Indukowane Reguły Decyzyjne I

Wykład 11

IRD Wykład 11

- Plan
  - **Powtórka**
    - **Krzywa Lift i współczynnik Kappa**
    - Reguły asocjacyjne – ocena jakości reguł
  - Reguły asocjacyjne
    - Poszukiwanie reguł
    - Metoda A priori
    - Przykłady

Krzywa Lift – postać nieskumulowana

P(X)

pd_kat1	kategoria
0,620987	1
0,620987	1
0,620987	1
0,620987	1
0,620987	0
0,620987	1
0,620987	1
0,620987	0
0,620987	1
0,595925	0
0,56335	0
0,559108	1
0,559108	0
0,559108	1
0,559108	1
0,559108	0
0,559108	1
0,559108	0
0,559108	0
0,559108	0
0,559108	1
0,559108	1
0,559108	0
0,559108	0
0,559108	0
0,559108	0
0,559108	0
0,559108	0
0,559108	0
0,559108	0

0,01

0,01

0,01

---

udz1 = 100%

---

udz1 = 67%

---

udz1 = 67%

---

udz1 = 33%

Krzywa Lift – postać procentowa i liczbowa

P(X)

pd_kat1	kategoria
0,620987	1
0,620987	1
0,620987	1
0,620987	1
0,620987	0
0,620987	1
0,620987	1
0,620987	0
0,620987	1
0,595925	0
0,56335	0
0,559108	1
0,559108	0
0,559108	1
0,559108	1
0,559108	0
0,559108	1
0,559108	0
0,559108	0
0,559108	0
0,559108	1
0,559108	1
0,559108	0
0,559108	0
0,559108	0
0,559108	0
0,559108	0
0,559108	0
0,559108	0
0,559108	0

