DS.DM - przykładowe zadania sprawdzające

1. Niech {{*bcde*}, {*bcdf*}, {*bcef*}, {*bdef*}, {*cdef*}} będą wszystkim częstymi zbiorami pozycji o długości 4. Które z poniższych zdań jest (są) prawdziwe?

   a) {*de*} jest częsty
   b) być może {*de*} jest częsty, ale nie jest to pewne
   c) {*de*} nie jest częsty
   d) {*bcdef*} jest częsty
   e) być może {*bcdef*} jest częsty, ale nie jest to pewne
   f) {*bcdef*} nie jest częsty

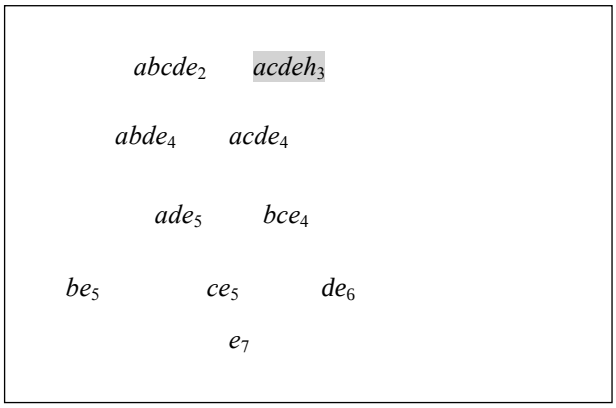
2. Rozważmy transakcyjny zbiór danych z tab. 1. Wyznacz wartości dla następujących miar:

TId	*Pozycje*
1	{*abce*}
2	{*abdeh*}
3	{*abefh*}
4	{*bcefh*}
5	{*acde*}
6	{*abcdefh*}
7	{*aefh*}
8	{*bcefh*}

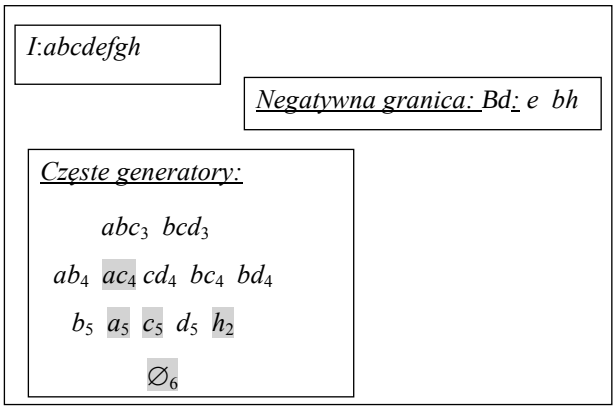
Tab. 1. Transakcyjny zbiór danych

- wsparcie *sup*({*fh*}→{*b*}) = 4
- wsparcie względne *rSup*({*fh*}) = 5/8
- wsparcie względne *rSup*({*b*}) = 6/8
- zaufanie *conf*({*fh*}→{*b*}) = 4/5
- lift *lift*({*fh*}→{*b*}) = (4/5) : (6/8)
- współczynnik pewności *cf*({*fh*}→{*b*}) = (4/5 - 6/8) : (1 - 6/8)
- współczynnik zależności *df*({*fh*}→{*b*}) = $\left(\frac{4}{5}-\frac{6}{8}\right):\left(\frac{\min\{\frac{5}{8},\frac{6}{8}\}}{\frac{5}{8}}-\frac{6}{8}\right)$
- Czy {*b*} zależy od {*fh*}?

