



Drzewo decyzyjne

- Klasyfikator (model klasyfikacji) stosowany w klasyfikacji zapalczyczej (gorliwej).
- Jego każdy węzeł wewnętrzny zawiera warunek zbudowany z jednego atrybutu warunkowego i wartości z dziedziny tego atrybutu.
- Jego liście zawierają przewidywaną wartość atrybutu decyzyjnego.
- Jest używany do przypisywania do stosownych klas decyzyjnych obiektów, których wartości atrybutu decyzyjnego nie są znane.

Tworzenie drzewa decyzyjnego

- Jest tworzone na podstawie charakterystyki obiektów o znanych wartościach atrybutu decyzyjnego.
- Do budowy drzewa decyzyjnego wykorzystuje się pewną miarę różnorodności klas decyzyjnych w analizowanych zbiorach obiektów. Typowo stosowaną miarą jest *zysk informacyjny* (ang. *information gain, IG*) lub miara *Gini*.

Przykład: Drzewo decyzyjne

Id	Age	Car type	Risk
0	23	Family	High
1	17	Sport	High
2	43	Sport	High
3	68	Family	Low
4	32	Truck	Low
5	20	Family	High

```


graph TD
    A["Car type ∈ {Sport}?"] -- nie --> B["Age < 30?"]
    A -- tak --> C["Risk = High"]
    B -- nie --> D["Risk = Low"]
    B -- tak --> E["Risk = High"]
    
```

Przykład: Niedefinitywne drzewo decyzyjne

Id	Age	Car type	Risk
0	20	Family	High
1	17	Sport	High
2	43	Sport	High
3	68	Family	Low
4	32	Truck	Low
5	20	Family	High
6	20	Family	Low


```

graph TD
    A["Car type ∈ {Sport}?"] -- nie --> B["Age < 30?"]
    A -- tak --> C["Risk = High"]
    B -- nie --> D["Risk = Low"]
    B -- tak --> E["Risk = High[2/3] ∨ Low[1/3]"]
    
```




Miara Giniego

- $Gini(S) = 1 - \sum_{i=1..n} (|S_i|/|S|)^2$, gdzie:
 - $S = S_1 \cup \dots \cup S_n$ jest zbiorem obiektów, a
 - $S_i, i=1..n$, są jego podzbiorymi wyznaczonymi przez wartości atrybutu decyzyjnego.
- Gini przyjmuje wartości z przedziału $[0, 1)$.
- $Gini(S) = 0$ implikuje, że wszystkie obiekty w zbiorze S należą do tej samej klasy decyzyjnej.



Miara Giniego po podziale zbioru obiektów


- Po podziale (*rozszczepieniu*, ang. *split*) zbioru obiektów S na wzajemnie rozłączne podzbiory $S_i, i=1..n$, takie, że $S = S_1 \cup \dots \cup S_n$, miara Gini dla zbioru S (oznaczana jako $Gini_i$) jest wyznaczana następująco:

$$Gini_i(S) = \sum_{i=1..n} (|S_i|/|S|) \times Gini(S_i).$$


SPRINT: Tworzenie drzewa decyzyjnego...

SPRINT:

- Tworzy binarne drzewo decyzyjne;
- Używa miary Giniego;
- Stosuje warunki dwóch rodzajów:
 - Dla atrybutu ciągłego $c: c < v, v \in V_c$;
 - Dla atrybutu nominalnego $c: c \in X, X \subseteq V_c$.




SPRINT: Tworzenie drzewa decyzyjnego...

- Niech D będzie zbiorem obiektów (tablicą decyzyjną) o atrybucie Id zawierającym identyfikatory obiektów, n atrybutach warunkowych i atrybucie decyzyjnym d .


Inicjalizacja:

- Dla każdego atrybutu warunkowego c , na podstawie tablicy decyzyjnej D jest tworzona jej podtablica decyzyjna D^c o trzech atrybutach:

$$(Id, c, d)$$


SPRINT: Tworzenie drzewa decyzyjnego...

- Dla każdego atrybutu warunkowego niezależnie jest znajdowany najlepszy warunek podziału obiektów w oparciu o podtablicę decyzyjną D^c .
- Spośród tych warunków wybierany jest (globalnie) **najlepszy warunek podziału** (NWP), czyli skutkujący **minimalną wartością $Gini_i$** . NWP jest używany do tworzenia nowego węzła wewnętrznego drzewa dec.
- Ostatnie 2 kroki powtarzane są rekurencyjnie dla:
 - podzbioru D_{tak} obiektów w D^w , spełniających warunek NWP;
 - podzbioru D_{nie} obiektów w D^w , niespełniających warunku NWP; gdzie w oznacza atrybut występujący w warunku NWP.



SPRINT: Atrybuty ciągłe...

- **Inicjalizacja:** D^c jest sortowany względem atrybutu c .
- Rozważane warunki podziału obiektów są tworzone w oparciu o wartości atrybutu c w D^c .
- Wyznaczenie najlepszego warunku podziału ze względu na atrybut c wymaga jednokrotnego odczytania D^c .

