Zadania z eksploracji danych i uczenia maszynowego

Marcin Korzeń

Katedra Metod Sztucznej Inteligencji i Matematyki Stosowanej
Wydział Informatyki, ZUT,
mkorzen@wi.zut.edu.pl

1 Rachunek prawdopodobieństwa i statystyka

1. Student zdaje egzamin, który wygląda następująco: są przygotowane dwie urny(I - łatwiejsza i II - trudniejsza) w każdej 10 pytań, egzaminator rzuca kostką do gry - jeżeli wypadnie 1 losuje pytanie z urny I, w przeciwnym wypadku losuje pytanie z urny II. Wiadomo że student zna odpowiedź na 9 pytań z urny I zna 2 z urny II. Następnego dnia okazuje się, że student jednak zdał egzamin (jakie jest tego prawdopodobieństwo?) ale nie może sobie przypomnieć, którą urnę wylosował. Co jest bardziej prawdopodobne, że odpowiedział na pytanie trudniejsze czy łatwiejsze ? A gdyby student nie zdał egzaminu: co jest tego przyczyną? jakie jest prawdopodobieństwo tego, że nie odpowiedział na pytanie prostsze z urny I ?.

2. Co jest bardziej prawdopodobne wygrać z równorzędnym przeciwnikiem dwie partie z trzech czy trzy z pięciu? (partii zakończonych remisem nie bierzemy pod uwagę)?

3. (Prawo następstw Laplace'a) Danych jest $N+1$ urn, w i-tej urnie ($i = 0, ..., N$) jest i kul białych oraz $N - i$ kul czarnych. Wylosowano urnę, a następnie z tej urny wylosowano ze zwracaniem n-razy jedną kulę. Załóżmy, że za każdym wylosowano to kulę białą. Jaka jest szansa, że kolejna kula wylosowana z tej urny będzie biała? Przenanalizować sytuację, gdy rozmiar próby n jest mały, a liczba urn N może być dowolnie duża (wskazówka: formuła Faulhabera, lub def. całki Riemanna).

4. (Problem Monty'ego Halla) Prowadzący ukrył nagrodę w jednym trzech pudełek. Wybieramy pudełko, ale nie otwieramy go, następnie prowadzący otwiera jedno z pustych pudełek i daje możliwość zmiany pudełka na drugie nieodkryte lub pozostania przy swoim wyborze. Należy rozstrzygnąć co się bardziej opłaca: a) zmienić pudełko na drugie nieodkryte, b) pozostać przy swoimi dotychczasowym wyborze, ewentualnie: c) bez znaczenia.

2 Wnioskowanie

1. Rzucamy n razy niesymetryczną monetą (o asymetrii - prawdopodobieństwie orła p), k razy wypadł orzeł $n - k$ reszka. Znaleźć estymator parametru p stosując metodę największej wiarygodności.

2. Podobnie jak w poprzednim zadaniu próba zawiera k orłów i $n-k$ reszek. Podać rozkład posteriori parametru p, zakładając jako prior na parametr p rozkład Beta(2,2).

3. Znaleźć estymator parametru p stosując metodę maksimum posteriori, zakładając jako prior na parametr p rozkład:

   (a) rozkład jednostajny określony na przedziale [0,1]

   (b) rozkład Beta$(2, 2)$

   (c) symetryczny rozkład trójkątny określony na przedziale $[0, 1]$

3 Entropia, przyrost informacji

Własności entropii (H) oraz indeksu Gini'ego (G) oraz odpowiadających im miar przyrostu informacji (I_H oraz I_G). W dalszej części, niech X, Y będą zmiennymi losowymi przyjmującymi odpowiednio wartości: $\{x_1, \ldots, x_n\}$, $\{y_1, \ldots, y_m\}$ z prawdopodobieństwami $\{p_{1,\cdot}, \ldots p_{n,\cdot}\}$, $\{p_{\cdot,1}, \ldots p_{\cdot,m}\}$ oraz prawdopodobieństwa łączne p_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, m$. Entropia:

$$I_H(X) = -\sum_{i=1}^{n} \Pr\{X = x_i\} \log_2(\Pr\{X = x_i\})$$

Informacja wzajemna i przyrost informacji:

$$\begin{aligned} I_H(X,Y) &= H(X) + H(Y) - H(XY) \\ I_H(Y,X) &= H(Y) - H(Y|X) \\ H(Y|X) &= \sum_{i=1}^{k} \Pr\{X = x_i\} H(Y|X = x_i) \end{aligned}$$