0,01 — udz1 = 100% → lift = 8,3

0,01 — udz1 = 67% → lift = 5,5

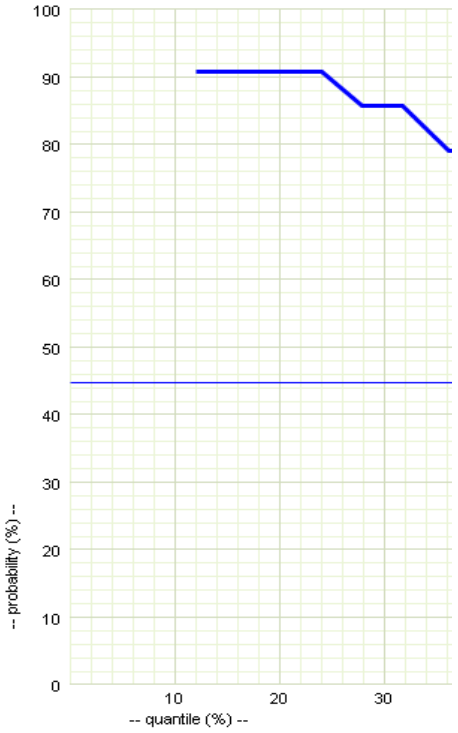
0,01 — udz1 = 67% → lift = 5,5

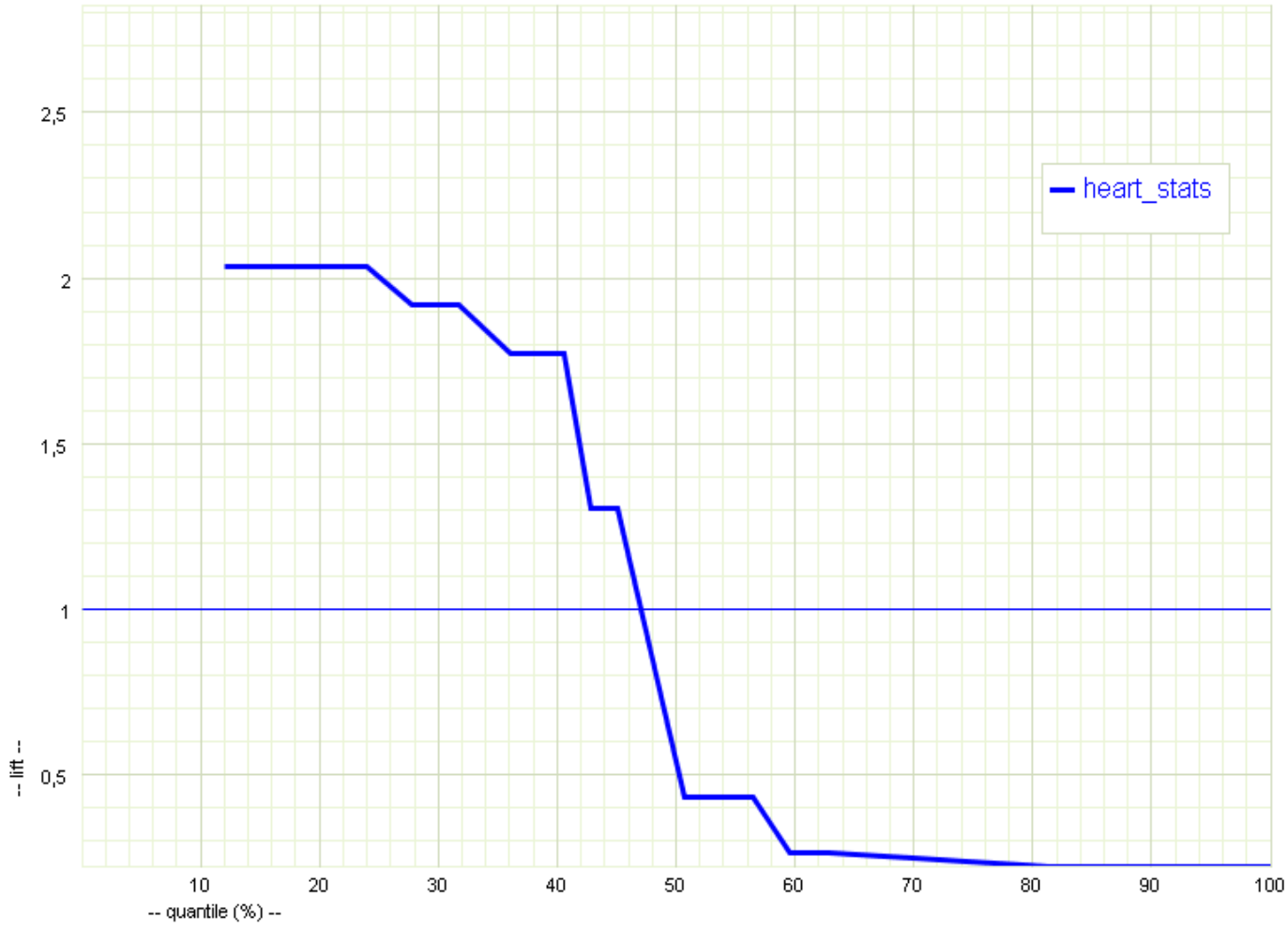
udz1 = 33% → lift = 2,75

a priori = 12%

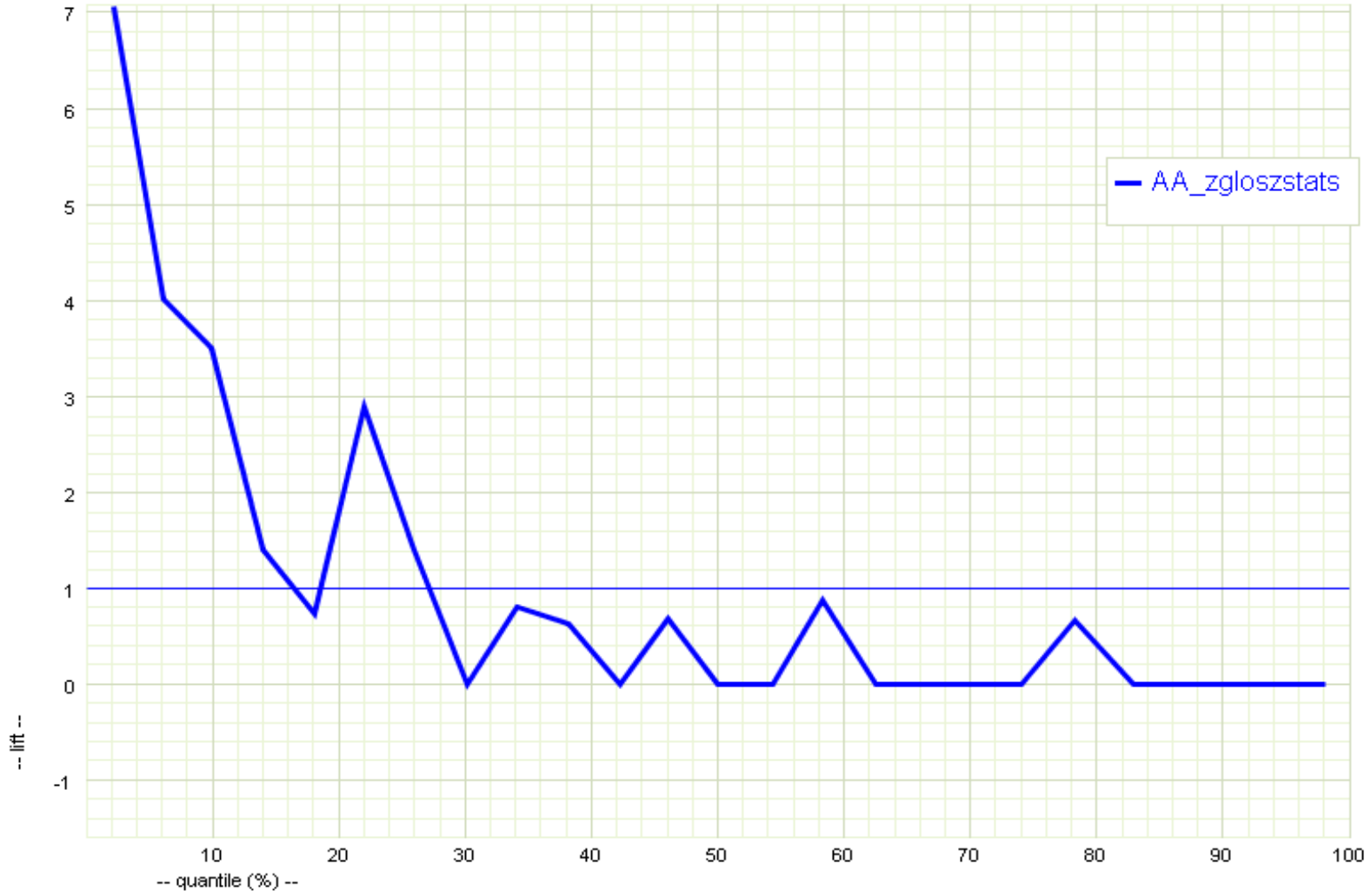
Lift = udz1 / a priori

Krzywa Lift – postać nieskumulowana

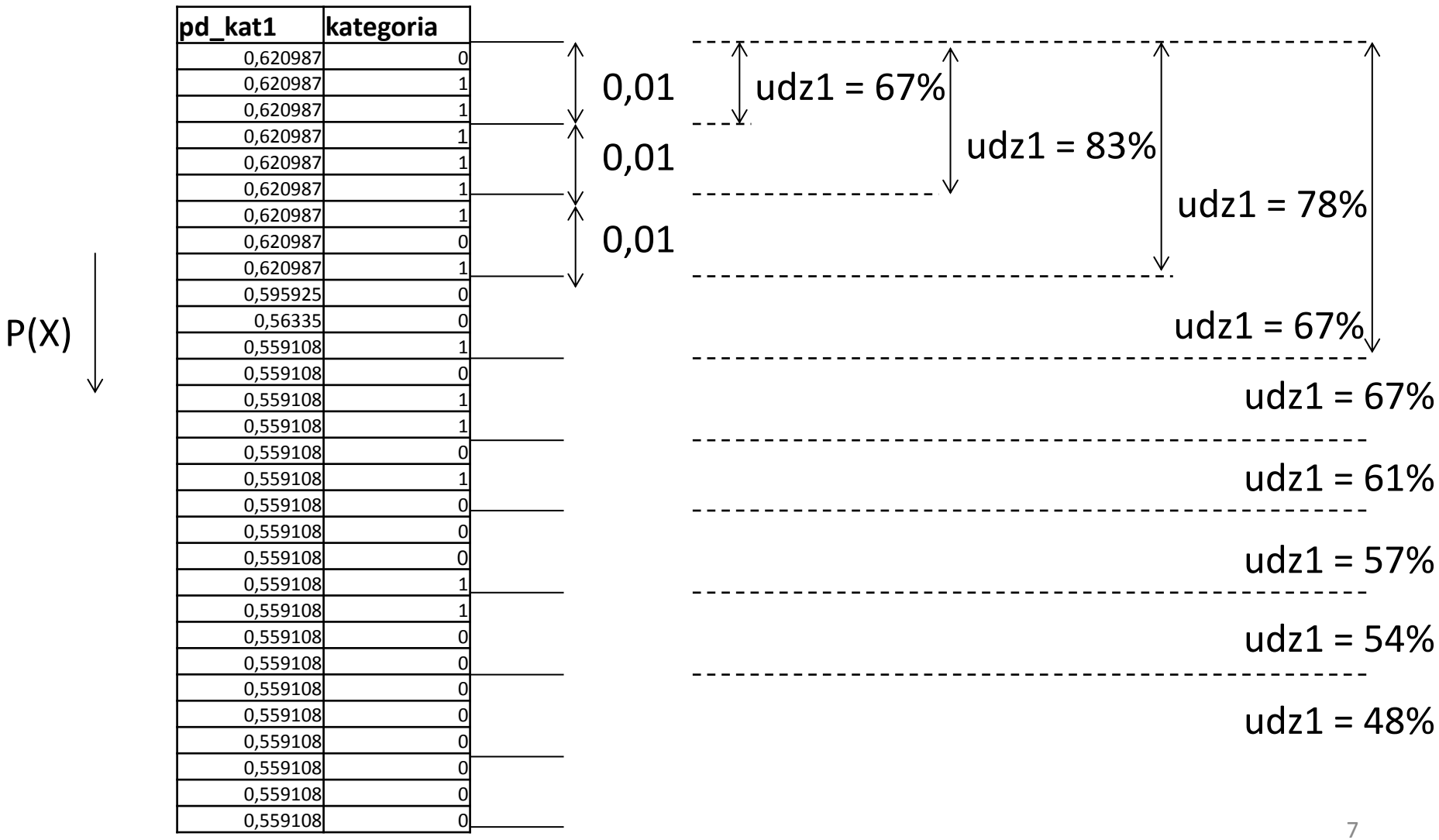




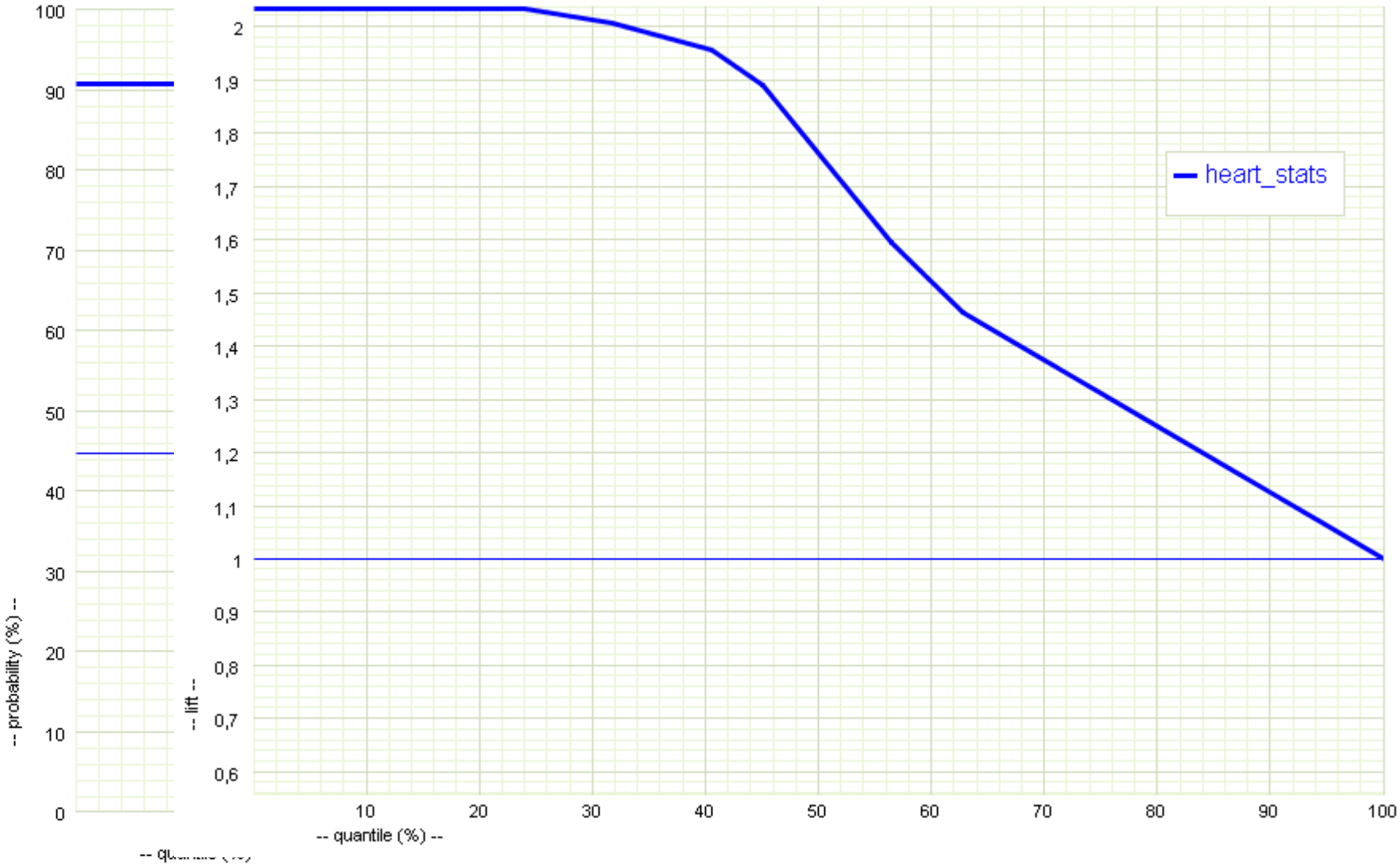
Krzywa Lift – postać nieskumulowana