Przykład: Atrybuty ciągłe...

D^{Age} po posortowaniu względem Age

Id	Age	Risk
1	17	High
5	20	High
0	23	High
4	32	Low
2	43	High
3	68	Low

Rozważane warunki podziału:

- Age ≤ 17?
- Age ≤ 20?
- Age ≤ 23?
- Age ≤ 32?
- Age ≤ 43?

Przykład: Atrybuty ciągłe...

- Age ≤ 17?

Id	Age	Risk
1	17	High
5	20	High
0	23	High
4	32	Low
2	43	High
3	68	Low

Risk	High	Low
$ D^{Age}_{tak} $	1	0
$ D^{Age}_{nie} $	3	2

$gini_{tak} = 1 - [(1/1)^2 + (0/1)^2] = 0$
 $gini_{nie} = 1 - [(3/5)^2 + (2/5)^2] = 12/25$
 $gini_r = (1/6) \times 0 + (5/6) \times (12/25)$
 Stąd, $gini_r = 2/5$.

- Age ≤ 20?

Risk	High	Low
$ D^{Age}_{tak} $	2	0
$ D^{Age}_{nie} $	2	2

$gini_r = 1/3$.

...

Przykład: Atrybuty ciągłe

Id	Age	Risk
1	17	High
5	20	High
0	23	High
4	32	Low
2	43	High
3	68	Low

- Znaleziony najlepszy warunek podziału ze względu na atrybut Age:
 $Age \leq 23?$
 $z Gini_r = 2/9$.
- Gdyby warunek ten okazał się globalnie najlepszym, w węźle drzewa zostałby zapisany w postaci:
 $Age < 27.5?$
 gdzie $27.5 = (23 + 32) / 2$.

SPRINT: Atrybuty nominalne...

- **Inicjalizacja:** Dla atrybutu warunkowego c wyznaczany jest histogram względem atrybutu decyzyjnego.
- Wyznaczenie histogramu wymaga jednokrotnego odczytania podtablicy decyzyjnej D^c .
- Warunki podziału względem atrybutu nominalnego c są oceniane na podstawie tego histogramu.

Przykład: Atrybuty nominalne...

$D^{car\ type}$

Id	Car type	Risk
0	Family	High
1	Sport	High
2	Sport	High
3	Family	Low
4	Truck	Low
5	Family	High

Histogram dla D^{car}

Car type	High	Low
Family	2	1
Sport	2	0
Truck	0	1

Przykład: Atrybuty nominalne...

- Histogram dla D^{car}

Car type	High	Low
Family	2	1
Sport	2	0
Truck	0	1

Rozważane warunki podziału:

- Car type ∈ {Family}?
- Car type ∈ {Sport}?
- Car type ∈ {Truck}?
- Car type ∈ {Family, Sport}?
- Car type ∈ {Family, Truck}?
- Car type ∈ {Sport, Truck}?

Przykład: Atrybuty nominalne

Car type	High	Low
Family	2	1
Sport	2	0
Truck	0	1

- Car type $\in \{\text{Family}\}$

Risk	High	Low
$ D_{\text{Car type tak}} $	2	1
$ D_{\text{Car type nie}} $	2	1

 $\text{gini}_r = 4/9$
- Car type $\in \{\text{Family, Sport}\}$

Risk	High	Low
$ D_{\text{Car type tak}} $	4	1
$ D_{\text{Car type nie}} $	0	1

 $\text{gini}_r = 4/15$

...

Przykład: Najlepszy warunek podziału

- Najlepszy warunek podziału względem Age:
Age < 27.5? ($\text{gini}_r = 2/9$).
- Najlepszy warunek podziału względem Car type:
Car type $\in \{\text{Truck}\}$? ($\text{gini}_r = 4/15$).

↓

- Globalnie najlepszy warunek podziału, to:
Age < 27.5?
- Warunek ten zostanie zapisany w nowo utworzonym węźle drzewa decyzyjnego.

Miara Giniego a entropia (H) i zysk informacyjny (IG)

Gini	$\text{Gini}(S) = 1 - \sum_i \left(\frac{ S_i }{ S }\right)^2$
Gini po podziale	$\text{Gini}_r(S) = \frac{ S_{\text{yes}} }{ S } \times \text{Gini}(S_{\text{yes}}) + \frac{ S_{\text{no}} }{ S } \times \text{Gini}(S_{\text{no}})$
Najlepszy warunek podziału	z min. $\text{Gini}_r(S)$
Entropia	$H(S) = - \sum_i \frac{ S_i }{ S } \times \log_2 \left(\frac{ S_i }{ S }\right)$
Zysk informacyjny	$\text{IG}(S) = H(S) - H_r(S)$, gdzie $H_r(S) = \frac{ S_{\text{yes}} }{ S } \times H(S_{\text{yes}}) + \frac{ S_{\text{no}} }{ S } \times H(S_{\text{no}})$
Najlepszy warunek podziału	z maks. $\text{IG}(S)$ (czyli min. $H_r(S)$)

Rodziny/zespoły klasyfikatorów: Lasy losowe...