Indeks Gini'ego jako miara rozproszenia

$$\begin{aligned} G_{CART}(X) &= 1 - \sum_{i=1}^{k} Pr(X = x_i\}^2 \\ I_{G_{CART}}(Y,X) &= G_{CART}(Y) - G_{CART}(Y|X) \\ G_{CART}(Y|X) &= \sum_{p=1}^{k} Pr\{a = p\} G_{CART}(d|a = p) \end{aligned}$$

1. Znaleźć rozkład, który maksymalizuje H oraz G
2. Uzasadnić, następujące własności[1]:
   (a) $I_H(X,Y) = H(X) + H(Y) - H(X,Y) = I(Y) - I(Y|X)$
   (b) $X \perp Y \Rightarrow H(X,Y) = H(X) + H(Y)$
   (c) $X \perp Y \Rightarrow I_H(Y,X) = 0$
   (d) $X \perp Y \Rightarrow I_G(Y,X) = 0$
   (e) $0 \leq H(X) \leq \log_2(n)$
   (f) $0 \leq I(X,Y) \leq H(Y)$
3. Czy dla indeksu Giniegi można zdefiniować informacje wzajemną, tzn. czy zachodzi: $I_G(X,Y) = G(X) + G(Y) - G(X,Y)$.
4. Na podstawie danych z tabeli 1 wyznaczyć ile informacji na temat udomowienie dostarcza informacja o upierzeniu zwierzęcia.

4 Przetwarzanie wstępne wizualizacja

1. Dla zbioru danych `challengerRing` (tab. 2) wykonać następujące czynności:
   (a) narysować histogramy dla atrybutów,
   (b) zdyskretyzować zmienną temperatura pod warunkiem liczby uszkodzeń,
   (c) napisać w postaci tabeli kontyngencji rozkład liczności tych dwóch zmiennych,

[1] $X \perp Y$ oznacza, że atrybuty są niezależne

Tablica 1: Tablica kontyngencji (zbiór `zoo`)

pióra \ domowe	false	true	łącznie:
false	71	10	81
true	17	3	20
łącznie:	88	13	101

Tablica 2: Zbiór `challengerRing`

Temporal order of flight	Number of O-rings at risk on a given flight	Number experiencing thermal distress	Launch temperature (degrees F)	Leak-check pressure (psi)
1	6	0	66	50
2	6	1	70	50
3	6	0	69	50
4	6	0	68	50
5	6	0	67	50
6	6	0	72	50
7	6	0	73	100
8	6	0	70	100
9	6	1	57	200
10	6	1	63	200
11	6	1	70	200
12	6	0	78	200
13	6	0	67	200
14	6	2	53	200
15	6	0	67	200
16	6	0	75	200
17	6	0	70	200
18	6	0	81	200
19	6	0	76	200
20	6	0	79	200
21	6	2	75	200
22	6	0	76	200
23	6	1	58	200

(d) znaleźć rozkład częstości i rozkłady brzegowe tych dwóch atrybutów,

(e) porównać empiryczny rozkład częstości z rozkładem produktowym,

(f) postawić test χ^2 o zależności zmiennych (znaleźć wartość statystyki),

(g) wziąć dowolny pakiet statystyczny i porównać powyższy wynik testu χ^2 z testem dokładnym Fishera, (np. R, `fisher.test`),

(h) jaka decyzję podjęlibyście, gdyby temperatura w momencie startu wyniosła by 35F ?

2. Dla zbioru danych `contactLens` (tab. 3):

(a) dokonać binaryzacji pierwszego atrybutu,

(b) zamienić wybrane trzy atrybuty w skali nominalnej na numeryczne z zachowaniem porządku,

