DATA MINING 2014 **Laboratorium 10 (MIARY ODMIENNOŚCI MIĘDZY OBIEKTAMI, ANALIZA SKUPIEŃ)**

10.0

a) Wygeneruj 2 wektory x i y którego współrzędnymi są zmienne losowe normalne o średniej równej 5 i odchyleniu standardowym równym 2. Oblicz odmienność między obiektami x i y używając jako miary:

- odległości euklidesowej,
- metryki maksimum,
- odległości Canberry,
- odległości w normie L_3.

b) Wygeneruj 2 wektory binarne x i y. Niech i-ta współrzędna przyjmuje wartość 1 z prawdopodobieństwem 0.7. Oblicz odmienność między obiektami x i y używając jako miary:

- odległości Hamminga,
- unormowanej odległości Hamminga,
- miara Jaccarda.

c) Wygeneruj następujące dane mieszane:

```
x1 <- as.logical(rbinom(10,1,0.5))
x2 <- sample(letters, 10, replace=TRUE)
x3 <- rnorm(10)
x4 <- ordered(cut(x3, -4:4, include.lowest=TRUE))
xx <- data.frame(x1, x2, x3, x4, stringsAsFactors = FALSE)
```

Oblicz miarę odległości między 10 obiektami opisanymi przez dane `xx` używając współczynnika Gowera. Skorzystaj n.p. z funkcji `gower.dist` w pakiecie `StatMatch`.

10.1

Zbiór *kwadraty.txt* zawiera sztucznie wygenerowane dane dwuwymiarowe, składające się z czterech niezależnych prób (każda o liczności 20) z rozkładów jednostajnych na 4 kwadratach o środkach usytuowanych na przekątnej większego kwadratu. Na podstawie wykresu rozproszenia danych można zaproponować prawdopodobną liczbę skupień $k = 4$.

a) Zastosuj metodę k-średnich. Na wykresie rozproszenia zaznacz kolorami otrzymane skupienia. Oblicz: sumę kwadratów odległości między punktami w skupieniach. Podaj liczbę elementów w poszczególnych skupieniach.

b) Zastosuj metodę hierarchiczną dla odmienności:

- typu najbliższego sąsiada (ang. single linkage),
- typu najdalszego sąsiada (ang. complete linkage),

- typu średnia (ang. average linkage).

Dokonaj przycięcia dendrogramu na poziomie skupień $k = 4$. Sporządź wykres rozproszenia z zaznaczeniem kolorami skupień. Sporządź dendrogram.

10.2
Dane *wrecord.dat* zawierają rekordy krajowe w wybranych konkurencjach biegowych- są to zmienne: **100m**, **200m**, **400m**, **800m**, **1500m**, **3000m**, **marathon**. Czasy dla trzech najkrótszych biegów podane są w sekundach, pozostałe czasy w minutach.

Dokonaj zamiany minut na sekundy dla czasów biegów długich. Następnie wartości każdej zmiennej podziel przez odchylenie standardowe tej zmiennej.

a) Zastosuj metodę hierarchiczną z odmiennością najdalszego sąsiada. Przytnij dendrogram aby otrzymać 3 skupienia. Wyznacz środki 3 skupień.

b) Zastosuj metodę k-średnich, podając jako wartości początkowe parametru `centers` środki skupień wyznaczone w punkcie (a).

c) Wyznacz dwie pierwsze składowe główne. Sporządź wykres (w pierwszych dwóch składowych); każde ze skupień oznacz innym kolorem.

10.3
Zastosuj metodę `Mclust` do danych *kwadraty.txt*.

10.4
Dane *congress.txt* zawierają macierz rozbieżności w głosowaniach dla 15 kongresmenów, dotyczących 19 głosowań (3 możliwe wyniki głosowania: za, przeciw, wstrzymanie się od głosu). Odległość między dwoma kongresmenami jest wyznaczona przez liczbę głosowań, w których głosowali różnie. Wśród głosujących znajdują się: demokraci (D) i republikanie (R).

a) Wykonaj hierarchiczną analizę skupień z odległością średnią, przedstaw wyniki za pomocą dendrogramu.

b) Dokonaj skalowania wielowymiarowego (funkcja `cmdscale`) i wyrysuj wykres na płaszczyźnie, podając nazwisko każdego z głosujących.

10.5
Dane w pliku *nci.data* zawierają wartości ekspresji genów. W eksperymencie mamy 6380 genów (wiersze macierzy) oraz 64 rodzaje nowotworów obserwowanych u różnych pacjentów. Celem analizy jest stwierdzenie: które rodzaje nowotworów są do siebie najbardziej podobne? Wykonaj analizę skupień korzystając z metody hierarchicznej z odległością średnią.