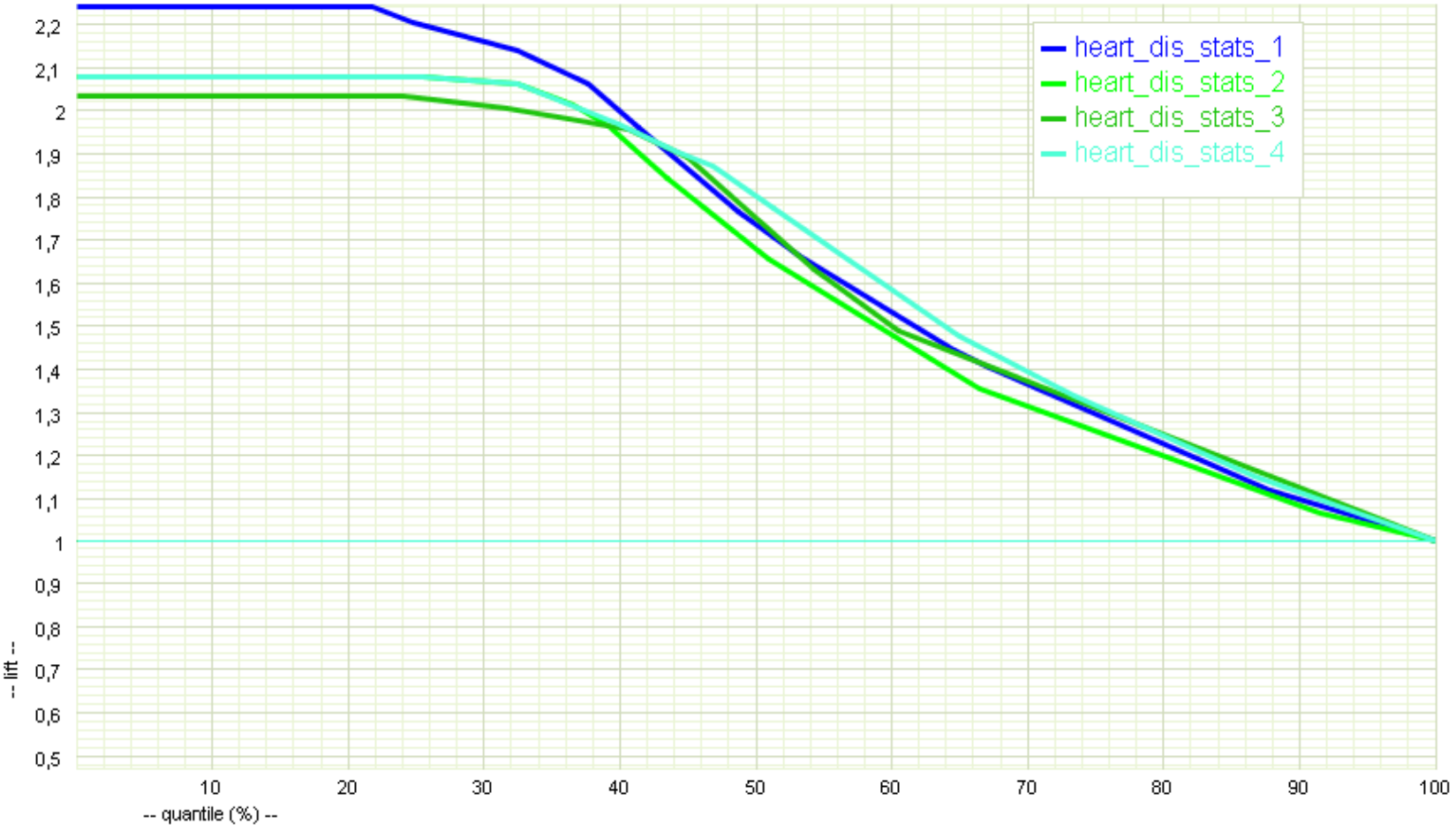
Krzywa Lift – postać skumulowana



Krzywa Lift – postać skumulowana



Choroby serca - Lift



Współczynnik Kappa

$$kappa = \frac{|X|_k - |X|_l}{|X| - |X|_l}$$

- gdzie

  $|X|_k$ - liczba poprawnych klasyfikacji zbudowanego klasyfikatora

  $|X|_l$ - liczba poprawnych klasyfikacji modelu losowego

  $|X|$ - całkowita liczba obserwacji.
- Kappa $\in$ <0,1> /z reguły/
- Kappa = 1, gdy klasyfikator jest idealny
- Kappa = 0, gdy klasyfikator jest losowy
- Interpretacja: jaką część błędnych klasyfikacji modelu losowego wyjaśnia badany model

IRD Wykład 11

- Plan
  - **Powtórka**
    - Krzywa Lift i współczynnik Kappa
    - **Reguły asocjacyjne – ocena jakości reguł**
  - Reguły asocjacyjne
    - Poszukiwanie reguł
    - Metoda A priori
    - Przykłady

Problem - formalnie

Transakcja	Artykuły
1	tygodnik, fantastyka, film, batonik
2	przewodnik
3	kosmetyk, miesięcznik, film, tygodnik