- Niech D będzie zbiorem danych o znanych wartościach atrybutu decyzyjnego i $n = |D|$.
- D jest używane do utworzenia k zbiorów danych: $D_1 \dots D_k$.
- Każdy zbiór D_i jest tworzony poprzez n-krotne losowanie ze zwracaniem rekordów ze zbioru D. W konsekwencji:
 - każdy D_i ma n rekordów, ale
 - liczba unikatowych rekordów w $D_i \leq n$.

Rodziny klasyfikatorów: Lasy losowe...

- Drzewo decyzyjne DecTree_i jest tworzone na podstawie zbioru D_i , $i = 1..k$.
- Klasyfikowany obiekt jest przypisywany do klasy decyzyjnej wskazywanej przez większość klasyfikatorów $\text{DecTree}_1, \dots, \text{DecTree}_k$.

Lasy losowe: Tworzenie drzew

Tworzenie DecTree_i na podstawie D_i :

- Dla każdego tworzonego węzła wewnętrznego niezależnie losuj mały podzbiór A pełnego zbioru atrybutów warunkowych AT (zwykle $|A| = \sqrt{|AT|}$).
- Twórz warunki podziału wyłącznie ze względu na atrybuty w A.
- W aktualnie tworzonym węźle zapisz najlepszy z tych warunków.

Klasyfikacja z użyciem (minimalnych) wzorców kontrastowych



Wzorce kontrastowe (CPs)

Przykładowy zbiór danych D z dwiema klasami decyzyjnymi: D_P i D_N .

Outlook	Temperature	Humidity	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

• Wzorec kontrastowy (ang. *Contrast Pattern*) jest definiowany jako wzorec występujący w dokładnie jednej klasie decyzyjnej.

• {Overcast, Hot} jest wzorcem kontrastowym. Występuje on tylko w klasie decyzyjnej D_P .



Dane wejściowe

- D – tablica decyzyjna, której część obiektów należy do pozytywnej klasy decyzyjnej D_P , a pozostałe do negatywnej klasy decyzyjnej D_N .
- T – klasyfikowany obiekt.



Przykład: Dane wejściowe i klasyfikowany obiekt

Przykładowy zbiór danych D z dwiema klasami decyzyjnymi: D_P i D_N .

Outlook	Temperature	Humidity	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

• Klasyfikowany obiekt:

T = {Sunny, Mild, High, True}



Leniwa klasyfikacja z użyciem CPs - kroki

- I. Utwórz odpowiadającą klasyfikowanemu obiektowi T, zredukowaną wersję D' tablicy decyzyjnej D.
- II. Znajdź minimalne wzorce kontrastowe $\text{Min}(\text{CP}(D'_P))$ i $\text{Min}(\text{CP}(D'_N))$ w D'.
- III. Użyj (min.) wzorców kontrastowych do wyznaczenia ocen przynależności obiektu T do poszczególnych klas dec.
- IV. Wybierz dla T klasę decyzyjną, dla której wartość oceny przynależności jest maks.



Krok I

- Utwórz tablicę decyzyjną D', która będzie zawierać wszystkie oryginalne wartości atrybutu decyzyjnego z D i tylko te wartości atrybutów warunkowych z D, które występują w T.

Przykład: Krok I

Oryginalna D & T = {Sunny, Mild, High, True} → zredukowana D'

Outlook	Temp.	Hum.	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

Outlook	Temp.	Hum.	Windy	Activity
		High		P
	Mild	High		P
				P
			True	P
Sunny				P
	Mild			P
Sunny	Mild		True	P
	Mild	High	True	P
				P
Sunny		High		N
Sunny		High	True	N
			True	N
Sunny	Mild	High		N
	Mild	High	True	N

Krok II

- Znajdź minimalne wzorce kontrastowe dla każdej klasy decyzyjnej: $\text{Min}(\text{CP}(D'_P))$ i $\text{Min}(\text{CP}(D'_N))$ in D' .
- Obserwacje:
 - Wzorce kontrastowe można wyznaczać stosując prostą modyfikację algorytmu Apriori.
 - Wsparcie wzorca można jednocześnie wyznaczać w D'_P i w D'_N .
 - Wzorec kandydujący jest wzorcem kontrastowym, jeżeli jego wsparcie jest różne od 0 w dokładnie jednej klasie decyzyjnej.

Przykład: Krok II...

Zredukowana D'

Outlook	Temp.	Hum.	Windy	Activity
		High		P
	Mild	High		P
				P
			True	P
Sunny				P
	Mild			P
Sunny	Mild		True	P
	Mild	High	True	P
				P
Sunny		High		N
Sunny		High	True	N
			True	N
Sunny	Mild	High		N
	Mild	High	True	N

- Kandydaci o długości 1:
 - $\{\text{Sunny}\}_{2/3}$ $\{\text{Mild}\}_{4/2}$ $\{\text{High}\}_{3/4}$ $\{\text{True}\}_{3/3}$
- Rezultat: Nie ma minimalnych CPs o długości 1.
- Kandydaci o długości 2:
 - $\{\text{Sunny, Mild}\}_{1/1}$ $\{\text{Sunny, High}\}_{0/3}$
 - $\{\text{Sunny, True}\}_{1/1}$ $\{\text{Mild, High}\}_{2/2}$
 - $\{\text{Mild, True}\}_{2/1}$ $\{\text{High, True}\}_{1/2}$

Przykład: Krok II...

