

Белорусский Государственный Университет
Информатики и Радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра ЭВМ

Лабораторная работа №3

Тема «Кластерный анализ»

Выполнил:

Студент группы 7М2431

Зайцев Ю.В.

Проверил:

Марченко В.В.

Минск, 2017

Задание:

Входные данные: n объектов, каждый из которых характеризуется двумя числовыми признаками: $\{x_i\}_{i=1}^n$ и $\{y_i\}_{i=1}^n$, а также номером класса $\{c_i\}_{i=1}^n$

Требуется исследовать работу алгоритмов кластеризации объектов наблюдения по двум признакам. Для каждого набора данных необходимо выполнить следующие задания:

1. Провести кластеризацию объектов наблюдения с помощью алгоритма k внутригрупповых средних.

2. Графически изобразить на плоскости разбиения объектов наблюдения в соответствии с кластерами. Также отметить центры каждого кластера. Количество кластеров должно соответствовать количеству классов.

3. Для разбиения на кластеры вычислить сумму квадратов расстояний от каждого объекта наблюдения до центра соответствующего кластера.

Исходные данные:

n_1	a_1	R_1	n_2	a_1	R_2
100	$\begin{pmatrix} -1 \\ 1 \end{pmatrix}$	$\begin{pmatrix} 1 & 0.1 \\ 0.1 & 2 \end{pmatrix}$	50	$\begin{pmatrix} 2 \\ 4 \end{pmatrix}$	$\begin{pmatrix} 2 & 0.1 \\ 0.1 & 1 \end{pmatrix}$

Где независимые случайные векторы (X, Y) , n_1 из которых относятся к первому классу, а n_2 – ко второму классу. Векторы, относящиеся к первому классу, распределены по гауссовскому закону с математическим ожиданием a_1 и корреляционной матрицей R_1 , а векторы, относящиеся ко второму классу – по гауссовскому закону с математическим ожиданием a_2 и корреляционной матрицей. R_2 .

Название файла: wine.csv

Ссылка: <http://archive.ics.uci.edu/ml/datasets/Wine>

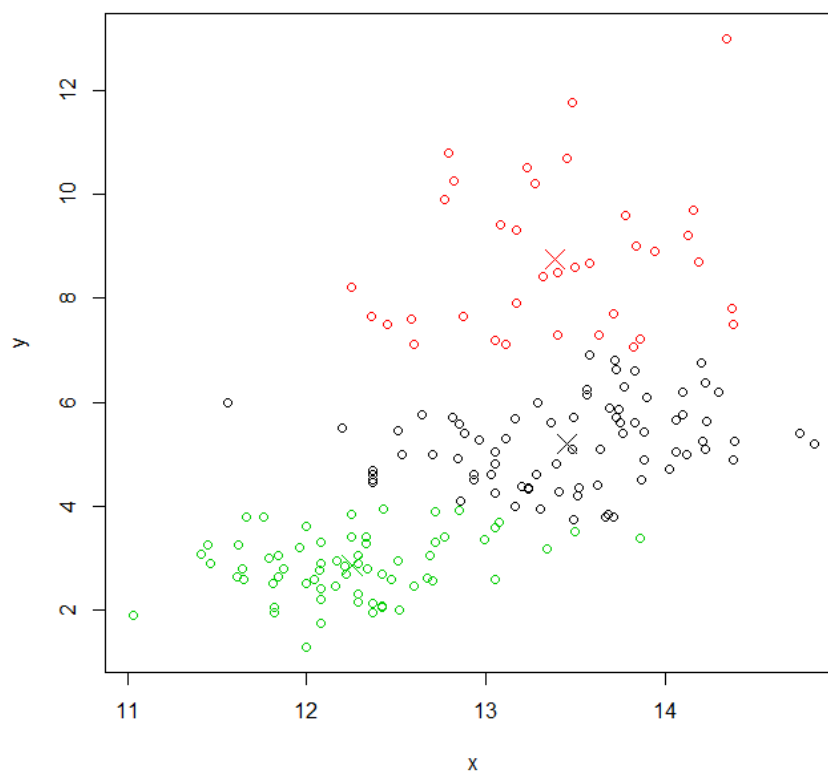
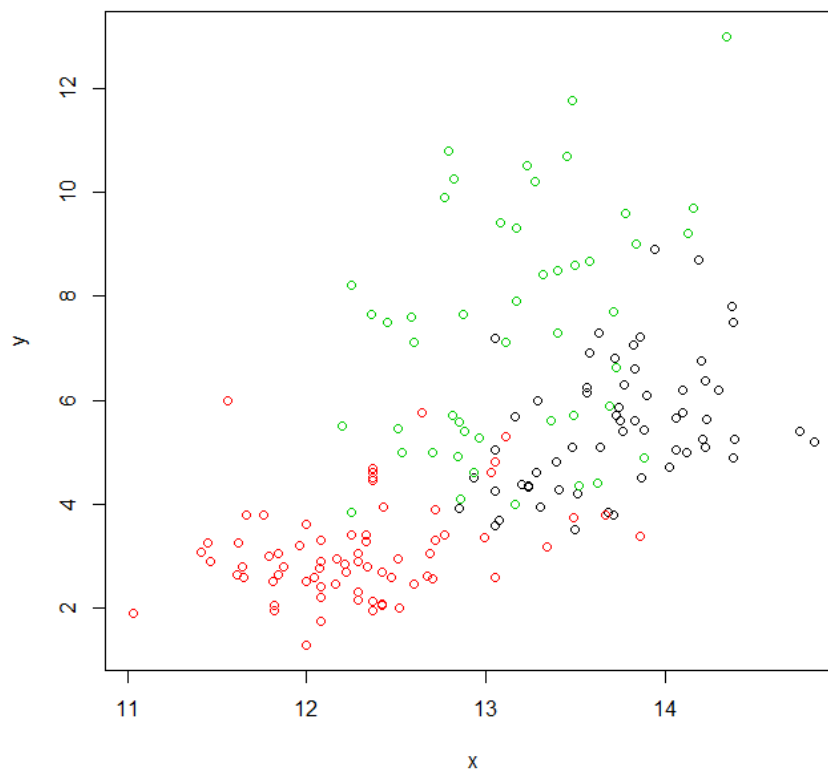
Первый признак: alcohol (столбец № 2)