- Nomenklatura
  - D – zbiór transakcji:
  - I – zbiór wszystkich artykułów
  - I={{tygodnik}, {fantastyka},{film},{batonik}, {przewodnik}, {kosmetyk}, {miesięcznik}}: |I|>>1
  - $T_t \subseteq I$ – zbiór artykułów kupionych w transakcji t
  - T_1={{tygodnik}, {fantastyka},{film},{batonik}}, T_2={{przewodnik}}
- Reguła asocjacyjna

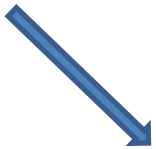
  Jeżeli A to B, $A \Rightarrow B$, {{tygodnik}}⇒{{film}}

  $A,B \subset I$; $A \cap B = \emptyset$

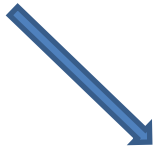
Algorytm A priori

Transakcja	Artykuły
1	tygodnik
1	fantastyka
1	film
1	batonik
2	film
2	przewodnik
3	kosmetyk
3	miesięcznik
3	film
3	tygodnik

Format transakcyjny



Dane binarne



Format macierzowy

transakcja	tygodnik	fantastyka	film	batonik	przewodnik	kosmetyk	miesięcznik
1	1	1	1	1	0	0	0
2	0	0	1	0	1	0	0
3	1	0	1	0	0	1	1

Ocena jakości reguł - support

- Wsparcie (ang. support) – miara częstości, wskazuje udział transakcji, w których występuje dany zestaw produktów:

  s(Z) = liczba transakcji zawierających zestaw Z /

  liczba wszystkich transakcji

transakcja	tygodnik	fantastyka	film	batonik	przewodnik	kosmetyk	miesięcznik
1	1	1	1	1	0	0	0
2	0	0	1	0	1	0	0
3	1	0	1	0	0	1	1

- Przykład: Z={{tygodnik}, {film}}

  s(Z) = 2/3

Support - własności

- Miara symetryczna

$$s(\text{tygodnik} \rightarrow \text{film}) = s(\text{film} \rightarrow \text{tygodnik}) = 0{,}67$$

transakcja	tygodnik	fantastyka	film	batonik	przewodnik	kosmetyk	miesięcznik
1	1	1	1	1	0	0	0
2	0	0	1	0	1	0	0
3	1	0	1	0	0	1	1

- Wartości nie są łatwo interpretowalne – uwaga na niskie wartości

$$s(A \text{ i } B) = 0{,}05 \text{ ale } s(A) = 0{,}45 \text{ i } s(B) = 0{,}55$$

Ocena jakości reguł - confidence

- Ufność c (ang. confidence) – udział transakcji spełniających regułę wśród transakcji zawierających poprzednik

$$c(A \rightarrow B) = \text{liczba transakcji spełniających regułę} / \text{liczba transakcji zawierających A}$$

transakcja	tygodnik	fantastyka	film	batonik	przewodnik	kosmetyk	miesięcznik
1	1	1	1	1	0	0	0
2	0	0	1	0	1	0	0
3	1	0	1	0	0	1	1

- Przykład:

$c(\text{film} \rightarrow \text{tygodnik}) = 2/3$

$c(\text{tygodnik} \rightarrow \text{film}) = 1/1=1$

Confidence - własności

- Wartości nie są łatwo interpretowalne – uwaga na niskie i na wysokie wartości
- Przykłady
  - c(A → B)=0,2 – niska wartość sugeruje, że produkt B jest rzadko kupowany z produktem A

  ale s(A → B)=0,02
  - c(A → B)=0,65 – wysoka wartość sugeruje, że produkt B jest często kupowany z produktem A

  ale s(A → B)=0,95

Ocena jakości reguł - lift

- Miara zależności / niezależności poprzednika i następnika reguły

$$Lift(A \rightarrow B) = s(A \rightarrow B) / (s(A) \cdot s(B))$$

- Określa
  - Występowanie zależności – czy kupno produktu A wpływa na zakup produktu B
  - Kierunek – czy wpływ jest pozytywny, czy negatywny
  - Siłę – w jakim stopniu wzrasta / maleje prawdopodobieństwo zakupu produktu B, jeśli klient kupił produkt A

Lift - własności

- Miara symetryczna

$$s(\text{tygodnik} \rightarrow \text{film}) = s(\text{film} \rightarrow \text{tygodnik}) = 0{,}67$$

$$s(\text{tygodnik}) = 0{,}67$$

$$s(\text{film}) = 1$$

$$\text{lift}(\text{tygodnik} \rightarrow \text{film}) = \text{lift}(\text{film} \rightarrow \text{tygodnik}) =$$

$$= 0{,}67 / (0{,}67 \cdot 1) = 1$$

transakcja	tygodnik	fantastyka	film	batonik	przewodnik	kosmetyk	miesięcznik
1	1	1	1	1	0	0	0
2	0	0	1	0	1	0	0
3	1	0	1	0	0	1	1

Lift - własności

- Tygodnik i batonik

$$s(\text{tygodnik} \rightarrow \text{batonik}) = s(\text{batonik} \rightarrow \text{tygodnik}) = 0{,}33$$

$$s(\text{tygodnik}) = 0{,}67 \qquad s(\text{batonik}) = 0{,}33$$

$$\text{lift}(\text{tygodnik} \rightarrow \text{batonik}) = \text{lift}(\text{batonik} \rightarrow \text{tygodnik}) =$$

$$= 0{,}33 / (0{,}67 \cdot 0{,}33) = 1{,}5 > 1$$

- Tygodnik i przewodnik

$$s(\text{tygodnik} \rightarrow \text{przewodnik}) = s(\text{przewodnik} \rightarrow \text{tygodnik}) = 0$$

$$\text{lift}(\text{tygodnik} \rightarrow \text{batonik}) = \text{lift}(\text{batonik} \rightarrow \text{tygodnik}) = 0$$

transakcja	tygodnik	fantastyka	film	batonik	przewodnik	kosmetyk	miesięcznik
1	1	1	1	1	0	0	0
2	0	0	1	0	1	0	0
3	1	0	1	0	0	1	1