Zredukowana D'

Outlook	Temp.	Hum.	Windy	Activity
		High		P
	Mild	High		P
				P
			True	P
Sunny				P
	Mild			P
Sunny	Mild		True	P
	Mild	High	True	P
				P
Sunny		High		N
Sunny		High	True	N
			True	N
Sunny	Mild	High		N
	Mild	High	True	N

- Kandydaci o długości 2:
 - $\{\text{Sunny, Mild}\}_{1/1}$ $\{\text{Sunny, High}\}_{0/3}$
 - $\{\text{Sunny, True}\}_{1/1}$ $\{\text{Mild, High}\}_{2/2}$
 - $\{\text{Mild, True}\}_{2/1}$ $\{\text{High, True}\}_{1/2}$
- Rezultat: $\{\text{Sunny, High}\} \in \text{Min}(\text{CP}(D'_N))$.
- Kandydaci o długości 3:
 - $\{\text{Sunny, Mild, True}\}_{1/0}$
 - $\{\text{Mild, High, True}\}_{1/1}$

Przykład: Krok II

Zredukowana D'

Outlook	Temp.	Hum.	Windy	Activity
		High		P
	Mild	High		P
				P
			True	P
Sunny				P
	Mild			P
Sunny	Mild		True	P
	Mild	High	True	P
				P
Sunny		High		N
Sunny		High	True	N
			True	N
Sunny	Mild	High		N
	Mild	High	True	N

- Kandydaci o długości 3:
 - $\{\text{Sunny, Mild, True}\}_{1/0}$
 - $\{\text{Mild, High, True}\}_{1/1}$
- Rezultat: $\{\text{Sunny, Mild, True}\} \in \text{Min}(\text{CP}(D'_P))$.
- Brak kandydatów o długości 4.

Krok III

- Wyznacz ocenę przynależności klasyfikowanego obiektu T do każdej klasy decyzyjnej D_i na podstawie liczności obiektów w tej klasie, wspierających jej minimalne wzorce kontrastowe:

$$\text{ocenaPrzynależności}(D_i) = \frac{\sup(\bigvee \text{Min}(\text{CP}(D'_i)))}{|D'_i|}$$

Przykład: Krok III

Zredukowana D'

Outlook	Temp.	Hum.	Windy	Activity
		High		P
	Mild	High		P
				P
			True	P
Sunny				P
	Mild			P
Sunny	Mild		True	P
	Mild	High	True	P
				P
Sunny		High		N
Sunny		High		N
Sunny	Mild	High		N
	Mild	High	True	N

- $|D'_p| = 9$ obiektów
- $\text{Min}(\text{CP}(D'_p)) = \{\{\text{Sunny, Mild, True}\}\}$
- $\text{ocenaPrzynależności}(D'_p) = \frac{\sup(\text{Min}(\text{CP}(D'_p)))}{|D'_p|} = 1/9$
- $|D'_N| = 5$ obiektów
- $\text{Min}(\text{CP}(D'_N)) = \{\{\text{Sunny, High}\}\}$
- $\text{ocenaPrzynależności}(D'_N) = \frac{\sup(\text{Min}(\text{CP}(D'_N)))}{|D'_N|} = 3/5$

Przykład: Krok IV

Tablica decyzyjna D

Outlook	Temp.	Hum.	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

- $\text{Min}(\text{CP}(D'_p)) = \{\{\text{Sunny, Mild, True}\}\}$
- $\text{ocenaPrzynależności}(D'_p) = 1/9$
- $\text{Min}(\text{CP}(D'_N)) = \{\{\text{Sunny, High}\}\}$
- $\text{ocenaPrzynależności}(D'_N) = 3/5$
- Obiekt T = {Sunny, Mild, High, True} będzie zaklasyfikowany do klasy decyzyjnej o wyższej ocenie przynależności, czyli do D_N .

Postępowanie w przypadku atrybutów ciągłych

- Wartości każdego atrybutu ciągłego c w:
 - tablicy decyzyjnej D i
 - każdym klasyfikowanym obiekcie T
 są przekształcane na wartości z przedziału [0, 1].
- zredukowana ze względu na T tablica decyzyjna D' nie powinna zawierać żadnej wartości atrybutu ciągłego c spoza przedziału $[v_c - \alpha\%, v_c + \alpha\%]$, gdzie $v_c \in T$.
- Pozostałe wartości atrybutu ciągłego c powinny być zastąpione wartością v_c w tablicy D' .