(c) zbudować macierz kowariancji dla zmiennych numerycznych,

(d) wyznaczyć pierwszą składową (kierunek, wektor) główną,

Tablica 3: Zbiór `contactLens`

	age	spectacleprescrip	astigmatism	tearprodrate	contactlenses
1	young	myope	no	reduced	none
2	young	myope	no	normal	soft
3	young	myope	yes	reduced	none
4	young	myope	yes	normal	hard
5	young	hypermetrope	no	reduced	none
6	young	hypermetrope	no	normal	soft
7	young	hypermetrope	yes	reduced	none
8	young	hypermetrope	yes	normal	hard
9	pre-presbyopic	myope	no	reduced	none
10	pre-presbyopic	myope	no	normal	soft
11	pre-presbyopic	myope	yes	reduced	none
12	pre-presbyopic	myope	yes	normal	hard
13	pre-presbyopic	hypermetrope	no	reduced	none
14	pre-presbyopic	hypermetrope	no	normal	soft
15	pre-presbyopic	hypermetrope	yes	reduced	none
16	pre-presbyopic	hypermetrope	yes	normal	none
17	presbyopic	myope	no	reduced	none
18	presbyopic	myope	no	normal	none
19	presbyopic	myope	yes	reduced	none
20	presbyopic	myope	yes	normal	hard
21	presbyopic	hypermetrope	no	reduced	none
22	presbyopic	hypermetrope	no	normal	soft
23	presbyopic	hypermetrope	yes	reduced	none
24	presbyopic	hypermetrope	yes	normal	none

(e) rzutować dane na ten kierunek i dokonać przedstawienia graficznego klas,

(f) czy w tym rzucie można dokonywać prognozowania o rodzaju soczewki kontaktowej?

(g) Wyznaczyć pozostałe składowe główne i ich wariancje, co można powiedzieć o zależności/niezależności tych zmiennych ?

3. Rozważmy tablicę kontyngencji w tab. 1 wyznaczyć informację wzajemną, co możemy powiedzieć na temat zależności atrybutów?

4. wybrać atrybut, który dostarcza najwięcej informacji o zmiennej decyzyjnej, (wskazówka: wziąć pod uwagę miary przyrostu informacji I_H i I_G)

5 Grupowanie danych.

Przypomnienie: niech, $x, y \in \mathbb{R}^n$, $\langle x, y\rangle = x^T y = \sum_i x_i, y_i$ $\|x\| = \sqrt{\langle x, x\rangle}$. Niech $A = \{x_1, \ldots, x_{n_A}\}, x_i \in \mathbb{R}^n$ $\overline{A} = \frac{1}{n_A} \sum_{i=1}^{n_A} x_i$.

1. Dane są punkty $x_i \in \mathbb{R}$, $i = 1, \ldots, p$ znaleźć taki punkt c, dla którego suma kwadratów odległości punktów x_i od punktu c jest najmniejsza.

2. Dane są liczby $x_1, x_2, \ldots, x_n$ znaleźć taki punkt c dla którego suma odległości punktów x_i od punktu c jest najmniejsza.

3. Przeanalizować analogiczną sytuację w R^n w obu powyższych przypadkach przypadkach.

4. Sprawdzić czy kwadrat odległości euklidesowej jest metryką.

5. Pokazać, że całkowita suma kwadratów suma kwadratów odległości $T = \sum_{i,j} \|x_i - x_j\|^2$ da się przedstawić jako sumę kwadratów odległości wewnątrz klasowych i zewnątrz klasowych.

6. Niech odległość pomiędzy skupieniami będzie dana wzorem $D(A, B) = d_2(\overline{A}, \overline{B})$. Pokazać, że środek ciężkości po złączeniu skupień jest dany wzorem

$$\overline{AB} = \frac{n_A\overline{A} + n_B\overline{B}}{n_A + n_B}.$$

7. Niech jak w metodzie Warda

$$D(A_1A_2) = SSE_{A_1A_2} - (SSE_{A_2} + SSE_{A_2}),$$

gdzie $SSE_X = \sum_{i=1}^{n}(x_i - \overline{x_X})^T(x_i - \overline{x_X})$. Uzasadnić, że

$$D(A_1, A_2) = \frac{n_1n_2}{n_1 + n_2}(\overline{A_1} - \overline{A_2})^T(\overline{A_1} - \overline{A_2})$$

8. Przeanalizować związek pomiędzy metoda Warda a metodą środków ciężkości.

9. Które z miar „odległości" pomiędzy skupieniami są metrykami w pełnym znaczeniu tego słowa.

10. Uzasadnić, że algorytm K-środków zatrzyma się w skończonej liczbie kroków. W tym celu rozważyć funkcję: $F(\mathbf{S}) = \sum_{i=1}^{K}\sum_{x_j \in S_i} \|\mathbf{x}_j - \mathbf{c_i}\|^2$, gdzie $\mathbf{S} = \{S_1, \ldots, S_K\}$, $\mathbf{c_i} = \text{mean}(S_i) = \overline{S_i}$, oraz popatrzeć jak zmienia się ta funkcja w kolejnych krokach algorytmu.

