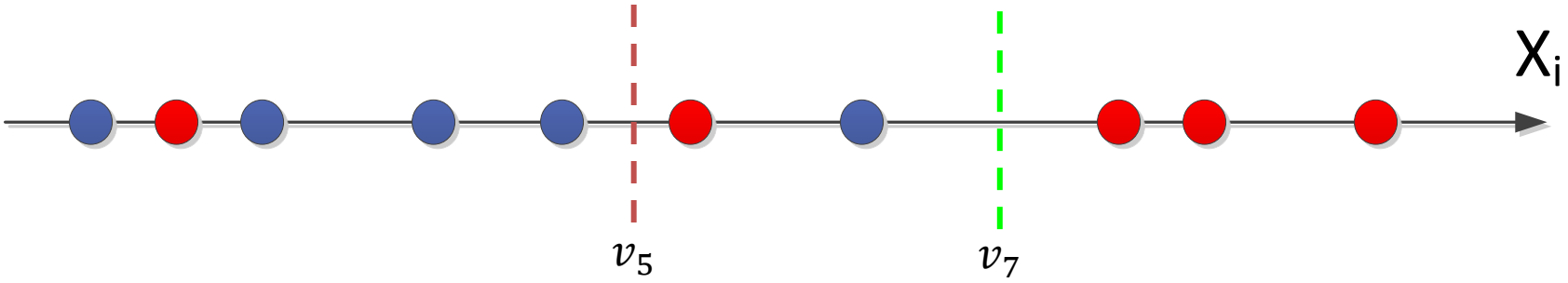
Atrybuty numeryczne



- Jako punkty podziału wybierane są wartości środkowe pomiędzy posortowanymi według wartości X_i obserwacjami w D
- Dla m obserwacji w zbiorze D możliwych jest $m-1$ punktów

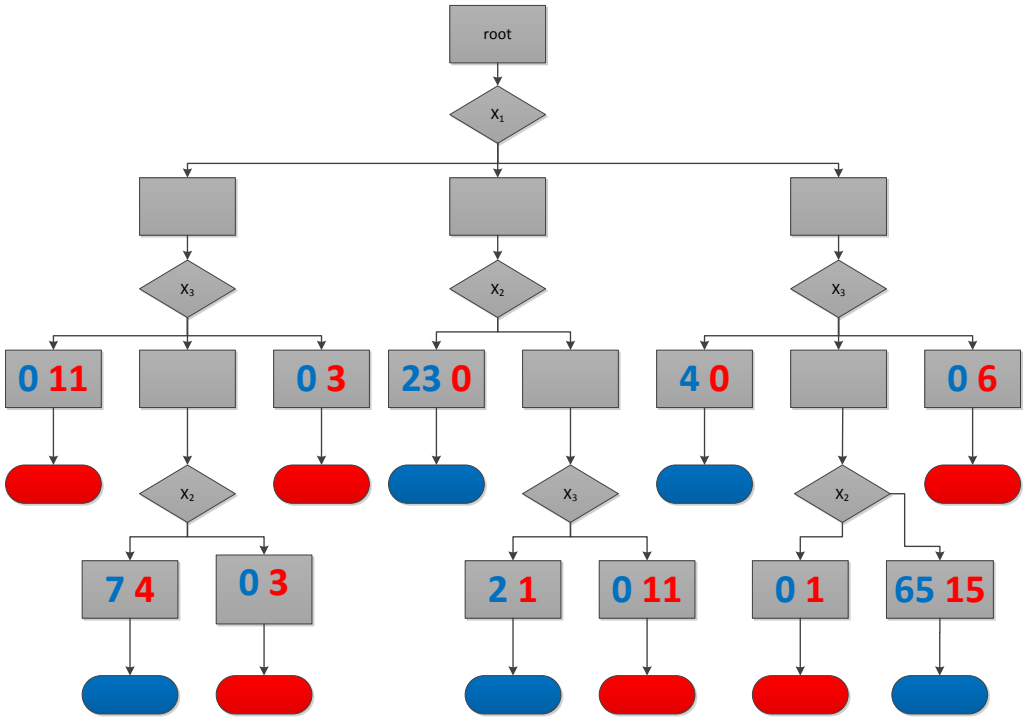
  podziału: $v_j = \frac{x_i^j + x_i^{j+1}}{2}, j = 1, \ldots, m-1$
- Dla każdego z nich obliczany jest wskaźnik oceny jakości podziału i wybierany najlepszy

