

ZeroR

- Odpowiada zawsze tak samo
- Decyzja to klasa większościowa ze zbioru uczącego

A	B	X
1	5	T
1	7	T
1	5	T
1	5	F
2	7	F

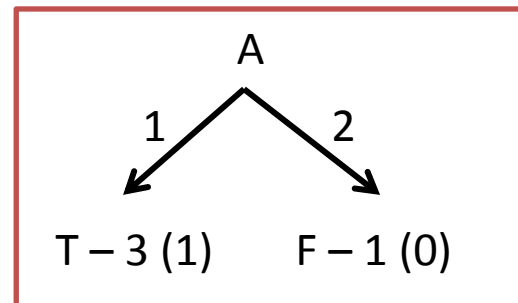
Tutaj jest więcej obiektów klasy T, więc klasyfikator ZeroR będzie zawsze odpowiadał T niezależnie od danych wejściowych

One rule model (OneR)

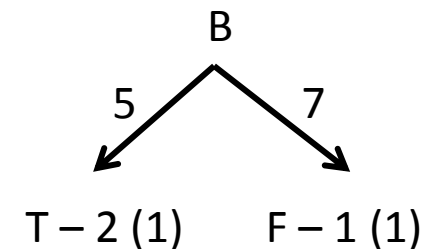
- Bardzo prosty mechanizm klasyfikacji
- Tworzy 1-dno poziomowe drzewo decyzyjne.