11. Algorytm K-środków składa się z dwóch kroków: 1) aktualizacja środków 2) wyznaczenie nowych sąsiedztw $\mathbf{S}$ (przypisań). Zastanowić się czy przypisania mają jakąś szczególną postać czy też w ogólności mogą być zupełnie dowolną klasteryzacją. Spróbować oszacować na tej podstawie liczbę kroków algorytmu przynajmniej w pewnych szczególnych przypadkach (np. dane jednowymiarowe $K = 2, 3, \ldots$, dane dwuwymiarowe $K = 2$, itp.).

5.1 Metoda PCA

1. Wyznaczyć macierz kowariancji dla tab. 4

2. Wyznaczyć wektory własne i wartości własne macierzy kowariancji.

3. Rzutować zmienne na dwie pierwsze składowe główne i wynik przedstawić graficznie.

Tablica 4: Przykładowy zbiór danych.

X_1	X_1	X_1
0.28	1.41	0.82
0.47	0.54	0.27
0.39	1.23	0.67
0.35	0.84	0.51
0.60	1.68	0.95
0.24	1.11	0.71
0.45	0.70	0.41
0.42	0.44	0.20
0.24	0.74	0.37
0.34	0.71	0.37

Tablica 5: Decyzje i prawdopodobieństwa predykcji na przykładowym zbiorze testującycm.

klasa	0	1	1	0	1	0	0	1	0	0
predykcja $\Pr(d=1\|x)$	0.95	0.8	0.75	0.6	0.55	0.4	0.3	0.25	0.2	0.1

6 Klasyfikacja.

6.1 Miary jakości klasyfikacji

1. Dla danych z tabeli 5 wyznaczyć:
   (a) Przyjąć próg 0.5 oraz wyznaczyć macierz konfuzji.
   (b) Przyjąć próg 0.5 oraz wyznaczyć: czułość, precyzję oraz miarę F_1.
   (c) Przyjąć próg 0.5 oraz wyznaczyć: czułość, specyficzność oraz zbalansowaną dokładność.
   (d) Narysować krzywą ROC.
   (e) Wybrać na podstawie krzywej ROC klasyfikator optymalny z punku widzenia kosztów podejmowania błędnych decyzji, jeżeli za przypadki fałszywie negatywne płacimy trzy razy więcej niż za przypadki fałszywie pozytywne.

2. Jak wygląda krzywa ROC dla klasyfikatora bezregułowego (wskazówka: krzywa jest wyznaczone w tym przypadku przez dwa punkty.)

3. Załóżmy prawdopodobieństwa apriori klas decyzyjnych $\{+,-\}$ są równe odpowiednio p_+ oraz $p_- = 1-p_+$ oraz złóżmy dodatkowo, że punkt punkt współrzędnych (x,y) leży na krzywej ROC. Podać formułę na błąd klasyfikacji w terminach: x, y, p_+, p_-.

4. Załóżmy prawdopodobieństwa apriori klas decyzyjnych $\{+,-\}$ są równe odpowiednio p_+ oraz $p_- = 1-p_+$. Znajdź w przestrzeni ROC (w układzie: czułość, 1-specyficzność) miejsce geometryczne punktów o stałym błędzie klasyfikacji równym e.

5. Na podstawie wyników z poprzednich punktów podaj procedurę graficzna znajdowania klasyfikatora o największej dokładności, dla danych p_+ i p_-.

6.2 Klasyfikator bezregułowy

1. Przypuśćmy, że w populacji jest 90% osób zdrowych oraz 10% osób chorujących na pewną chorobę. Przypuśćmy, że klasyfikujemy pacjenta w ten sposób że rzucamy symetryczną monetą oraz jeżeli wypadł orzeł diagnozujemy: zdrowy, jeżeli reszka stawiamy diagnozę: chory. Obliczyć dokładność klasyfikacji takiego klasyfikatora (czyli prawdopodobieństwo pojęcia poprawnej decyzji).