Przykład: Normalizacja

Normalizacja tablicy dec. D

Age	...	Activity	Age	...	Activity
18	...	P	0.18	...	P
45	...	P	0.45	...	P
38	...	P	0.38	...	P
60	...	P	0.60	...	P
19	...	P	0.19	...	P
50	...	P	0.50	...	P
39	...	P	0.39	...	P
22	...	P	0.22	...	P
44	...	P	0.44	...	P
48	...	N	0.48	...	N
25	...	N	0.25	...	N
32	...	N	0.32	...	N
56	...	N	0.56	...	N
40	...	N	0.40	...	N

Normalizacja obiektu T

$T = \{28, \dots\} \Rightarrow T = \{0.28, \dots\}$

Przykład: Wyznaczanie zredukowanej tablicy decyzyjnej D'

Znormalizowana D $\rightarrow D'$

Age	...	Activity	Age	...	Activity
0.18	...	P	P
0.45	...	P	P
0.38	...	P	P
0.60	...	P	P
0.19	...	P	P
0.50	...	P	P
0.39	...	P	P
0.22	...	P	P
0.44	...	P	P
0.48	...	N	N
0.25	...	N	0.28	...	N
0.32	...	N	0.28	...	N
0.56	...	N	N
0.40	...	N	N

- Niech:
 - znormalizowany obiekt $T = \{0.28, \dots\}$,
 - $\alpha\% = 5\%$.
- Wtedy wartości uznane za (wystarczająco) nierozróżnialne z 0.28 należą do przedziału:

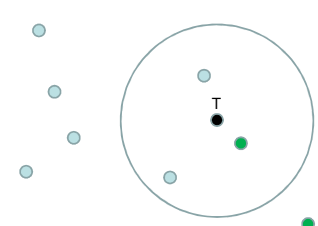
$$[0.23, 0.33]$$

Klasyfikacja z użyciem k najbliższych sąsiadów

Klasyfikacja z użyciem k najbliższych sąsiadów (kNN)

- Niech T będzie klasyfikowanym obiektem. Jego klasyfikacja z użyciem kNN polega na:
 - wyszukaniu k najbliższych sąsiadów obiektu T w tablicy decyzyjnej,
 - przypisaniu obiektu T do klasy decyzyjnej reprezentowanej najliczniej wśród znalezionych k najbliższych sąsiadów.

Przykład: klasyfikacja z użyciem kNN



Niech $k = 3$. Wtedy obiekt T zostanie zakwalifikowany do klasy obiektów oznaczonych kolorem niebieskim.

Współczynnik podobieństwa Gowera



Współczynnik podobieństwa Gowera...

Współczynnik podobieństwa Gowera dla obiektów p i q :

$$G(p, q) = \frac{\sum_{d=1}^m w(p_d, q_d) \times s(p_d, q_d)}{\sum_{d=1}^m w(p_d, q_d)}, \text{ gdzie}$$

- m jest liczbą atrybutów;
- $s(p_d, q_d)$ jest podobieństwem obiektów p i q ze względu na atrybut d ;
- $w(p_d, q_d)$ jest wagą atrybutu d dla obiektów p i q . Jeżeli p i q są nieporównywalne ze względu na atrybut d (np. p_d lub q_d jest nieznanne), to $w(p_d, q_d) = 0$. Wpp., zazwyczaj $w(p_d, q_d)$ przyjmuje się za równe 1.

Współczynnik podobieństwa Gowera

- Dla atrybutów numerycznych: $s(p_d, q_d) = 1 - \frac{|p_d - q_d|}{range_d}$.
- Dla atrybutów nominalnych: $s(p_d, q_d) = 1$, jeśli $p_d = q_d$; $s(p_d, q_d) = 0$, wpp.
- Dla atrybutów binarych (dychotomicznych):
 - Jeśli $p_d = 1$ i $q_d = 1$, to $s(p_d, q_d) = 1$ i $w(p_d, q_d) = 1$.
 - Jeśli $p_d = 1$ i $q_d = 0$, to $s(p_d, q_d) = 0$ i $w(p_d, q_d) = 1$.
 - Jeśli $p_d = 0$ i $q_d = 1$, to $s(p_d, q_d) = 0$ i $w(p_d, q_d) = 1$.
 - Jeśli $p_d = 0$ i $q_d = 0$, to $s(p_d, q_d) = 0$ i $w(p_d, q_d) = 0$.

Przykład: Obliczanie współczynnika podobieństwa Gowera...