A	B	X
1	5	T
1	7	T
1	5	T
1	5	F
2	7	F

		X	
		T	F
A	1	3	1
	2	0	1

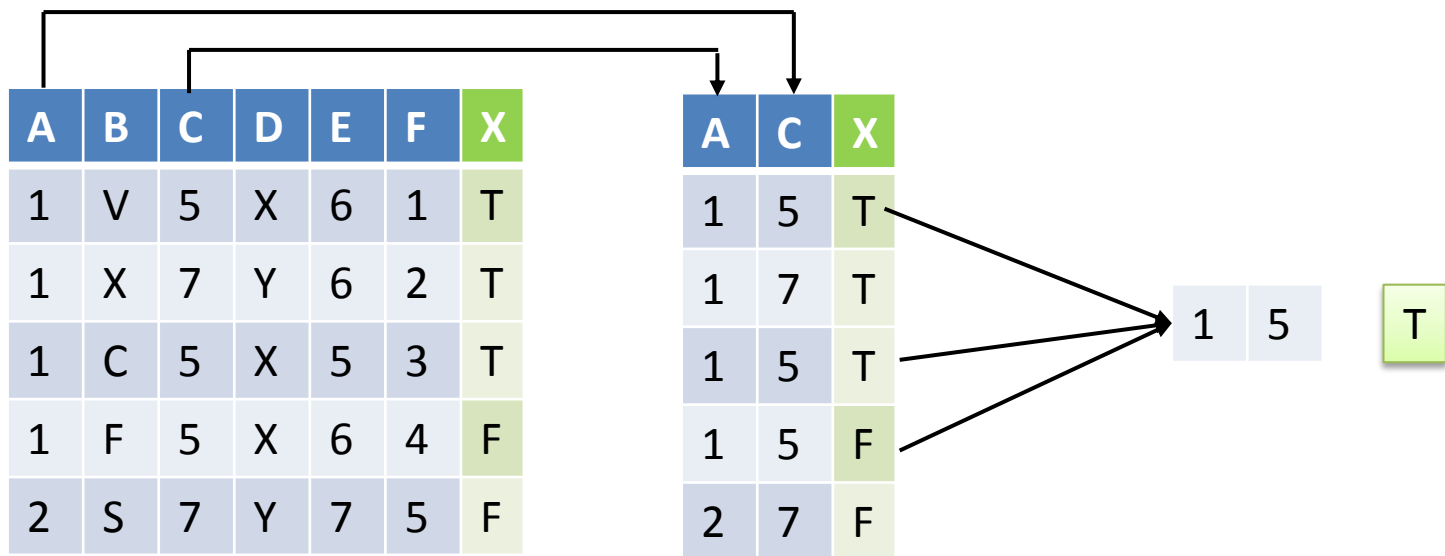


		X	
		T	F
B	5	2	1
	7	1	1



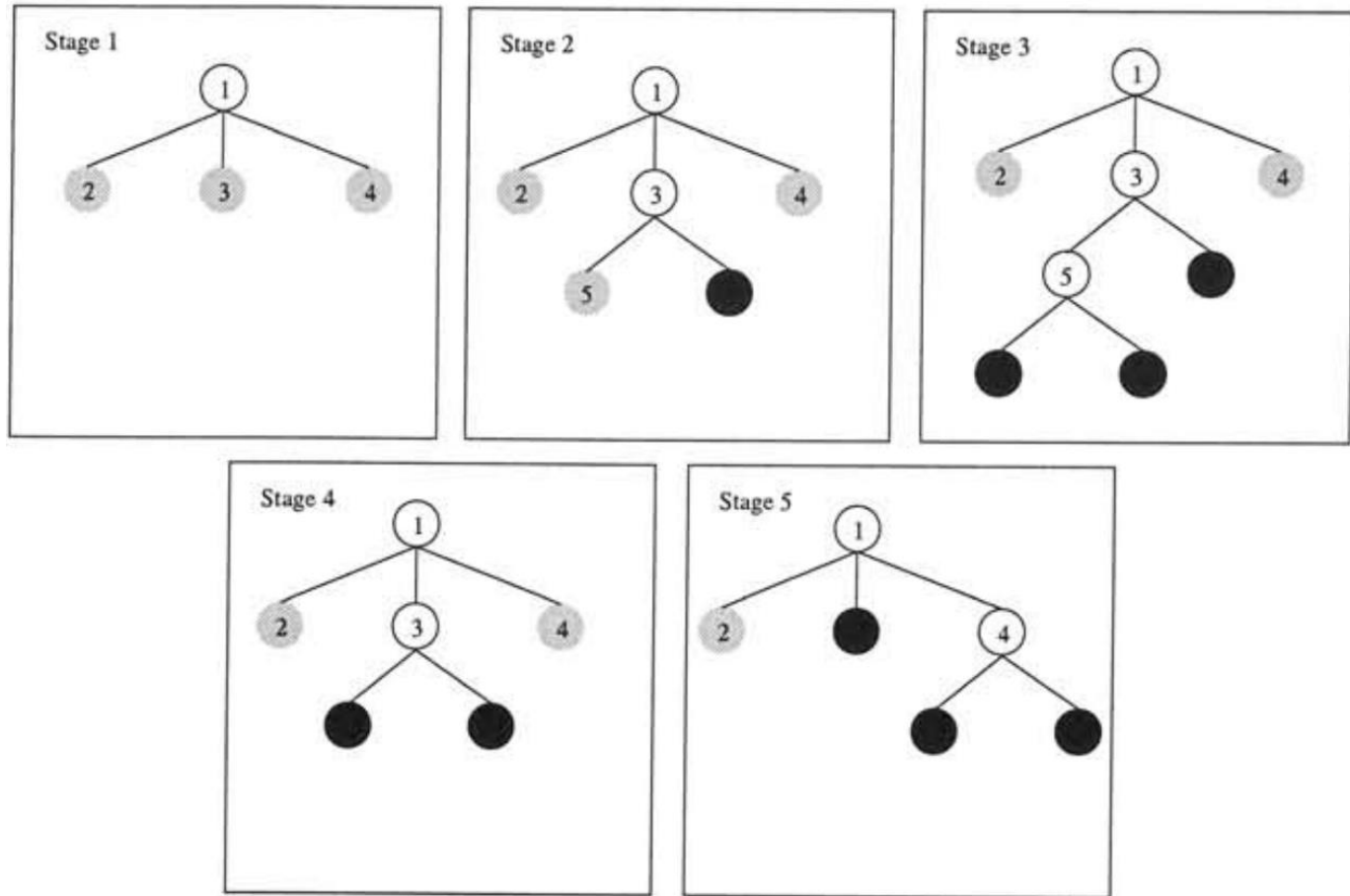
Decision table

- Mechanizm klasyfikacji trochę podobny do OneR i iBk
- Wykorzystuje „okrojone” przykłady uczące do klasyfikacji