2. Przypuśćmy, że mamy n klas decyzyjnych z prawdopodobieństwami $p_1, \ldots, p_n$[2, 3]. Rozważmy regułę decyzyjną która z prawdopodobieństwem q_1 podejmuje decyzje 1, z prawdopodobieństwem q_2 podejmuje decyzję 2, ..., z prawdopodobieństwem q_n podejmuje decyzję n. Należy wyznaczyć:
   (a) prawdopodobieństwo podjęcia poprawnej decyzji jako funkcję p i q ($Pr(p,q)$),
   (b) dla jakiego wektora q wyrażenie $Pr(p,q)$ przyjmuje wartość największą ?

3. Weźmy pod uwagę zmienną decyzyjną zbioru `contactLens`. Załóżmy, że pojawił się pacjent dla którego nie pomierzono żadnego z atrybutów warunkowych, jaką decyzje o wyborze soczewki rozsądnie byłoby podjąć?

[2] Rozważania te prowadzą do tzw. klasyfikatora bezregułowego (*Zero Rule*). Rozsądnie jest oceniać jakość dowolnego klasyfikatora (zwłaszcza tych słabej jakości) w stosunku do klasyfikatora bezregułowego.

[3] Należy mieć świadomość, że – mimo czasami wysokiej dokładności klasyfikacji – praktyczna przydatność klasyfikatorów tego typu jest bardzo ograniczona.

6.3 Naiwny klasyfikator bayesowski

1. Dla zbioru danych `contactLens` (tab. 3):
   (a) znaleźć odpowiednie prawdopodobieństwa brzegowe i warunkowe,
   (b) zbudować naiwny klasyfikator Bayesa,
   (c) policzyć przykładowo: $p(\text{none}|p_{13})$, $p(\text{soft}|p_{13})$, $p(\text{hard}|p_{13})$,
   (d) zbadać zależność dwóch wybranych atrybutów.

6.4 Regresja liniowa i logistyczna

1. Rozważmy dwie klasy decyzyjne $+1$ i -1 o rozkładzie normalnym $N(M_1, \Sigma)$ oraz $N(M_2, \Sigma)$ wyznaczyć:
   (a) równanie funkcji dyskryminującej obie klasy.
   (b) prawdopodobieństwa decyzji pod warunkiem x: $p(y = +1|x)$, oraz $p(y = -1|x)$
2. Rozważmy model regresji logistycznej

$$p(y = +1|x, w) = f(yw^T x) = \frac{1}{1 + \exp(-yw^T x)}$$

   wyznaczyć
   (a) funkcję wiarygodności
   (b) wektor pierwszych pochodnych (gradient) funkcji logistycznej, porównać wynik z krokiem aktualizacji wag algorytmu „reguła perceptronu"
   (c) macierz drugich pochodnych funkcji logistycznej

6.5 Drzewa decyzyjne

1. Wielkość $H(Y|X)$ nosi nazwę zanieczyszczenia (ang. *impurity*). Uzasadnić, że minimalizacja zanieczyszczenia jest równoważna maksymalizacji przyrostu informacji $I(X, Y)$.
2. Przeprowadzić dyskusją, którą z miar (zanieczyszczenie ($H(Y|X)$), przyrost informacji $(H(Y) - H(Y|X))$, czy informację wzajemna) jest najkorzystniejsza z obliczeniowego punktu widzenia przy konstrukcji drzewa decyzyjnego.
3. Dla danych `contactLens` (tab. 3) zbudować pełne drzewo decyzyjne stosując miarę zanieczyszczenia opartą na I_H lub I_G.
4. Przyciąc drzewo korzystając wykorzystując minimalizującą koszt $E + \alpha T$, gdzie E jest błędem na zbiorze uczącym a T jest rozmiarem drzewa, przyjąć $\alpha = 0.1$.