Lift - własności

- Lift(A → B) > 0
- Jeżeli lift(A → B) = 1, to zakup produktu A nie wpływa na kupno produktu B
- Jeżeli lift(A → B) > 1, to zakup produktu A zwiększa prawdopodobieństwo kupna produktu B
- Jeżeli lift(A → B) < 1, to zakup produktu A zmniejsza prawdopodobieństwo kupna produktu B
- ZADANIE 2

IRD Wykład 11

- Plan
  - Powtórka
    - Krzywa Lift i współczynnik Kappa
    - Reguły asocjacyjne – ocena jakości reguł
  - **Reguły asocjacyjne**
    - **Poszukiwanie reguł**
    - Metoda A priori
    - Przykłady

Jakość reguł, A⇒B

- Wsparcie s (ang. *support*) – odsetek transakcj zawierający A oraz B

$$s = P(A \subset T \wedge B \subset T) = \frac{Liczba\ transakcji\ zawierajacych\ A\ i\ B}{Liczba\ wszystkich\ transakcji}$$

1	A	B	
2	A	B	
3	A	B	
4	A	B	
5	A	B	
6	A	Y	Z
7	A	Y	
8	Z	Y	
9	Z	Y	
10	Z	Y	

- Ufność c (ang. *confidence*) – odsetek transakc zawierających A, który jednocześnie zawiera B

$$c = P(B \subset T \mid A \subset T) = \frac{P(A \subset T\ \&\ B \subset T)}{P(A \subset T)} = \frac{Liczba\ transakcji\ zawierajacych\ A\ i\ B}{Liczba\ transakcji\ zawierajacych\ A}$$

c(A⇒B)=5/7, c(B⇒A)=1
s(A⇒B)=0,5, s(B⇒A)=0,5
c(A⇒Y)=2/7, c(Y⇒A)=2/5
s(A⇒Y)=0,2, s(Y⇒A)=0,2
c(Z⇒Y)=1, c(Y⇒Z)=4/5
s(Z⇒Y)=0,4, s(Y⇒Z)=0,4

Poszukiwanie reguł asocjacyjnych

Założenia:

- Szukamy reguł **mocnych** – o odpowiednio wysokim wsparciu i ufności
  - *Sprzedaż sklepowa*: s>20%, c>70% oznacza, że …

  reguła dotyczy przynajmniej 20% klientów i jest trafna na 70%
  - *Wykrywanie nadużyć, terroryzmu*: s>1%, bo …

  jest mało takich transakcji

Poszukiwanie reguł asocjacyjnych

- Szukamy w **częstych** zbiorach zdarzeń
  - *Zbiór zdarzeń* – podzbiór I (zbioru wszystkich artykułów)
  - *Częstość zbioru zdarzeń* – l. transakcji zawierająca dany zbiór
  - *Zbiór częsty* - o częstości większej od wielkości progowej ϕ
  - F_k – zbiór częstych zbiorów zdarzeń o k elementach

IRD Wykład 11

- Plan
  - Powtórka
    - Krzywa Lift i współczynnik Kappa
    - Reguły asocjacyjne – ocena jakości reguł
  - **Reguły asocjacyjne**
    - Poszukiwanie reguł
    - **Metoda A priori**
    - Przykłady

Poszukiwanie reguł asocjacyjnych

- Właściwość *A priori*
  - Jeśli zbiór zdarzeń Z nie jest częsty, to dla dowolnego elementu A, zbiór Z$\cup$A nie będzie częsty.

Metoda A priori

- Koncepcja
  1. Znajdź wszystkie częste zbiory zdarzeń (o częstości $\geq\phi$)
  2. Na podstawie pkt. 1 utwórz reguły spełniające warunki na c i s
- Algorytm
  1. k=1. W zbiorze atrybutów I znajdź częste zbiory 1-elementowe F_1
  2. Zmień k=k+1,
  3. Ustal zbiór kandydatów na zbiory częste C_k poprzez sumy zbiorów F_{k-1} (jeśli zbiory różnią się jednym tylko elementem),
  4. Przytnij C_k do F_k przy pomocy właściwości A priori. Przejdź do pkt. 2, chyba że zbiór F_k jest pusty,
  5. Dla zbiorów F_2-$F_{|I|}$ wyznacz wszystkie reguły asocjacyjne i znajdź najlepsze przy pomocy granicznych c i s.

IRD Wykład 11

- Plan
  - Powtórka
    - Krzywa Lift i współczynnik Kappa
    - Reguły asocjacyjne – ocena jakości reguł
  - **Reguły asocjacyjne**
    - Poszukiwanie reguł
    - Metoda A priori
    - **Przykłady**

Przykład działania metody A priori

Transakcja	film			historia	języki	muzyka	poradnik
1	0			1	1	1	1
2	0			0	0	0	0
3	1			0	0	1	1
4	1			0	0	1	1
5	1			1	1	0	0
Freq.	3			2	2	3	3

1. ϕ=2 Zbiory częste jednoelementowe

F_1={{film}, {historia}, {języki}, {muzyka}, {poradnik}}

Przykład działania metody A priori

2. k=2, Zbiór kandydatów na zbiory częste C_2:

	film	**historia**	**języki**	**muzyka**	**poradnik**
film	–	1	1	2	2
historia		–	2	1	1
języki			–	1	1
muzyka				–	3
poradnik					–

Przykład działania metody A priori

3. Zbiory częste:

   F_2={{film, muzyka}, {film, poradnik}, {historia, języki}, {muzyka, poradnik}}

4. Reguła 1: Jeżeli film to muzyka

   s(R1)=0,4; c(R1)=0,67

5. Reguła 2: Jeżeli film to poradnik

   s(R2)=0,4; c(R2)=0,67

Podaj i oceń kolejne reguły dla k=2

Przykład działania metody A priori

6. k=3, Zbiór kandydatów na zbiory częste

C_3={{film, muzyka, historia}, {film, muzyka, języki}, {film, muzyka, poradnik}, {film, poradnik, historia}, {film, poradnik, języki}, …}