Id	Age	Car type	Risk
0	23	Family	High
1	17	Sport	High
2	43	Sport	Low
3	68	Family	Low
4	32	Truck	Low
5	20	Family	High

Object to be classified:

Id	Age	Car type	Risk
6	20	Sport	?

$$G(6,0) = \frac{1 * \left(1 - \frac{|20 - 23|}{100}\right) + 1 * 0}{\frac{0.97}{2} + 1} = 0.485;$$

$$G(6,1) = \frac{1 * \left(1 - \frac{|20 - 17|}{100}\right) + 1 * 1}{\frac{1.97}{2} + 1} = 0.985$$

Przykład: Obliczanie współczynnika podobieństwa Gowera

Id	Age	Car type	Risk
0	23	Family	High
1	17	Sport	High
2	43	Sport	Low
3	68	Family	Low
4	32	Truck	Low
5	20	Family	High

Object to be classified:

Id	Age	Car type	Risk
7		Sport	?

$$G(p, q) = \frac{\sum_{d=1}^m w(p_d, q_d) \times s(p_d, q_d)}{\sum_{d=1}^m w(p_d, q_d)}$$

$$G(7, 0) = \frac{0 + \text{undefined} + 1 + 0}{0 + 1} = \frac{1}{1} = 1$$

$$G(7, 1) = \frac{0 + \text{undefined} + 1 + 1}{0 + 1} = \frac{2}{1} = 2$$

Klasyfikacja z użyciem naiwnego klasyfikatora Bayesowskiego

Twierdzenie Bayesa...

- $P(D|C) = \frac{P(C \cap D)}{P(C)}$
- $P(C|D) = \frac{P(C \cap D)}{P(D)}$
- Stąd:

$$P(D|C) \times P(C) = P(C \cap D) = P(C|D) \times P(D).$$

Twierdzenie Bayesa

- $P(D|C) \times P(C) = P(C|D) \times P(D).$
- Stąd:

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

Klasyfikator Bayesowski...

Przykładowy zbiór danych D z dwiema klasami decyzyjnymi.

Outlook	Temperature	Humidity	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

Niech $T = \{\text{Sunny, Mild, High, True}\}$ będzie klasyfikowanym obiektem.

Jeśli $P(D_P|C_T) > P(D_N|C_T)$, to zwróć decyzję P.
Wpp. zwróć decyzję N.

Klasyfikator Bayesowski

Klasyfikowany obiekt: $T = \{\text{Sunny, Mild, High, True}\}$.

Jeżeli $P(D_P|C_T) > P(D_N|C_T)$,
to zwróć decyzję P.
Wpp. zwróć decyzję N.

Własność:

$$P(D_P|C_T) > P(D_N|C_T) \Leftrightarrow \frac{P(C_T|D_P) \times P(D_P)}{P(C_T)} > \frac{P(C_T|D_N) \times P(D_N)}{P(C_T)} \Leftrightarrow P(C_T|D_P) \times P(D_P) > P(C_T|D_N) \times P(D_N).$$

Naiwny klasyfikator Bayesowski...

Przykładowy zbiór danych D z dwiema klasami decyzyjnymi.

Outlook	Temperature	Humidity	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

$T = \{\text{Sunny, Mild, High, True}\}.$

Jeśli $P(C_T|D_P) \times P(D_P) > P(C_T|D_N) \times P(D_N),$
to zwróć P, wpp. N.

$P(D_P) = \frac{9}{14}.$
 $P(C_T|D_P) = \frac{2}{9} \times \frac{4}{9} \times \frac{3}{9} \times \frac{3}{9}.$
 Stąd:
 $P(C_T|D_P) \times P(D_P) \approx 0.007.$

Naiwny klasyfikator Bayesowski...

Przykładowy zbiór danych D z dwiema klasami decyzyjnymi.

Outlook	Temperature	Humidity	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

$T = \{\text{Sunny, Mild, High, True}\}.$

Jeśli $P(C_T|D_P) \times P(D_P) > P(C_T|D_N) \times P(D_N),$
to zwróć P, wpp. N.

$P(D_N) = \frac{5}{14}.$
 $P(C_T|D_N) = \frac{3}{5} \times \frac{2}{5} \times \frac{4}{5} \times \frac{3}{5}.$
 Stąd:
 $P(C_T|D_N) \times P(D_N) \approx 0.041.$

Naiwny klasyfikator Bayesowski

Przykładowy zbiór danych D z dwiema klasami decyzyjnymi.

Outlook	Temperature	Humidity	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

$T = \{\text{Sunny, Mild, High, True}\}.$

Jeśli $P(C_T|D_P) \times P(D_P) > P(C_T|D_N) \times P(D_N),$
to zwróć P, wpp. N.

$P(C_T|D_P) \times P(D_P) \approx 0.007.$
 $P(C_T|D_N) \times P(D_N) \approx 0.041.$
 Stąd, T zostanie zaklasyfikowany do klasy decyzyjnej N.

Problem z częstością zero

Przykładowy zbiór danych D z dwiema klasami decyzyjnymi.

Outlook	Temperature	Humidity	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

- **Outlook = Overcast nie występuje w klasie dec. N!**
- Dziedzina atrybutu Outlook ma $|V_{\text{Outlook}}| = 3$ wartości.
- **Rozwiązanie:** Zmodyfikuj wszystkie prawdopodobieństwa warunkowe $P(\text{Outlook} = v|D_P)$ i $P(\text{Outlook} = v|D_N)$, gdzie $v \in V_{\text{Outlook}}$, zwiększając licznik o małą wartość λ i zwiększając mianownik o $|V_{\text{Outlook}}| \times \lambda$. Typowo, $\lambda = 1$ (lub $\lambda = 1/|D|$).