6.6 Sieci neuronowe, perceptron

1. Rozważmy algorytm uczenia perceptronu przedstawiony na rysunku 1.
   (a) Sprawdzić czy następujące dane:

$$x = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \qquad d = \begin{bmatrix} -1 & -1 & -1 & 1 \end{bmatrix}, \tag{1}$$

   są liniowo separowane.
   (b) Wykonać do końca algorytm dla danych 1d; jako wagi początkowe można przyjąć $b_0 = 1$ oraz $w_0 = [0 \quad 0]$,
   (c) wyznaczyć margines separacji.

```
1: function [W,B]=LEARNP(\{(X^i, d^i)\}_{i=1,...,p})
2:     n = 0;
3:     while n < p do (czy wszystkie próbki są poprawnie klasyfikowane)
4:         if d^i(⟨w_k, x^i⟩ + b_k) < 0 then (Czy próbka jest po właściwej stronie)
5:             w_{k+1} = w_k + d^i · X^i; (nie - korygujemy wagi)
6:             b_{k+1} = b_k + d^i · 1;
7:             n = 0; (zerujemy licznik stopu)
8:         else
9:             n++; (tak - zwiększamy licznik stopu)
10:        end if
11:    end while
12: end function
```

Rysunek 1: Algorytm uczenia perceptronu

(d) Uzasadnić, że algorytm nie zatrzyma się nigdy dla danych:

$$x^T = \begin{bmatrix} 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}, \qquad d^T = \begin{bmatrix} -1 & 1 & 1 & -1 \end{bmatrix}.$$

(e) Udowodnić, że algorytm ten zatrzyma się zawsze, jeżeli tylko zbiór wzorców będzie liniowo separowany. (wskazówka: oszacować z dołu rzut wektora wag na dowolną prostą separacją, oszacować z góry normę wektora wag, skorzystać z twierdzenia Cauchy'ego-Schwartz'a.)

7 Wyszukiwanie reguł i wzorców

7.1 Algorytm apriori

1. Dla tabeli danych 6 wykonać następujące czynności:
   (a) Znaleźć wsparcie następujących zbiorów: $\{A_1\}, \{A_1, A_2\}, \{A_1, A_2, A_4\}$.
   (b) Znaleźć zaufanie następujących reguł: $\{\emptyset \to A_1\}, \{A_1, A_2 \to A_5\}, \{A_1 \to A_2A_4\}, \{A_1A_2 \to A_4\}$.
   (c) Znaleźć wszystkie zbiory częste o $\min_{supp} = 4$ używając klasycznego algorytmu Apriori.
   (d) Znaleźć wszystkie zbiory częste o $\min_{supp} = 1$ używając drzewka wyliczającego podzbiory od prawej do lewej lub odwrotnie.
   (e) Wyznaczyć reguły o $min_{conf} > 0.6$.
   (f) Dla reguł znalezionych w poprzednich dwóch punktach: zaznaczyć na wykresie (wsparcie, zaufanie) reguły jako pojedyncze punkty.
   (g) Wyznaczyć wszystkiew reguły paretooptymalne.

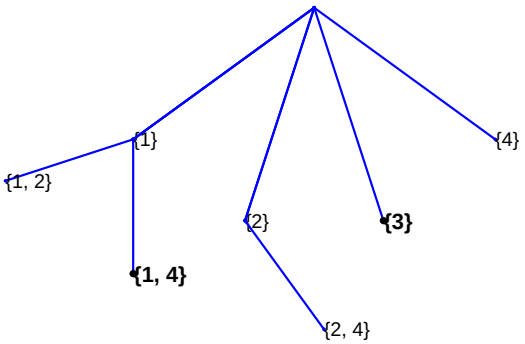
2. Dla tabeli danych 6 wykonac następujące czynności:
   (a) znaleźć wszystkie zbiory częste o $\min_{supp} = 5$ używając klasycznego algorytmu Apriori,
   (b) znaleźć wszystkie zbiory częste o $\min_{supp} = 1$ używając drzewka wyliczającego podzbiory od prawej do lewej lub odwrotnie,
   (c) wyznaczyć reguły $min_{conf} > 0.6$
   (d) dla reguł znalezionych w poprzednich dwóch punktach: zaznaczyć na wykresie reguły jako punkty (wsparcie, zaufanie) oraz wyznaczyć reguły pareto-optymalne.