PART (1)

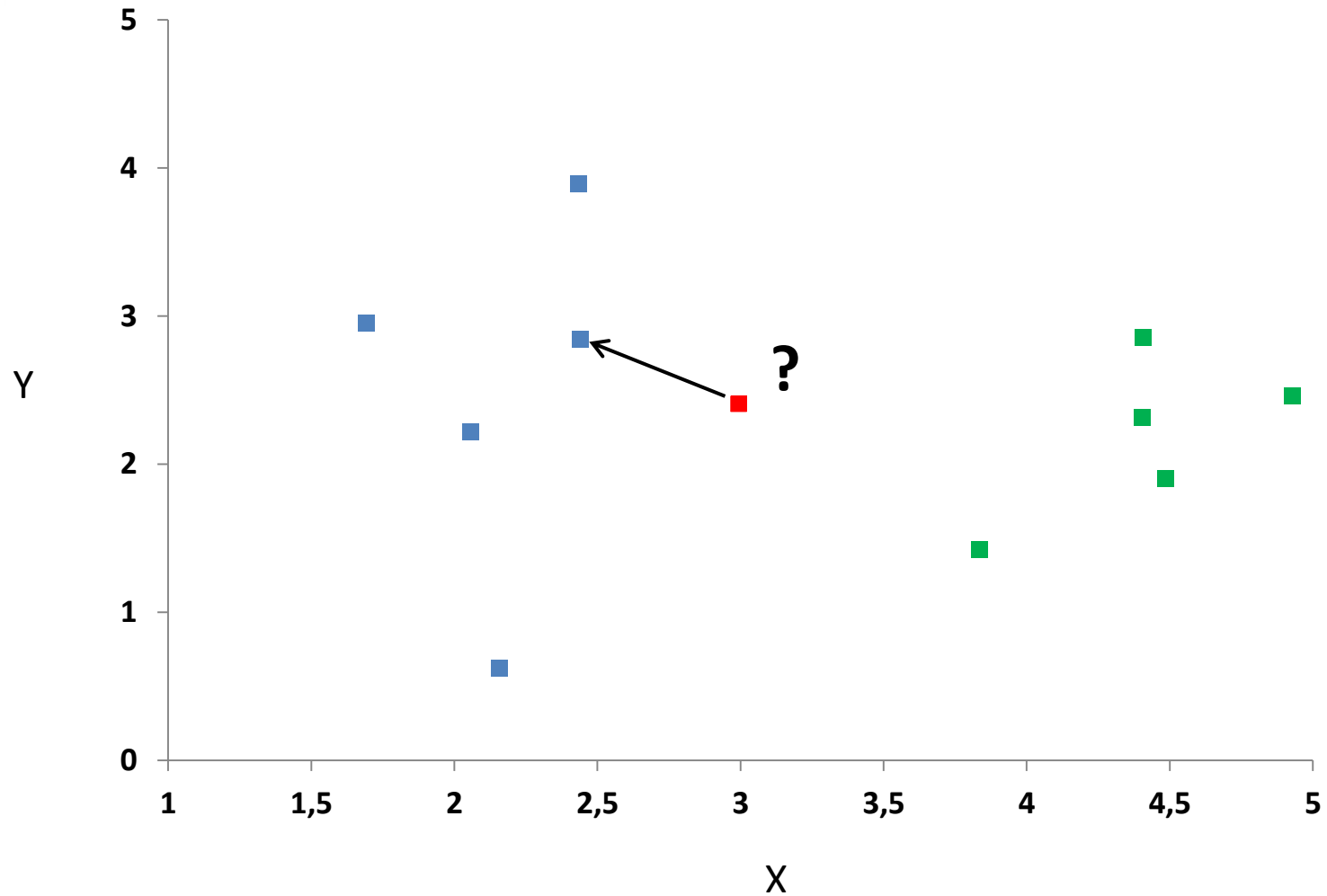
- Skomplikowany algorytm generujący reguły decyzyjne.