Atrybuty numeryczne



Podział	$x_i \leq v_i$ $P(\boldsymbol{red})$	$x_i \leq v_i$ $P(blue)$	$x_i > v_i$ $P(\boldsymbol{red})$	$x_i > v_i$ $P(blue)$	Info Gain	Gain Ratio	Gini
1	0.00	1.00	0.56	0.44	0.11	0.23	0.06
2	0.50	0.50	0.50	0.50	0.00	0.00	0.00
3	0.33	0.67	0.57	0.43	0.03	0.04	0.04
4	0.25	0.75	0.67	0.33	0.12	0.13	0.13
5	0.20	0.80	0.80	0.20	0.28	0.28	0.27
6	0.33	0.67	0.75	0.25	0.12	0.13	0.13
7	0.29	0.71	1.00	0.00	0.40	0.45	0.25
8	0.38	0.63	1.00	0.00	0.24	0.33	0.14
9	0.44	0.56	1.00	0.00	0.11	0.23	0.06

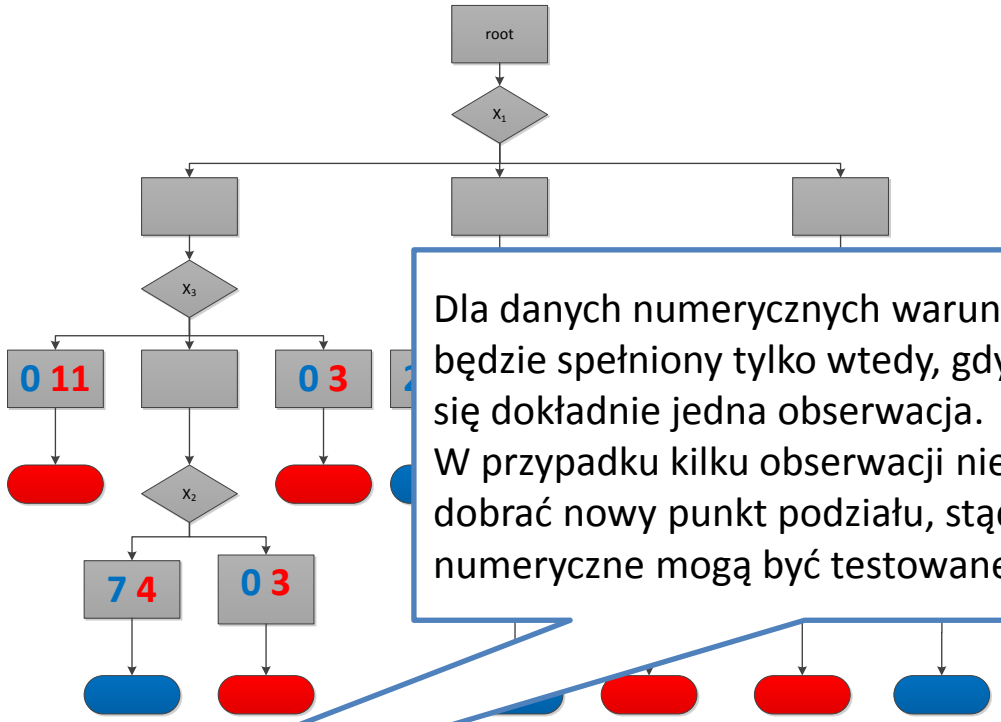
Kryteria stopu



Kryteria stopu decydują, kiedy gałąź z danego węzła n nie będzie rozwijana

1. Jeżeli wszystkie obserwacje w zbiorze $D(n)$ mają tę samą wartość y
2. Jeżeli wszystkie obserwacje mają identyczne wartości atrybutów. (Brak jednoznacznej klasyfikacji wynika z szumu danych.)
3. Potencjalne kryterium – brak przyrostu wskaźnika zysku dla każdego możliwego podziału (np. IG = 0)

Kryteria stopu



Dla danych numerycznych warunek 2. na ogół będzie spełniony tylko wtedy, gdy w węźle znajdzie się dokładnie jedna obserwacja.
W przypadku kilku obserwacji niemal zawsze można dobrać nowy punkt podziału, stąd atrybuty numeryczne mogą być testowane wielokrotnie.