Tablica 6: Zbiór danych, poszukiwanie podzbiorów częstych

lp.	A_1	A_2	A_3	A_4
1	1	1	0	0
2	0	0	0	1
3	1	1	0	0
4	1	1	0	1
5	1	0	0	0
6	0	0	0	0
7	0	1	0	1
8	0	1	0	1
9	0	1	0	1
10	0	0	0	1
11	1	1	0	0
12	0	1	0	1
13	0	0	0	0
14	0	0	0	0
15	0	0	0	0
16	0	1	0	1
17	0	0	1	0
18	1	0	0	0
19	0	0	0	1
20	1	1	0	1

Zbiory częste	wsparcie
$\emptyset$	20
$\{A_4\}$	10
$\{A_2\}$	10
$\{A_2, A_4\}$	7
$\{A_1\}$	7
$\{A_1, A_2\}$	5

Brzeg negatywny
$\{A_3\}$
$\{A_1, A_4\}$



Rys. Drzewko przeszukiwań wraz z brzegiem

Przypomnienie

Entropia

$$H(a) = -\sum_{p=1}^{k} \Pr\{a = p\} \log_2(\Pr\{a = p\})$$

Informacja wzajemna (przyrost informacji)

$$\begin{aligned} I_H(d,a) &= H(d) - H(d|a) \\ I_H(d,a) &= H(a) + H(d) - H(ad) \\ H(d|a) &= \sum_{p=1}^{k} \Pr\{a = p\} H(d|a = p) \end{aligned}$$

Indeks Gini'ego jako miara rozproszenia

$$G_{CART}(a) = 1 - \sum_{p=1}^{k} Pr\{a = p\}^2$$

$$I_{G_{CART}}(d,a) = G_{CART}(d) - G_{CART}(d|a)$$

$$G_{CART}(d|a) = \sum_{p=1}^{k} Pr\{a = p\} G_{CART}(d|a = p)$$

Tablica 7: Zbiór danych nr 2

A_1	A_2	A_3	A_4	A_5	A_6	A_7
0	1	1	1	1	1	1
1	1	0	0	1	1	1
0	0	0	1	0	0	1
1	1	1	0	1	1	0
1	1	0	1	1	1	0
1	0	0	0	1	0	0
0	0	1	0	0	0	1
0	0	1	1	0	0	1
0	0	1	1	0	1	1
1	0	0	0	1	0	0
1	1	0	1	1	1	1
1	0	0	1	1	1	0
0	1	1	1	1	1	1
0	1	0	1	1	1	1
0	1	0	1	0	0	0
0	1	0	0	0	0	0
1	0	1	1	1	1	1
1	1	1	1	1	1	1
1	0	1	1	1	1	1
0	0	1	1	0	0	1

χ^2

$$Q = \sum_{i=1}^{k} \sum_{j=1}^{l} \frac{n_{ij} - n_{i.} n_{.j}/n}{n_{i.} n_{.j}/n}$$

Statystyka Q ma rozkład χ^2 z $(k-1)(l-1)$ stopniami swobody.

Liniowa separowalność, marginesem separacji Mówimy, że zbiór wzorców $\{(X^i, d^i)\}_{i=1,\ldots,p}$ jest **linowo separowalny**, jeżeli istnieje prosta o parametrach (w, b) taka, że dla każdego $i = 1, \ldots, p$ zachodzi:

$$d^{(i)} \left(\sum_{j=1}^{n} w_j {x_j}^{(i)} + b \right) > 0$$

Niech $\gamma^{(i)} = d^{(i)} \left(\sum_{j=1}^{n} w_j {x_j}^{(i)} + b \right)$. **Marginesem separacji** prostej o parametrach (w, b), nazywamy liczbę

$$\gamma_{(w,b)} = \min_{i=1,\ldots,p} \gamma^{(i)}.$$

Jeżeli $\|w\| = 1$ wtedy margines jest minimalną odległością punku od prostej wziętą ze znakiem

Norma, iloczyn skalarny

- Iloczyn skalarny

$$\langle w, x \rangle = \sum_{i=1}^{n} w_i x_i$$

- Norma wektora

$$\|w\| = \sqrt{\sum_{i=1}^{n} {w_i}^2}$$

- Twierdzenie (Cauchy-Schwartz)

$$\langle w, x \rangle \leqslant \|w\| \cdot \|x\|$$

Wyliczanie podzbiorów

```
function rekurencjaL(atr,mm)
disp(num2str(atr))
if isempty(atr)
    mx=0;
else
    mx=max(atr);
end
for ii=mm:-1:mx+1
    rekurencjaL([atr ii],mm);
end
```