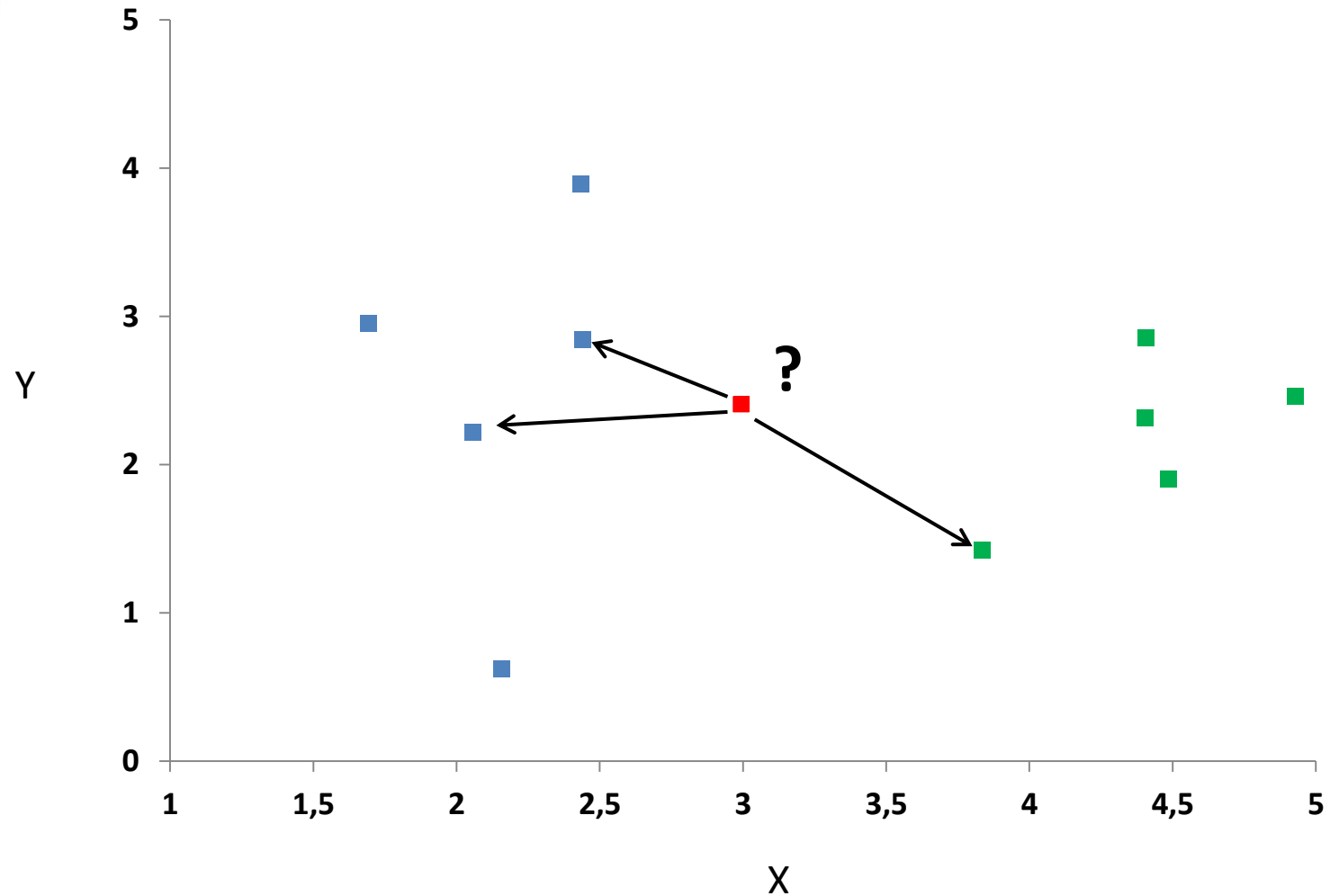
PART (2)

