

Zadanie 1

Dana jest tablica przedstawiona poniżej. CustID oznacza identyfikator klienta, TID oznacza identyfikator transakcji.

CustID	TID	Items
1	1	A, D, E
1	2	A, B, C, E
2	3	A, B, D, E
2	4	A, C, D, E
3	5	B, C, E
3	6	B, D, E
4	7	C, D
4	8	A, B, C
5	9	A, D, E
5	10	A, B, E

- a) Oblicz wsparcie zbiorów: {E}, {B, D} oraz {B, D, E}.
b) Oblicz ufność następujących reguł asocjacyjnych:

$B, D \rightarrow E$

$E \rightarrow B, D$

- c) Odnajdź wszystkie zbiory częste dla minsup 0.4

Zadanie 2

Dany jest następujący zbiór przykładów.

X	Y	Atrybut_decyzyjny
T	F	B
T	F	B
F	F	A
F	T	B
F	T	A
T	T	B
F	T	A
F	T	A
F	F	A
T	F	B

- a) Skonstruuj drzewo decyzyjne dla podanych przykładów za pomocą algorytmu ID3 i miary information gain.
b) Zbuduj klasyfikator Bayesa dla tego zestawu. Jeżeli prawdopodobieństwo warunkowe wyjdzie 0, to przyjmij wartość 0.01. Zaklasyfikuj przykład $X=F, Y=T$.

Zadanie 3

Dany jest zbiór punktów w przestrzeni jednowymiarowej: (6, 12, 18, 24, 30, 42, 48).

- (a) Dla podanego zbioru centroidów (means) (18, 45) utwórz 2 klastry poprzez przypisanie obiektów (punktów) do najbliższego centroidu oraz oblicz sumaryczny błąd średniokwadratowy dla każdego z klastrów.
- (b) Czy podział zbiorów obiektów na otrzymane 2 klastry jest stabilny? To znaczy, czy jeżeli algorytm k-means zacznie swoje działanie przyjmując jako początkowe środki klastrów podane powyżej centroidy, to nastąpi zmiana przypisania obiektów do klastrów? Uzasadnij krótko odpowiedź.
- (c) Przeprowadź algorytm k-means rozpoczynając od centroidów 42 i 48 aż dojdiesz do podziału stabilnego.

Zadanie 4

Dana jest tablica podobieństwa (similarity matrix) pomiędzy obiektami p1, p2, p3, p4 i p5.

	p1	p2	p3	p4	p5
p1	1	0.1	0.41	0.55	0.35
p2	0.1	1	0.64	0.47	0.98
p3	0.41	0.64	1	0.44	0.85
p4	0.55	0.47	0.44	1	0.76
p5	0.35	0.98	0.85	0.76	1

Zastosuj metodę grupowania hierarchicznego do pogrupowania obiektów przyjmując dwie miary odległości dwóch klastrów: „minimum distance” oraz „maximum distance”.

Przedstaw wynik grupowania w postaci dendrogramów!

Dendrogramy powinny czytelnie wskazać w jakiej kolejności są grupowane obiekty dla tych dwóch algorytmów grupowania.

Zadanie 5

Dana jest następująca relacja opisująca klientów naszego sklepu internetowego zainteresowanych produktami muzycznymi:

ID	Wiek	Dochód	Żonaty/Zamężna	kupi_komputer
1	21	wysoki	nie	nie
2	29	wysoki	nie	tak
3	33	wysoki	nie	tak
4	41	niski	nie	nie
5	44	niski	tak	tak
6	34	niski	tak	tak
7	28	średni	tak	tak
8	34	średni	nie	nie
9	20	wysoki	nie	nie
10	39	wysoki	nie	tak

Problem: Znajdź wielowymiarowe reguły asocjacyjne, które można wyprowadzić z powyżej relacji zakładając minsup=40 % (co najmniej 4 transakcje) oraz minconf=50 %

Przyjmij schemat dyskretyzacji atrybutu **Wiek** „przedziały o równej szerokości” (tzw. equiwidth discretization schema): (20 – 35), (36 - 50).

Zadanie 6

Dany jest następujący zbiór trenujący klasyfikatora k-NN.

X	Y	KLASA
1	5	C1
6	2	C1
4	4	C1
7	6	C1
2	1	C1
5	1	C1
4	8	C2
7	3	C2
5	5	C2
3	2	C2
8	7	C2
8	4	C2

Zaklasyfikuj obiekt (6,6) metodą 1-NN za pomocą metryki L_1 (Manhattan) i 6-NN za pomocą metryki L_∞ (maksimum).

Zadanie 7

Przeprowadzono test dokładności na klasyfikatorze binarny wykrywającym czy pacjent jest chory na anginę (poniżej tabela z wynikami). Zbuduj macierz pomyłek i oblicz precision i recall.

Klasa	Zakl jako	Klasa	Zakl jako	Klasa	Zakl jako	Klasa	Zakl jako
Chory	Chory	Chory	Zdrowy	Zdrowy	Zdrowy	Chory	Zdrowy
Zdrowy	Zdrowy	Chory	Chory	Zdrowy	Zdrowy	Chory	Chory
Zdrowy	Chory	Zdrowy	Zdrowy	Chory	Zdrowy	Chory	Chory
Chory	Chory	Chory	Zdrowy	Zdrowy	Zdrowy	Zdrowy	Zdrowy
Zdrowy	Zdrowy	Zdrowy	Zdrowy	Chory	Chory	Zdrowy	Zdrowy
Chory	Chory	Zdrowy	Zdrowy	Chory	Chory	Zdrowy	Chory

Zadanie 8

Dany jest następujący zbiór danych przedstawiony w tabeli poniżej.

Tid	Items
1	A, B, D, E
2	B, C, E
3	A, B, D, E
4	A, B, C, E
5	A, B, C, D, E
6	B, C, D

Tid oznacza identyfikator klienta, Items oznacza listę produktów zakupionych przez klienta.

Znajdź wszystkie zbiory częste w powyższej tabeli zakładając minsup=30%. Znajdź wszystkie reguły asocjacyjne z największych zbiorów częstych jakie znajdziesz, przy minconf=80%

Zadanie 9

Dany jest następujący zbiór danych przedstawiony w tabeli poniżej.

Tid	Items
1	A, B, D, E
2	B, C, E
3	A, B, D, E
4	A, B, C, E
5	A, B, C, D, E
6	B, C, D

Oblicz macierz podobieństwa każdej pary klientów stosując współczynnik Jaccarda. Pogrupuj klientów hierarchicznym algorytmem aglomeracyjnym obliczając podobieństwo pomiędzy zgrupowaniami za pomocą metody minimal similarity.