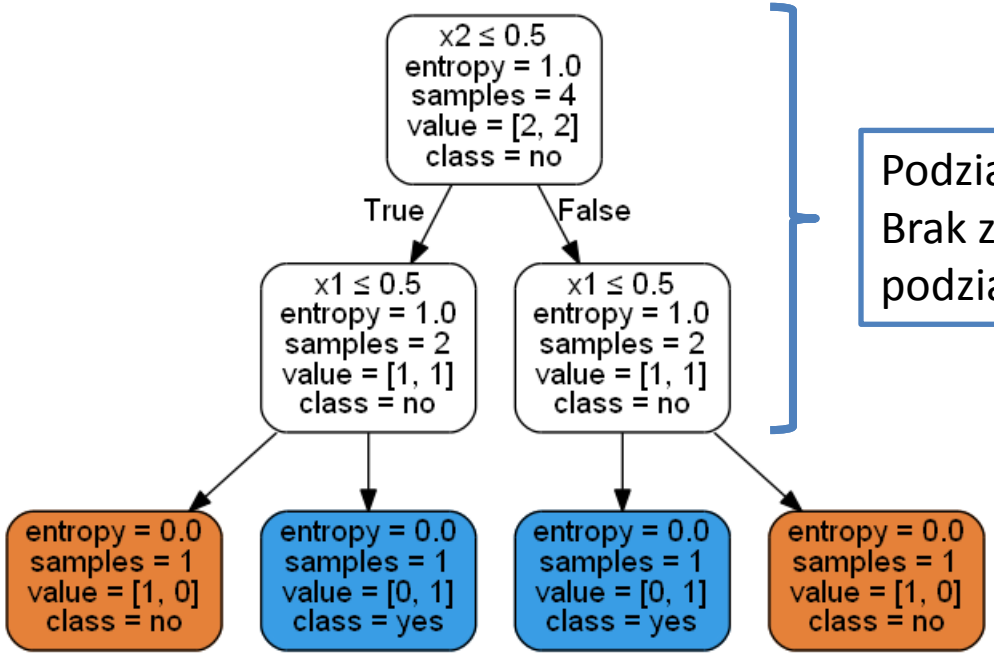
Kryteria stopu decydu[illegible]edy gałąź z danego węzła n nie będzie rozwijana

1. Jeżeli ws[illegible]stkie obserwacje w zbiorze $D(n)$ mają tę samą wartość y
2. Jeżeli wszystkie obserwacje mają identyczne wartości atrybutów. (Brak jednoznacznej klasyfikacji wynika z szumu danych.)
3. Potencjalne kryterium – brak przyrostu wskaźnika zysku dla każdego możliwego podziału (np. IG = 0)

Brak przyrostu wskaźnika zysku?

X_1	X_2	Y
1	1	no
1	0	yes
0	1	yes
0	0	no

- Wybór X_1 lub X_2, jako atrybutu do podziału nie przynosi żadnego zysku
- Jednakże w kolejnych 2 krokach po dokonaniu podziału można uzyskać idealne dopasowanie

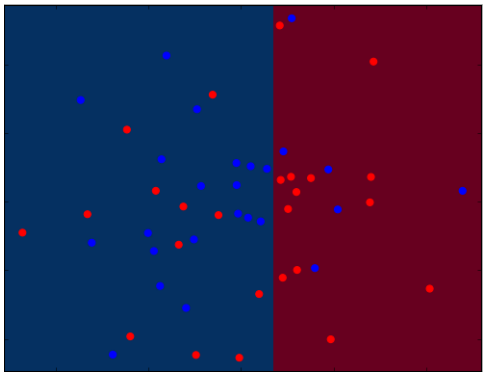


Podział zgodnie z warunkiem $x_2 \leq 0.5$
Brak zysku: $entropy = 1.0$ przed i po podziale