- Wygenerowanie jednej reguły działa następująco:
 1. Zaczynamy budowę drzewa tak jak normalnie się to robi, np. w C4.5
 2. Wybieramy liść w którym zbiór przykładów ma najmniejszą entropię
 3. Rekurencyjnie budujemy drzewo dalej.
 4. Kiedy dany węzeł w którym się znajdujemy posiada jedynie dzieci obcinamy go stosując standardowe algorytmy obcinania węzłów drzewa.
 5. Powtarzamy obcinanie aż jakieś obcięcie nie zostanie wykonane.
 6. Z powstałych liści wybieramy taki, który opisuje największą liczbę przykładów i tworzymy z niego regułę.
 7. Usuwamy przykłady pokryte przez utworzoną regułę ze zbioru i kontynuujemy aż pokryjemy wszystkie przykłady bądź utworzymy zadaną liczbę reguł.

IB1



$$IBk/K^*$$



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

- (ang. confusion matrix)

		Stan faktyczny		Razem
		chory	zdrowy	
Wynik klasyfikacji	chory (wynik testu pozytywny)	TP prawidłowa decyzja pozytywna	FP błędna decyzja pozytywna (błąd I rodzaju)	P'
	zdrowy (wynik testu negatywny)	FN błędna decyzja negatywna (błąd II rodzaju)	TN prawidłowa decyzja negatywna	N'
Razem		P	N	

Miary oceny klasyfikatora

- **trafność** (accuracy)

$$ACC = \frac{TP + TN}{P + N}$$

- **miara F** (F1 score, F-measure)

$$F1 = 2 \frac{PPV * TPR}{PPV + TPR}$$

Miary oceny klasyfikatora

– **czułość** (sensitivity, true positive rate, **recall**)

prawdopodobieństwo, że test dla osoby chorej da wynik pozytywny (tzn. że jest chora)

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN}$$

– **specyficzność** (specificity, true negative rate)

*prawdopodobieństwo, że test dla osoby zdrowiej da wynik negatywny
(tzn. że jest zdrowa)*

$$SPC = \frac{TN}{N} = \frac{TN}{FP + TN}$$

Miary oceny klasyfikatora

– fałszywe alarmy (false positive rate)

Jaka część osób zdrowych została zaklasyfikowana jako osoby chore?

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} = 1 - SPC$$

– fałszywe odkrycia (false discovery rate)

Jaka część spośród pozytywnych wyników klasyfikacji jest fałszywa?

$$FDR = \frac{FP}{FP + TP} = 1 - PPV$$

Miary oceny klasyfikatora

- **precyzja pozytywna** (positive predictive value, **precision**)

Jakie jest prawdopodobieństwo, że badana osoba jest chora, jeżeli została sklasyfikowana jako chora?