7. F_3={{film, muzyka, poradnik}}

	film & muzyka	**film & poradnik**	**historia & języki**	**muzyka & poradnik**
film	–	–	1	1
historia	0	0	–	1
języki	0	0	–	1
muzyka	–	2	1	–
poradnik	2	–	1	–

Przykład działania metody A priori

8. Reguły dla F_3={{film, muzyka, poradnik}}

Reguła 9: Jeżeli film & muzyka to poradnik

s(R9) = 0,4 c(R9) = 1

Reguła 10: Jeżeli film & poradnik to muzyka

s(R10) = 0,4 c(R10) = 1

Reguła 11: Jeżeli muzyka & poradnik to film

s(R11) = 0,4 c(R11) = 0,67

Transakcja	film	tygodnik	miesięcznik	historia	języki	muzyka	poradnik
1	0	0	0	1	1	1	1
2	0	1	0	0	0	0	0
3	1	0	0	0	0	1	1
4	1	0	0	0	0	1	1
5	1	0	1	1	1	0	0
Freq.	3	1	1	2	2	3	3

Zadanie 7

Transakcja	Produkty
1	muzyka, poradnik, film
2	przewodnik, beletrystyka, muzyka
3	film, beletrystyka
4	film, muzyka, beletrystyka, poradnik
5	muzyka, film
6	beletrystyka
7	beletrystyka, przewodnik, film
8	beletrystyka, przewodnik, film, poradnik
9	film, muzyka, przewodnik, poradnik
10	przewodnik, beletrystyka, muzyka, film

Zadanie 7

Transakcja	muzyka	film	poradnik	przewodnik	beletrystyka
1	1	1	1	0	0
2	1	0	0	1	1
3	0	1	0	0	1
4	1	1	1	0	1
5	1	1	0	0	0
6	0	0	0	0	1
7	0	1	0	1	1
8	0	1	1	1	1
9	1	1	1	1	0
10	1	1	0	1	1
suma	6	8	4	5	7

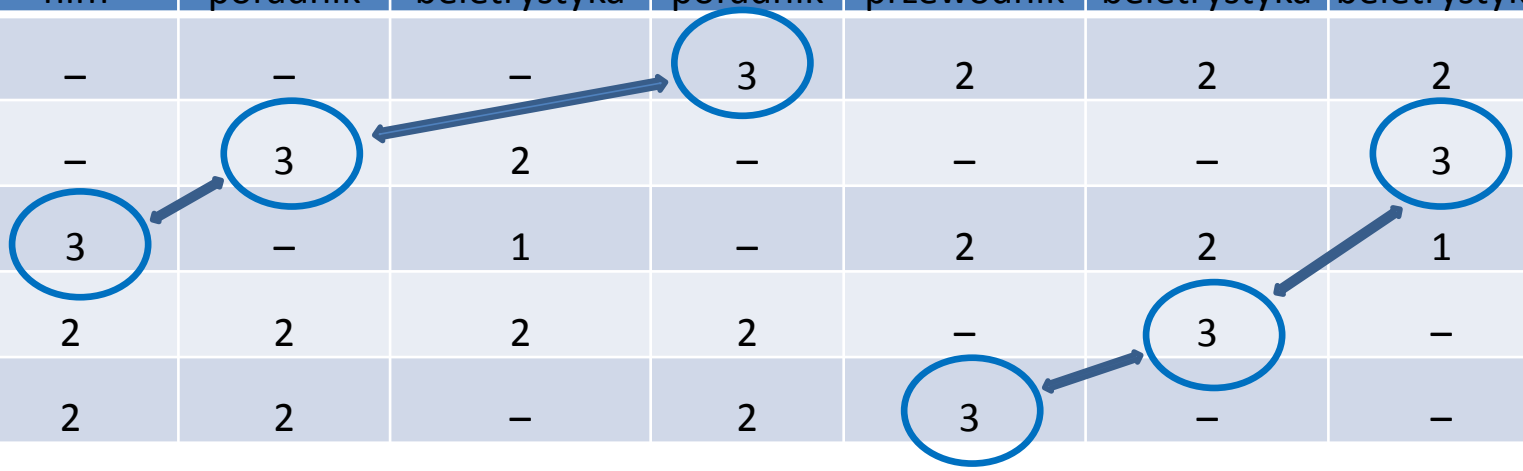
Zadanie 7

	muzyka	film	poradnik	przewodnik	beletrystyka
muzyka	–	5	3	2	3
film		–	4	4	5
poradnik			–	2	2
przewodnik				–	4
beletrystyka					–

- C_2 – wszystkie pary
- F_2 = {{muzyka, film}, {muzyka, poradnik}, {muzyka, beletrystyka}, {film, poradnik}, {film, przewodnik}, {film, beletrystyka}, {przewodnik, beletrystyka}}
- Reguły R1-R14

Zadanie 7

	muzyka & film	muzyka & poradnik	muzyka & beletrystyka	film & poradnik	film & przewodnik	film & beletrystyka	przewodnik & beletrystyka
muzyka	–	–	–	3	2	2	2
film	–	3	2	–	–	–	3
poradnik	3	–	1	–	2	2	1
przewodnik	2	2	2	2	–	3	–
beletrystyka	2	2	–	2	3	–	–



- C_3
- F_3 = {{muzyka, film, poradnik}, {film, przewodnik, beletrystyka}}
- Reguły R15-R20

Zadanie 7

	muzyka & film & poradnik	film & przewodnik & beletrystyka
muzyka	–	1
film	–	–
poradnik	–	1
przewodnik	1	–
beletrystyka	1	–