3. Rozważ transakcyjny zbiór danych z tab. 1. Wyznacz zaufanie reguły z negacją: {*e*}→{$\bar{h}$}: *conf*({*fh*}→{$\bar{b}$}) = 1/5
4. Rozważ transakcyjny zbiór danych z tab. 1. Które zbiory pozycji są domknięciami zbioru {*bfh*}? γ({*bfh*}) = {*befh*}
5. Rozważ transakcyjny zbiór danych z tab. 1. Które zbiory pozycji są generatorami zbioru {*bfh*}? *G*({*bfh*}) = {{*bf*}}
6. Na podstawie reprezentacji opartej na częstych zbiorach zamkniętych (CR) z rys. 1, określ, czy {*ach*} jest zbiorem częstym. Jeśli tak, określ jego wsparcie. Jeśli nie, podaj największą możliwą wartość wsparcia zbioru {*ach*}. Częsty, sup = 3
7. Na podstawie reprezentacji opartej na generatorach (GR) z rys. 2, określ, czy {*ach*} jest zbiorem częstym. Jeśli tak, określ jego wsparcie. Jeśli nie, podaj największą możliwą wartość wsparcia zbioru {*ach*}. Częsty, sup = 2.



Rys. 1. CR: Częste zbiory zamknięte *FC*



Rys. 2. GR: Częste generator *FG* & negatywna granica *Bd*

8. Dotyczy operatora pokrycia: Wyznacz liczbę reguł asocjacyjnych należących do pokrycia reguły: ∅→{*abcde*}? |*C*(∅→{*abcde*})| = $3^5 - 2^5$.
9. Rozważ reguły reprezentatywne RR z Tab. 2:

rule identifier	rule	support	confidence	
1	∅→{*abe*} [4,4/5]	4	4/5	
2	∅→{*bcde*} [4,4/5]	4	4/5	
3	{*a*}→{*bcde*} [3,3/4]	3	¾	
4	{*c*}→{*abde*} [3,3/4]	3	¾	
5	{*d*}→{*abce*} [3,3/4]	3	¾	

- Które z tych reguł pokrywają regułę {*ae*}→{*b*}? #1, #3
- Podaj oszacowanie wsparcia i zaufania reguły {*ae*}→{*b*} na podstawie wsparć i zaufań pokrywających ją reguł reprezentatywnych RR:

  support ≥ 4, confidence ≥ 4/5

Tab. 2. Reguły reprezentatywne RR

10. **Dotyczy algorytmu SPADE:** Poniżej zamieszczono listy identyfikatorów transakcji (tidlisty) wspierające wybrane sekwencje.

t<(*c*)(*b*)>	
Cid	**TId**
1	10
2	20
3	5
4	5
4	15

t<(*c*)(*c*)>	
CId	**TId**
1	10
1	20
2	5
2	10
2	15

t<(*b*)(*c*)>	
CId	**TId**
1	10
2	10
2	20
3	5
4	15

t<(*c*)(*bc*)>	
Cid	**TId**
1	10
2	10
2	15
3	20
4	25

t<(*c*)>	
Cid	**TId**
1	10
1	15
3	5
3	55
5	25

t<(*b*)>	
CId	**TId**
1	10
2	5
3	5
3	15
5	20

a) Jakie jest wsparcie sekwencji: <(*c*)(*c*)>? 2

b) Które tidlisty są używane w SPADE do utworzenia tidlisty *t*<(*c*)(*b*)(*c*)>?
t<(*c*)(*b*)>, *t*<(*c*)(*c*)>*

c) Utwórz *t*<(*c*)(*b*)(*c*)>:

CId	TId
1	20

11. **Dotyczy algorytmu GSP (Generalized Sequential Patterns):**

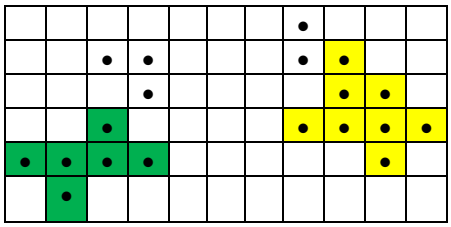
CId	TId	Items
1	1	*ac*
1	2	*ab*
1	5	*cdei*
1	8	*fgh*
1	9	*gi*

Niech *windowSize* = 2, *minGap* = 2, *maxGap* = 4.

Czy sekwencja danych klienta o identyfikatorze CId = 1 (zamieszczona w tabeli po lewej stronie) wspiera sekwencję kandydującą <(*bc*) (*ci*) (*hi*)>?