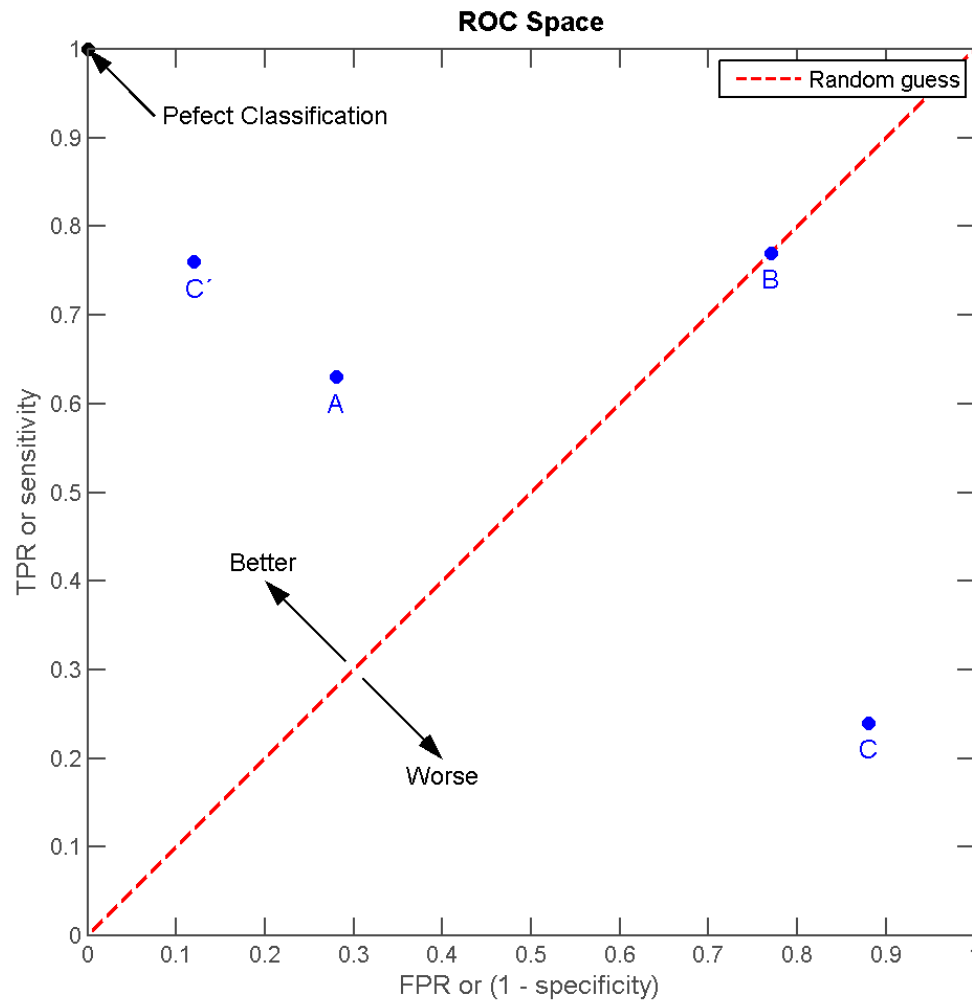
$$PPV = \frac{TP}{TP + FP}$$

- **precyzja negatywna** (negative predictive value)

Jakie jest prawdopodobieństwo, że badana osoba jest zdrowa, jeżeli została sklasyfikowana jako zdrowa?

$$NPV = \frac{TN}{TN + FN}$$

Przestrzeń ROC



Prawdopodobieństwo warunkowe (1)

- Niech Ania i Tomek grają w grę, w której każde z nich rzuca kostką. Pierwsza rzuca Ania. Ten kto uzyska więcej oczek na kostce wygrywa. Jeżeli jest remis to gra jest powtarzana.
- Niech będą dane oznaczenia:
 - A - Ania uzyskała 3 oczka
 - B - Tomek uzyskał 1 oczko
 - C - suma oczek to 8

$P(A)?$ **1/6**

$P(B)?$ **1/6**

$P(C)?$

Prawdopodobieństwo warunkowe (2)

- Możliwych pięć sytuacji dla C:
 $\{(2,6), (3,5), (4,4), (5,3), (6,2)\}$

$$P(C) = \frac{5}{6 * 6} = \frac{5}{36}$$

- $P(A \text{ i } C)?$ 1/36 (tylko sytuacja (3,5))
- $P(B \text{ i } C)?$ 0
- $P(A \text{ i } B)?$ 1/36
- Jakie jest prawdopodobieństwo, że Ania wylosowała 5, jeżeli wygrała z Tomkiem?

Prawdopodobieństwo warunkowe (3)

- Jakie jest prawdopodobieństwo, że Ania wylosowała 5 (zdarzenie A), jeżeli wygrała z Tomkiem (zdarzenie B)?

$P(B)=15/36$ $P(A \cap B)=4/36$

1, 1	2, 1	3, 1	4, 1	5, 1	6, 1
1, 2	2, 2	3, 2	4, 2	5, 2	6, 2
1, 3	2, 3	3, 3	4, 3	5, 3	6, 3
1, 4	2, 4	3, 4	4, 4	5, 4	6, 4
1, 5	2, 5	3, 5	4, 5	5, 5	6, 5
1, 6	2, 6	3, 6	4, 6	5, 6	6, 6