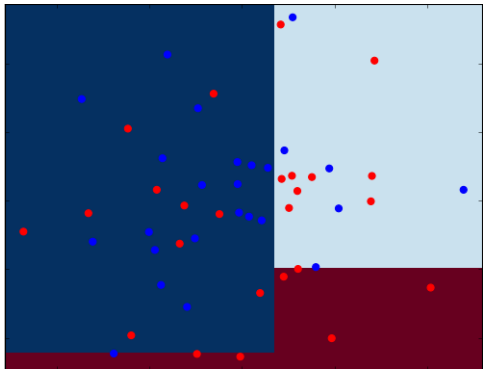
Nadmierne dopasowanie

- Drzewa decyzyjne są bardzo podatne na zjawisko nadmiernego dopasowania.
- Wraz ze wzrostem głębokości drzewa są w stanie bardzo dobrze dostosować się do danych uczących
- Równocześnie pogarsza się ich zdolność generalizacji (błąd testowy rośnie)

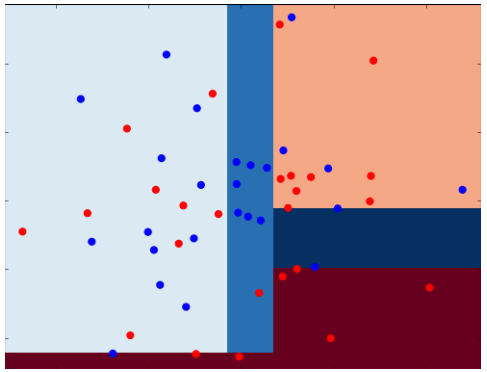
Przykład – różne głębokości drzewa



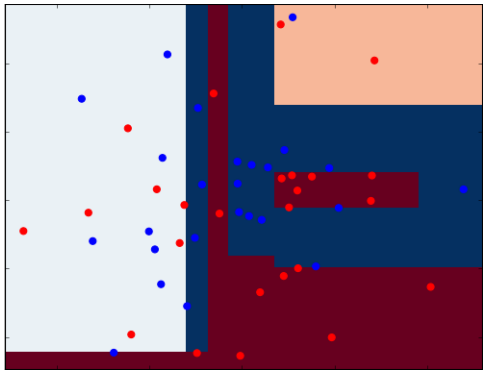
depth=1 (decision stump)
$e_{train} = 0.36$
$e_{test} = 0.34$



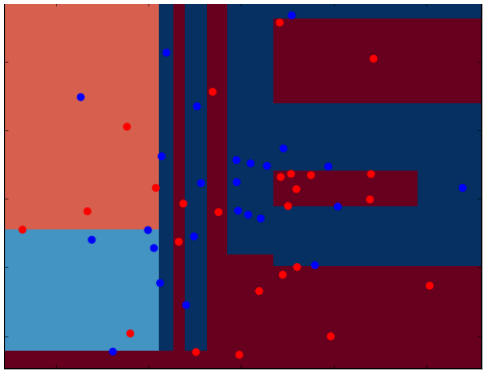
depth=2
$e_{train} = 0.32$
$e_{test} = 0.36$



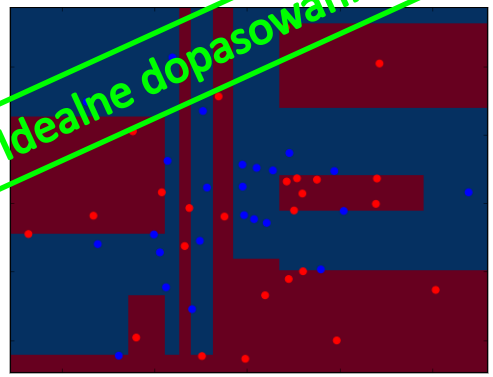
depth=3
$e_{train} = 0.28$
$e_{test} = 0.36$