Rozwiązanie problemu częstości zero

Przykładowy zbiór danych D z dwiema klasami decyzyjnymi.

Outlook	Temperature	Humidity	Windy	Activity
Overcast	Hot	High	False	P
Rain	Mild	High	False	P
Rain	Cool	Normal	False	P
Overcast	Cool	Normal	True	P
Sunny	Cool	Normal	False	P
Rain	Mild	Normal	False	P
Sunny	Mild	Normal	True	P
Overcast	Mild	High	True	P
Overcast	Hot	Normal	False	P
Sunny	Hot	High	False	N
Sunny	Hot	High	True	N
Rain	Cool	Normal	True	N
Sunny	Mild	High	False	N
Rain	Mild	High	True	N

- $V_{\text{Outlook}} = \{\text{Overcast, Rain, Sunny}\}.$
- $Stqd, |V_{\text{Outlook}}| = 3.$
- Niech $\lambda = 1$. Wtedy:
- $P(\text{Outlook} = \text{Overcast}|D_P) = \frac{4+1}{9+3}.$
- $P(\text{Outlook} = \text{Rain}|D_P) = \frac{3+1}{9+3}.$
- $P(\text{Outlook} = \text{Sunny}|D_P) = \frac{2+1}{9+3}.$
- $P(\text{Outlook} = \text{Overcast}|D_N) = \frac{0+1}{5+3}.$
- $P(\text{Outlook} = \text{Rain}|D_N) = \frac{2+1}{5+3}.$
- $P(\text{Outlook} = \text{Sunny}|D_N) = \frac{3+1}{5+3}.$

Naiwny klasyfikator Bayesowski a atrybuty o wartościach ciągłych

- Dla atrybutu c o wartościach ciągłych zakłada się rozkład Gaussowski. Wtedy:

$$P(c_i = x|D_i) = \frac{1}{\sqrt{2\pi}\sigma_i} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}},$$

gdzie:

- c_i - wartość atrybutu c dla i-tego obiektu w klasie dec. D_i ,
- średnia $\mu_i = \frac{1}{n_i} \sum_{l=1}^{n_i} c_{il},$
- odchylenie standardowe $\sigma_i = \sqrt{\frac{1}{n_i} \sum_{l=1}^{n_i} (c_{il} - \mu_i)^2},$
- n_i - liczba obiektów w klasie dec. D_i ($n_i = |D_i|$).

Ocena jakości klasyfikatora



Metoda wydzielania (ang. holdout method)...

- Zbiór danych o znanych wartościach atrybutu decyzyjnego jest dzielony na:
 - *zbiór uczący* (typowo: 2/3) i
 - *zbiór testowy* (typowo: 1/3).
- Zbiór uczący jest używany do *konstrukcji* klasyfikatora.
- Zbiór testowy jest używany do *testowania* klasyfikatora.
- Obydwa zbiory powinny być reprezentatywne.



Metoda wydzielania

- *Szacowana dokładność* = stosunek liczby poprawnie sklasyfikowanych obiektów ze zbioru testowego do liczności tego zbioru.



Losowe próbkowanie (ang. random subsampling)

- *Losowe próbkowanie* polega na *k*-krotnym wykonaniu metody wydzielania.
- *Oszacowanie dokładności klasyfikatora* jest wyznaczane jako *średnia* z oszacowań dokładności uzyskanych w każdej iteracji.



k-krotna walidacja krzyżowa


- Zbiór danych D jest losowo dzielony na k (często stosowane $k = 10$ lub $k = 5$) wzajemnie rozłącznych podzbiorów (części): $D_1 \dots D_k$.
- W iteracji i ($i = 1 \dots k$):
 - $D \setminus D_i$ pełni rolę *zbioru uczącego*,
 - D_i pełni rolę *zbioru testowego*.
- *Szacowana dokładność* = stosunek sumarycznej liczby poprawnych klasyfikacji z k iteracji do $|D|$.



Walidacja krzyżowa „Leave-One-Out”


- „*Leave-one-out*” jest *k*-krotną walidacją krzyżową, gdzie:

$$k = |D|.$$




Walidacja lasu losowego

- **Szacowana dokładność** = stosunek sumarycznej liczby poprawnych klasyfikacji obiektów z D do liczności D , z tym, że decyzja o zaklasyfikowaniu każdego pojedynczego obiektu z D jest podejmowana wyłącznie z użyciem drzew, w których budowie ten obiekt nie brał udziału.