$P(A)=6/36$

Ania Tomek

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

$$P(A|B) = \frac{\frac{4}{36}}{\frac{15}{36}} = \frac{4}{15}$$

Twierdzenie Bayesa (1)



prawdopodobieństwo
a posteriori

prawdopodobieństwo
a priori

$$P(h | D) = \frac{P(D | h) \cdot P(h)}{P(D)}$$

hipoteza

dane

$$P(D) = \sum_{i=1}^n P(D | h_i) \cdot P(h_i), \quad \sum_{i=1}^n P(h_i) = 1, P(h_i \wedge h_j) = 0$$

Twierdzenie Bayesa (2)

- Prawdopodobieństwo *a posteriori*

$$P(D | h) = P(a_1 = v_1 \wedge a_2 = v_2 \dots a_n = v_n | h) \approx$$

$$P(a_1 = v_1 | h) \cdot P(a_2 = v_2 | h) \cdot \dots \cdot P(a_n = v_n | h)$$

warunkowa niezależność

- Szacowanie prawdopodobieństw cząstkowych

– zmienne kategoryczne

$$P(a_i = v_i | h) = \frac{|\{D \in T : a_i = v_i \wedge l = h\}|}{|T|}$$

– zmienne numeryczne

$$P(a_i = v_i | h) = \frac{1}{\sigma_A \sqrt{2\pi}} e^{\left(\frac{-(a_i - \mu_A)^2}{2\sigma_A^2}\right)}$$

Idź na całość !



Marilyn vos Savant

Idź na całość – zmienić czy nie zmienić? (1)

- Założenie: wybieramy na początku bramkę 1
C – samochód w bramce 2
D – otwarcie bramki D3