depth=5
$e_{train} = 0.16$
$e_{test} = 0.48$

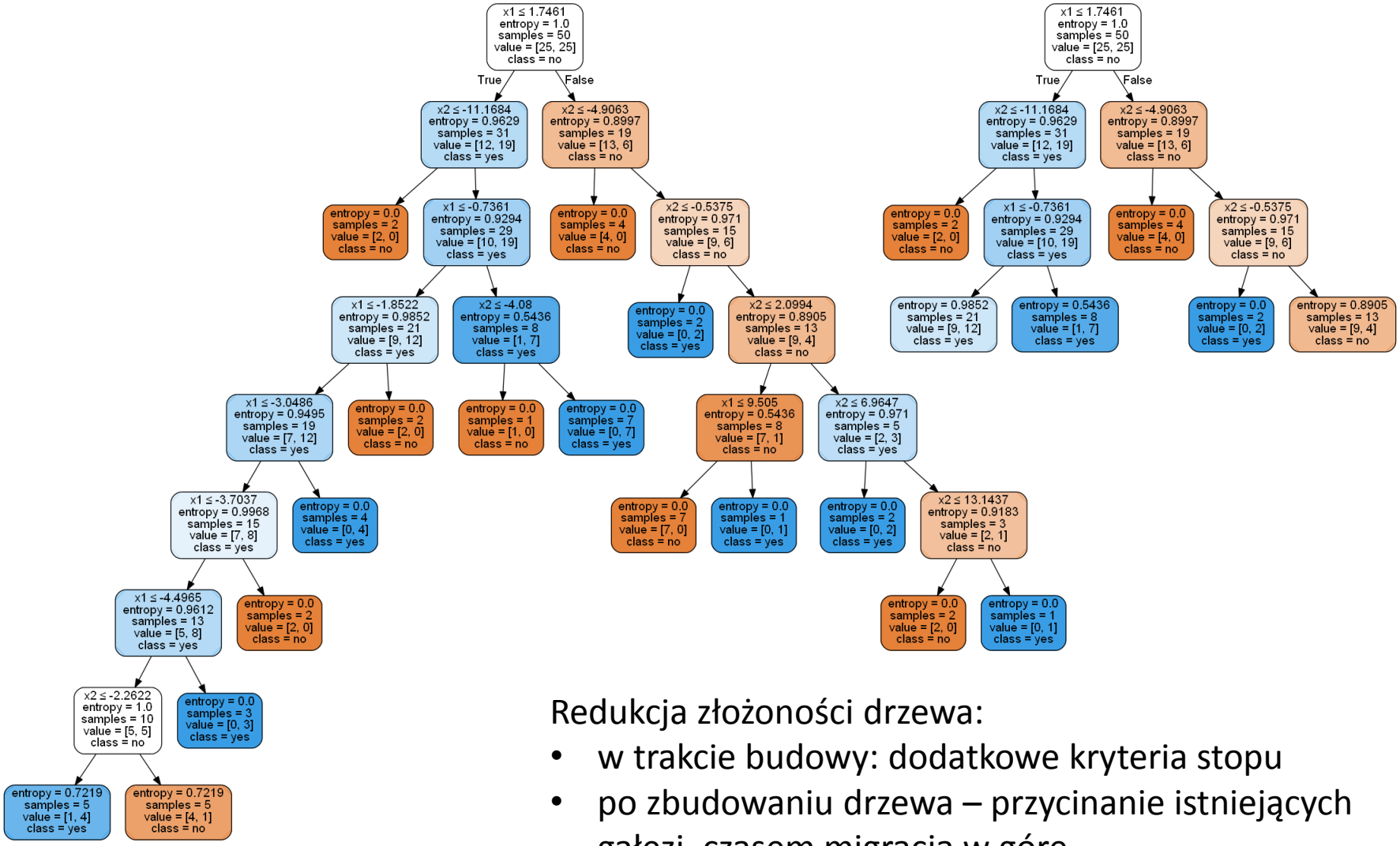


depth=8
$e_{train} = 0.04$
$e_{test} = 0.56$



depth=10
$e_{train} = 0.0$
$e_{test} = 0.54$

Redukcja złożoności drzewa



Redukcja złożoności drzewa:

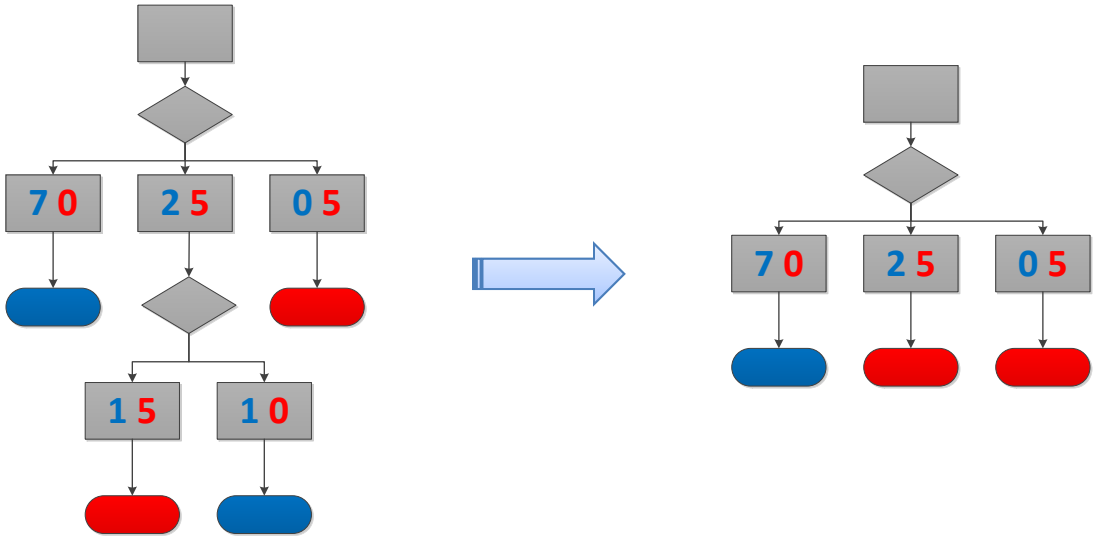
- w trakcie budowy: dodatkowe kryteria stopu
- po zbudowaniu drzewa – przycinanie istniejących gałęzi, czasem migracja w górę

Dodatkowe kryteria stopu

- **Ograniczenie głębokości drzewa** (parametr max_depth)
  - Dobór eksperymentalny: testowanie / walidacja krzyżowa
- **Brak poprawy trafności**: nie dziel węzła, jeśli w wyniku podziału błąd klasyfikacji nie spadnie.
  - Nie zawsze właściwe, wcześniejszy przykład XOR
- **Dolne ograniczenie liczby obserwacji** przypisanych do węzła, np. nie dziel węzła, jeżeli jest ona mniejsza niż 2

Przycinanie

- Etap przycinania drzewa następuje po jego zbudowaniu
- Polega ono na zastąpieniu węzła liściem decyzją wybraną na podstawie klasy większościowej
- Kryteria:
  - test chi kwadrat
  - ograniczenie liczby węzłów w drzewie



Przycinanie: test chi kwadrat

- Począwszy od dołu drzewa dla każdego węzła przeprowadzany jest test χ^2, aby określić prawdopodobieństwo $p(n)$, że widoczna korelacja pomiędzy wartościami atrybutu i zmiennej decyzyjnej była przypadkowa
- Jeśli $p(n) > p_{\max}$ (konfigurowlny parametr) węzeł jest usuwany i zastępowany liściem
- Proces jest kontynuowany, dopóki istnieją węzły spełniające warunek $p(n) > p_{\max}$

Ograniczenie liczby węzłów w drzewie

- Do oceny drzewa stosuje się funkcję agregującą dwa czynniki:
  - błąd klasyfikacji $err(T, D)$
  - liczbę liści drzewa $leafs(T)$
- Funkcja oceny:

$$J(T, D) = err(T, D) + \lambda \cdot \ leafs(T)$$

- Algorytm:
  1. Wybierz nieodwiedzony węzeł n u dołu drzewa
  2. Oblicz $J(T, D)$ dla pełnego drzewa T
  3. Oblicz $J(T', D)$ dla drzewa $T' = T \backslash \{n\}$
  4. Jeśli $J(T', D) < \ J(T, D)$ przypisz: $T \leftarrow \ T'$ i przejdź do 1
  5. Zaznacz węzeł n jako odwiedzony i przejdź do 1
- W zależności od podejścia, zbiór D użyty w przycinaniu może być zbiorem uczącym lub walidacyjnym (preferowane)