Tak

12. **Dotyczy algorytmu DBSCAN:**



Niech *minPts* = 4, a promień *Eps* = 1. Zaznacz na rysunku z lewej strony grupy i szum, które zostałyby wyznaczone przy użyciu algorytmu DBSCAN, przy założeniu, że zastosowano miarę odległości Euklidesowej (czyli, $distance(P_1,P_2) = [|x_1-x_2|^2 + |y_1-y_2|^2]^{½}$).

Grupa 1 – zb. punktów w zielonym obszarze, grupa 2 – zb. punktów w żółtym obszarze, szum – zb. pozostałych punktów.

13. **Dotyczy użycia nierówności trójkąta przy wyznaczaniu sąsiedztwa epsilonowego:** Niech D będzie zbiorem dwu-wymiarowych punktów jak w tab. 3, dla których wyznaczono ich odległości Euklidesowe do punktu referencyjnego *r*.

Tab. 3. Zbiór D dwu-wymiarowych punktów, uporządkowany względem ich odległości Euklidesowych do punktu referencyjnego *r* (wraz z informacją o tych odległościach)

point q	X	Y	distance(q,r)
K	0.9	0.0	0.9
L	1.0	1.5	1.9
G	0.0	2.4	2.4
H	2.4	2.0	3.1
F	1.1	3.0	3.2
C	2.8	3.5	5.0
A	**4.2**	**4.0**	**5.8**
B	5.9	3.9	6.1

Załóżmy, że punkty w D są uporządkowane ze względu na ich odległości Euklidesowe do punktu *r*. Niech *Eps* = 1.0, a A będzie punktem, którego otoczenie epsilonowe należy efektywnie wyznaczyć z użyciem własności nierówności trójkąta.

a) Dla których punktów w zbiorze D trzeba wyznaczyć pesymistyczne oszacowanie ich odległości Euklidesowych do punktu A? B, C, F

b) Dla których punktów w zbiorze D trzeba wyznaczyć ich rzeczywiste odległości Euklidesowe do punktu A? B, C

Tab. 4. Tablica decyzyjna

Car #	*Price*	*Mileage*	*Size*	*MaX-speed*	*d*
1	*medium*	*medium*	*full*	*low*	*poor*
2	*high*	*high*	*full*	*low*	*poor*
3	*low*	*low*	*full*	*low*	*good*
4	*medium*	*medium*	*full*	*low*	*good*
5	*high*	*low*	*compact*	*high*	*excellent*
6	*high*	*low*	*full*	*high*	*excellent*
7	*low*	*low*	*full*	*low*	*excellent*

Tab. 5. Tablica decyzyjna, zredukowana ze względu na obiekt klasyfikowany T

Car#	*Price*	*Mileage*	*Size*	*MaX-speed*	*d*	*ocenaPrzyna-leżności*
1					*poor*	½
2	*high*	*high*			*poor*	
3					*good*	0/2
4					*good*	
5	*high*		*compact*	*high*	*excellent*	2/3
6	*high*			*high*	*excellent*	
7					*excellent*	

14. **Dotyczy klasyfikacji leniwej z użyciem wzorców kontrastowych:** Niech *d* będzie atrybutem decyzyjnym w tabeli 4. Niech T będzie klasyfikowanym obiektem, o następujących wartościach atrybutów warunkowych: (*high, high, compact, high*).

a) Podaj w tabeli 5 zawartość tej tablicy decyzyjnej po redukcji ze względu na obiekt T.

b) W oparciu o tę zredukowaną tablicę decyzyjną wyznacz minimalne wzorce kontrastowe dla każdej z poniższych klas decyzyjnych:
- *poor*: $\{Mileage_{high}\}$
- *good*: brak
- *excellent*: $\{Size_{compact}\}$, $\{MaX\text{-}Speed_{high}\}$

c) Jaka jest wartość *ocenyPrzynależności* obiektu T do klasy decyzyjnej *Decision= poor*?

d) Jaka jest wartość *ocenyPrzynależności* obiektu T do klasy decyzyjnej *Decision= good*?

e) Jaka jest wartość *ocenyPrzynależności* obiektu T do klasy decyzyjnej *Decision= excellent*?

f) Do której klasy decyzyjnej zostanie zaklasyfikowany obiekt T? excellent