$$P(C|D) = \frac{P(D|C) * P(C)}{P(D)}$$

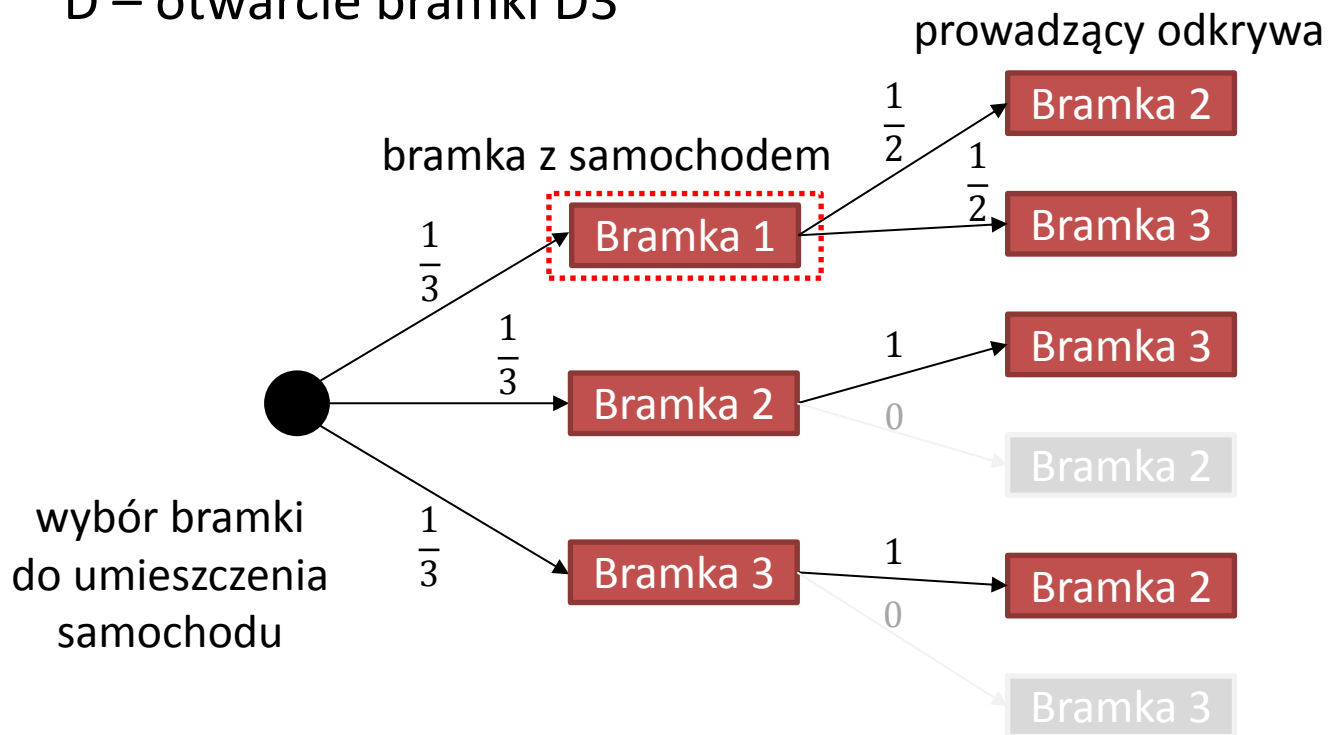
$$P(C) = \frac{1}{3}$$

$$P(D|C) = 1$$

$$P(D) = ?$$

Idź na całość – zmienić czy nie zmienić? (2)

- Założenie: wybieramy (jako gracz) na początku bramkę 1
D – otwarcie bramki D3



$$P(D) = \frac{1}{3} * \frac{1}{2} + \frac{1}{3} * 1 + \frac{1}{3} * 0 = \frac{1}{6} + \frac{1}{3} = \frac{1}{2}$$

$$P(C|D) = \frac{1 * \frac{1}{3}}{\frac{1}{2}} = \frac{1}{3} * \frac{2}{1} = \frac{2}{3}$$

NKB w akcji

Nazwa	Narodziny	Ssie mleko?	Liczba nóg	Czy lata?	Gromada
mysz	żywe	tak	4	nie	ssak
lew	żywe	tak	4	nie	ssak
wieloryb	żywe	tak	2	nie	ssak
kiwi	jajo	nie	2	nie	ptak
orzeł	jajo	nie	2	tak	ptak
bocian	jajo	nie	2	tak	ptak
<i>nietoperz</i>	<i>żywe</i>	<i>tak</i>	<i>2</i>	<i>tak</i>	<i>???</i>

Klasyfikatory regułowe (1)

- reguły: if ... then ...
- pokrycie reguły - % przykładów ze zbioru treningowego, które są pokrywane (spełniają poprzednik) przez regułę
- Jeden przykład może być pokrywany przez wiele reguł (chyba, że są to reguły rozłączne)
- Klasyfikacja nowego przypadku: Jeżeli jest pokrywany przez:
 - jedną regułę – stosujemy ją
 - wiele reguł – wybieramy regułę ustawioną najwyżej w rankingu (ranking np. po pokryciu, trafności etc.) lub stosujemy głosowanie (może być z wagami)
 - żadną z reguł – stosujemy regułę domyślną (czyli bez if'a) wskazującą na dominującą w zbiorze klasę

Klasyfikatory regułowe (2)

- Skąd reguły?
 - indukcja drzew i transformacja do reguł
 - bezpośrednia indukcja reguł
- Algorytm maksymalnego pokrycia (idea):
 - przetwarzaj kolejne klasy decyzyjne
 - w ramach danej klasy
 - szukaj najlepszej reguły
 - usuń przykłady pokrywane przez tę regułę
 - szukaj kolejnej reguły albo zatrzymaj się (brak dalszych przykładów, zbyt *słabej jakości* poprzednio odkryta reguła)
 - dodaj regułę domyślną

RIPPER

- Repeated Incremental Pruning to Produce Error Reduction)
- od ogółu do szczegółu
- tworzy ranking klas (na podstawie rosnącej liczby przykładów)
- Dla kolejnej klasy:
 - przykłady z tej klasy to przykłady pozytywne
 - przykłady z innych klas to przykłady negatywne
 - wszystkie przykłady dzielone na zbiór konstrukcyjny i zbiór walidacyjny (hold out 2/3-1/3)
 - generowanie najlepszej reguły na podstawie zbioru konstrukcyjnego (przez rozszerzanie reguły z użyciem information gain)
 - przycinanie reguły z wykorzystaniem zbioru walidacyjnego, przycięta reguła może pokrywać przykłady negatywne
 - warunek stopu wykorzystuje miarę długości kodu (ile informacji potrzeba do zakodowania zbioru reguł i przykładów pozytywnych i negatywnych)