- C_4
- F_4 – pusty

Zadanie 7

	muzyka & film	muzyka & poradnik	muzyka & beletrystyka	film & poradnik	film & przewodnik	film & beletrystyka	przewodnik & beletrystyka
muzyka & film	–	–	–	–	–	–	1
muzyka & poradnik		–	–	–	1	1	0
muzyka & beletrystyka			–	–		–	–
film & poradnik				–	–	–	1
film & przewodnik					–	–	–
film & beletrystyka						–	–
przewodnik & beletrystyka							–

- C_4
- F_4 – pusty

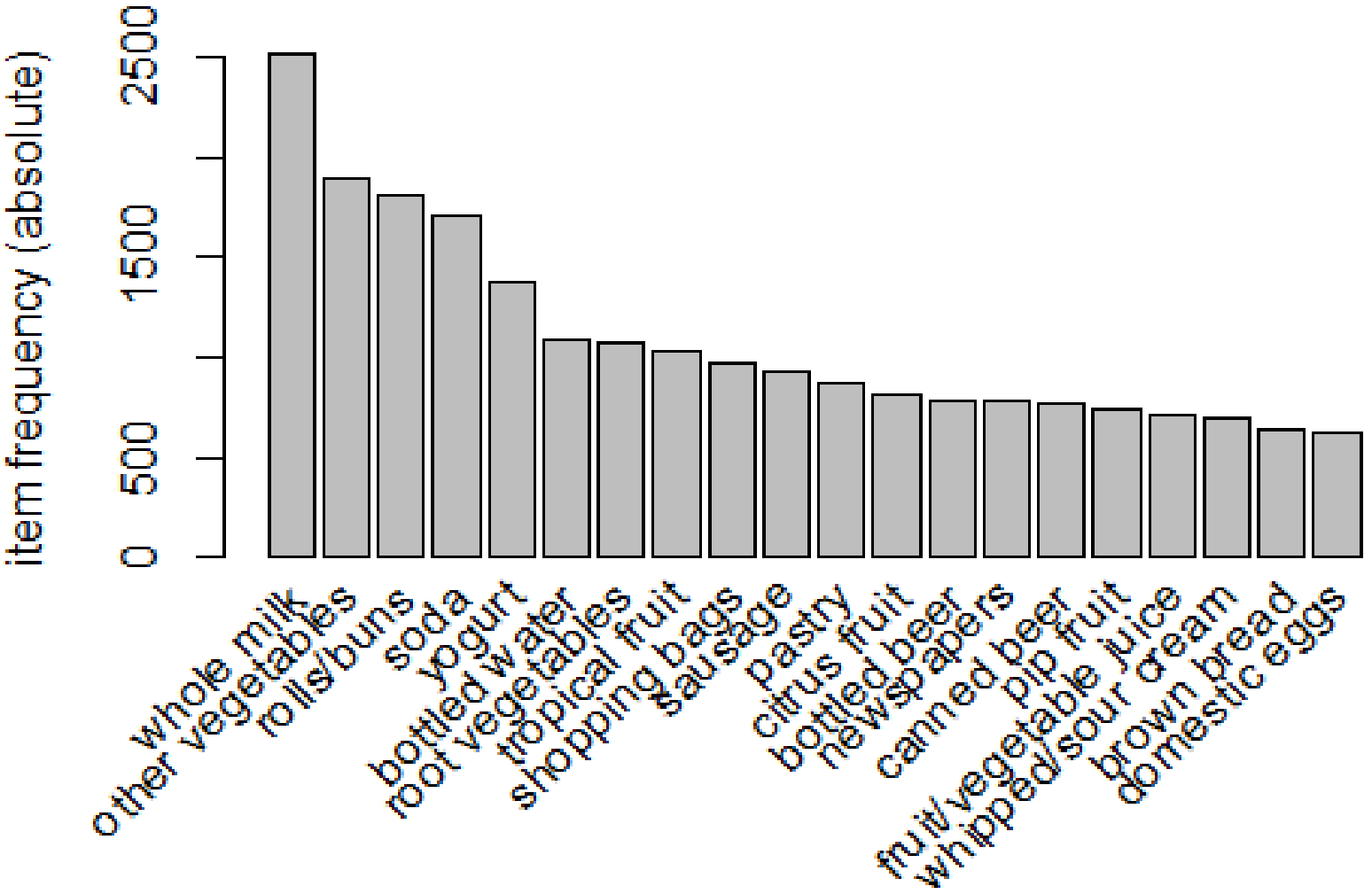
R – pakiet arulesViz

- Dane "Groceries " – 9835 wierszy, 169 kolumn

citrus fruit	semi-finished bread	margarine	ready soups	
tropical fruit	yogurt	coffee		
whole milk				
pip fruit	yogurt	cream cheese	meat spreads	
other vegetables	whole milk	condensed milk	long life bakery product	
whole milk	butter	yogurt	rice	abrasive cleaner
rolls/buns				
other vegetables	UHT-milk	rolls/buns	bottled beer	liquor (appetizer)
pot plants				
whole milk	cereals			
tropical fruit	other vegetables	white bread	bottled water	chocolate
citrus fruit	tropical fruit	whole milk	butter	curd

R – pakiet arulesViz

- Dane "Groceries " – 9835 wierszy, 169 kolumn



Uwagi końcowe

- Algorytmy zdefiniowano dla zmiennych binarnych, ale można go rozszerzyć na zmienne nominalne
- Algorytm generuje wiele reguł, należy je odfiltrować przy pomocy miar wsparcia i ufności
- Reguły bywają zwodnicze.

  [Przyrost trafności względem pierwotnego prawdopodobieństwo może być niewielki, stąd zamiast *ufności*, można używać *przyrostu ufności*]

- Jeśli z góry ustalimy następnik, to reguły asocjacyjne stanowią przykład uczenia nadzorowanego

Kolokwium

- Reguły Asocjacyjne
  - Wsparcie/support
  - Ufność/confidence
  - Lift
  - Właściwość Apriori
  - Zbiory częste
- Literatura:
  - Larose D. (2006) Odkrywanie wiedzy z danych, PWN
  - Witten I.H., Frank E. (2005) Data mining, Morgan Kauffman