Dodatkowe miary oceny dla klasyfikatora z 2 wartościami decyzyjnymi

- **czułość** = **zwrot** = $\frac{\# \text{poprawnie sklasyfikowanych w klasie Poz}}{\text{liczność klasy Poz}}$
- **specyficzność** = $\frac{\# \text{poprawnie sklasyfikowanych w klasie Neg}}{\text{liczność klasy Neg}}$
- **precyzja** = $\frac{\# \text{poprawnie sklasyfikowanych w klasie Poz}}{\# \text{zaklasyfikowanych (poprawnie lub nie) do klasy Poz}}$
- **F-miara** = $\frac{2 \times \text{precyzja} \times \text{zwrot}}{\text{precyzja} + \text{zwrot}}$



Krzywa ROC (Receiver Operating Characteristic)


- **czułość** = $\frac{\# \text{poprawnie sklasyfikowanych w klasie Poz}}{\text{liczność klasy Poz}}$
- **specyficzność** = $\frac{\# \text{poprawnie sklasyfikowanych w klasie Neg}}{\text{liczność klasy Neg}}$
- **Krzywa ROC** – wykres tworzony na podstawie dwuwymiarowych punktów (1-specyficzność, czułość).



Ocena klasyfikatora w przypadku silnie zróżnicowanych licznosci klas decyzyjnych


- **Zbalansowana_dokladność** =
$$\frac{1}{l} \sum_{i=1}^l \frac{\# \text{poprawnie sklasyfikowanych w klasie } D_i}{\text{liczność klasy } D_i}$$

gdzie l jest liczbą klas decyzyjnych.




Literatura...

- Hongjian Fan, Kotagiri Ramamohanarao: Fast Discovery and the Generalization of Strong Jumping Emerging Patterns for Building Compact and Accurate Classifiers. IEEE Trans. Knowl. Data Eng. 18(6): 721-737, 2006
- J. C. Gower, A general coefficient of similarity and some of its properties. Biometrics 27, 857-874 (1971)
- Jiawei Han, Micheline Kamber, Jian Pei: Data Mining: Concept and Techniques, The Morgan Kaufmann Series in Data Management Systems, 2011
- Jacek Koronacki, Jan Ćwik: Statystyczne systemy uczące się, Akademicka Oficyna Wydawnicza EXIT, 2008
- Marzena Kryszkiewicz: Virtual Balancing of Decision Classes. ACIIDS (1) 2017: 673-684




Literatura

- Marzena Kryszkiewicz, Przemysław Podsiadły: Explicit Contrast Patterns Versus Minimal Jumping Emerging Patterns for Lazy Classification in High Dimensional Data. IEA/AIE 2016: 80-94
- Jinyan Li, Guozhu Dong, Kotagiri Ramamohanarao: Instance-Based Classification by Emerging Patterns. PKDD 2000: 191-200
- Tadeusz Morzy, Eksploracja danych: Metody i algorytmy, Wydawnictwo Naukowe PWN, 2013
- John C. Shafer, Rakesh Agrawal, Manish Mehta: SPRINT: A Scalable Parallel Classifier for Data Mining [VLDB 1996](#): 544-555
- Paweł Terlecki, Krzysztof Walczak: Efficient Discovery of Top-K Minimal Jumping Emerging Patterns. RSCTC 2008: 438-447




Ćwiczenia...

1. Niech Activity będzie atrybutem decyzyjnym w tabeli na slajdzie 26. Postępując zgodnie z algorytmem SPRINT:
 - Wyznacz histogram dla atrybutu *Outlook* z uwzględnieniem atrybutu decyzyjnego.
 - Wyznacz miarę Gini dla warunku podziału: *Outlook* \in {Outlook}?
 - Wyznacz miarę Gini dla warunku podziału: *Outlook* \in {Rain}?
 - Wyznacz miarę Gini dla warunku podziału: *Outlook* \in {Sunny}?
 - Który z powyższych warunków podziału jest lepszy?



Ćwiczenia...

2. Niech Activity będzie atrybutem decyzyjnym w tabeli na slajdzie 26. Obiekt $T = \{\text{Sunny, Cool, High, True}\}$ ma być sklasyfikowany z użyciem minimalnych wzorców kontrastowych na podstawie tej tabeli.
 - Dokonaj redukcji tablicy decyzyjnej ze względu na obiekt T.
 - W oparciu o tę zredukowaną tablicę decyzyjną wyznacz minimalne wzorce kontrastowe dla klas decyzyjnych:
 - Activity = P
 - Activity = N
 - Jaka jest wartość *ocenyPrzynależności* obiektu T do klasy decyzyjnej Activity = P?
 - Jaka jest wartość *ocenyPrzynależności* obiektu T do klasy decyzyjnej Activity = N?
 - Do której klasy decyzyjnej zostanie zaklasyfikowany obiekt T?



Ćwiczenia

3. Niech Risk będzie atrybutem decyzyjnym w tabeli na slajdzie 5. Wyznacz klasę decyzyjną dla obiektu T o wartościach (Age = 25, Car_Type = Family) z użyciem k najbliższych sąsiadów dla:
 - k = 1,
 - k = 3.Do wyznaczenia k najbliższych sąsiadów użyj miary Gowera.
4. Niech Risk będzie atrybutem decyzyjnym w tabeli na slajdzie 5. Wyznacz klasę decyzyjną dla obiektu T o wartościach (Age = 25, Car_Type = Family) za pomocą naiwnego klasyfikatora Bayesowskiego.