Кластерный анализ

Домашнее задание 1

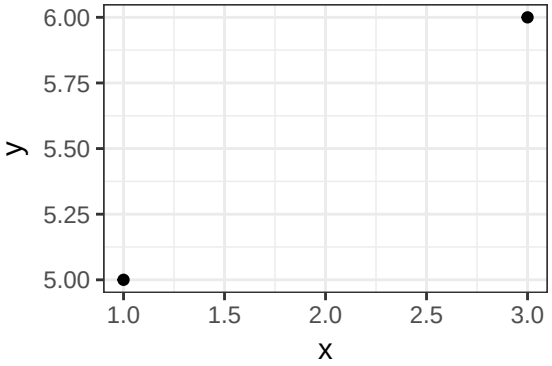
Алла Тамбовцева

Задание 1

Прочитайте в Analysis of Multivariate Social Science Data (Bartholomew et al.) страницы 17-29 (нумерация в pdf-файле 28-40).

Задание 2

Даны две точки x_1 и x_2:



Определите расстояние между этими точками, используя:

- евклидово расстояние
- квадрат евклидова расстояния
- манхэттенское расстояние
- расстояние Чебышёва

Задание 3

Проверьте, могут ли следующие показатели быть использованы в качестве метрики:

- коэффициент корреляции r
- $1 - r$, где r – коэффициент корреляции

Задание 4

Какие из перечисленных ниже матриц могут быть матрицами расстояний? Обоснуйте свой ответ.

$$A = \begin{pmatrix} 0 & 2 & 3 \\ 3 & 0 & 5 \\ 3 & 5 & 0 \end{pmatrix}; B = \begin{pmatrix} 0 & 2 & 4 \\ 2 & 1 & 5 \\ 4 & 5 & 0 \end{pmatrix}; C = \begin{pmatrix} 0 & 1.5 & 6 \\ 1.5 & 0 & 5 \\ 6 & 5 & 0 \end{pmatrix}; D = \begin{pmatrix} 0 & -1.5 & 6 \\ -1.5 & 0 & 5 \\ 6 & 5 & 0 \end{pmatrix}$$

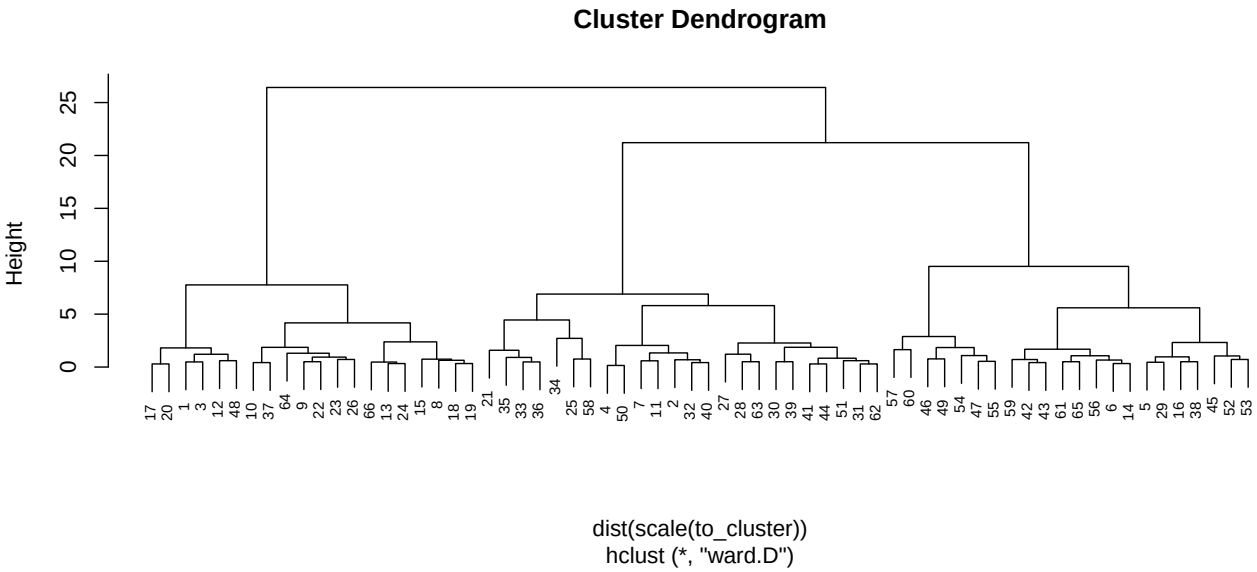
Задание 5

Какое максимальное число кластеров наблюдений можно выделить на основании представленной ниже дендрограммы, если

- в каждом кластере должно быть не менее 5 наблюдений?
- наблюдения номер 53 и 57 должны быть в одном кластере?
- наблюдения номер 4 и 31 должны быть в одном кластере?

Все три пункта выше – отдельные, не должны выполняться одновременно.

Дендрограмма:



Задание 6

Дана небольшая база данных aulatlong.csv по географическим координатам 10 городов Австралии.

```
df <- read.csv("aulatlong.csv")
```

Выберите из нее столбцы latitude и longitude (без использования библиотеки dplyr).

```
to_clust <- subset(df, select = c('latitude', 'longitude'))
rownames(to_clust) <- df$X # названия строк по названиям городов
```

Реализуйте иерархический кластерный анализ на основе данных в to_clust, используя евклидово расстояние (квадрат евклидова расстояния) и следующие способы агрегирования:

- метод ближнего соседа (метод одиночной связи, single)
- метод дальнего соседа (метод полной связи, complete)
- метод средней связи (average)
- метод Варда (ward.D)

Сравните полученные дендрограммы. Все пять диаграммы можно вывести одновременно следующим образом, по 3 графика в ряд:

```
par(mfrow = c(2, 3)) # для одновременного вывода нескольких графиков

# код для построения дендрограммы 1
# код для построения дендрограммы 2
```

```
# код для построения дендрограммы 3
# код для построения дендрограммы 4
# код для построения дендрограммы 5
```

Подумайте об особенностях каждого метода агрегирования, об их достоинствах и недостатках.