Второй признак: color-intensity (столбец № 11)

Класс cultivar (столбец №1)

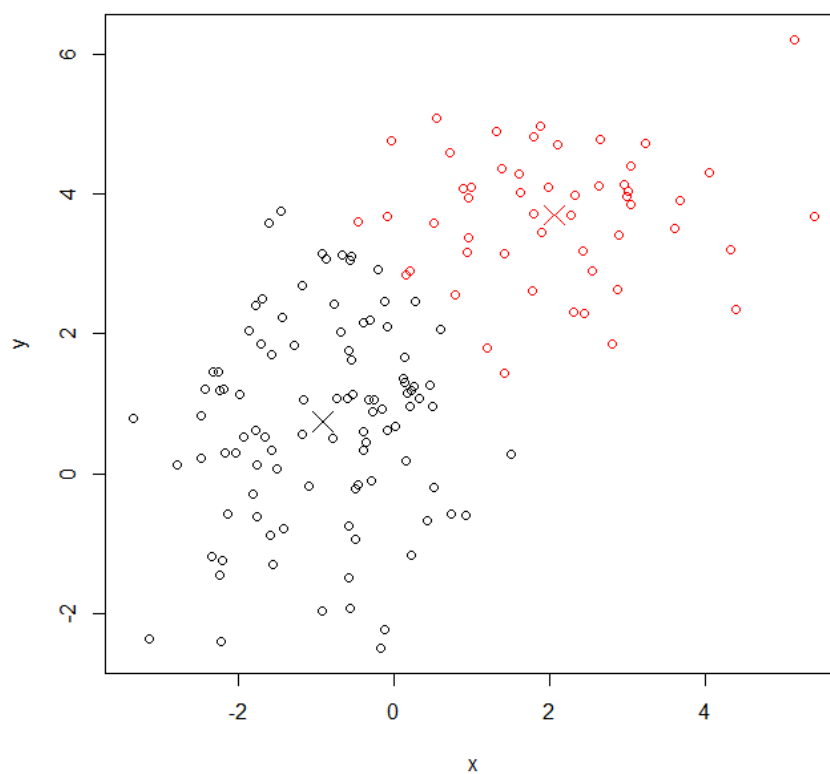
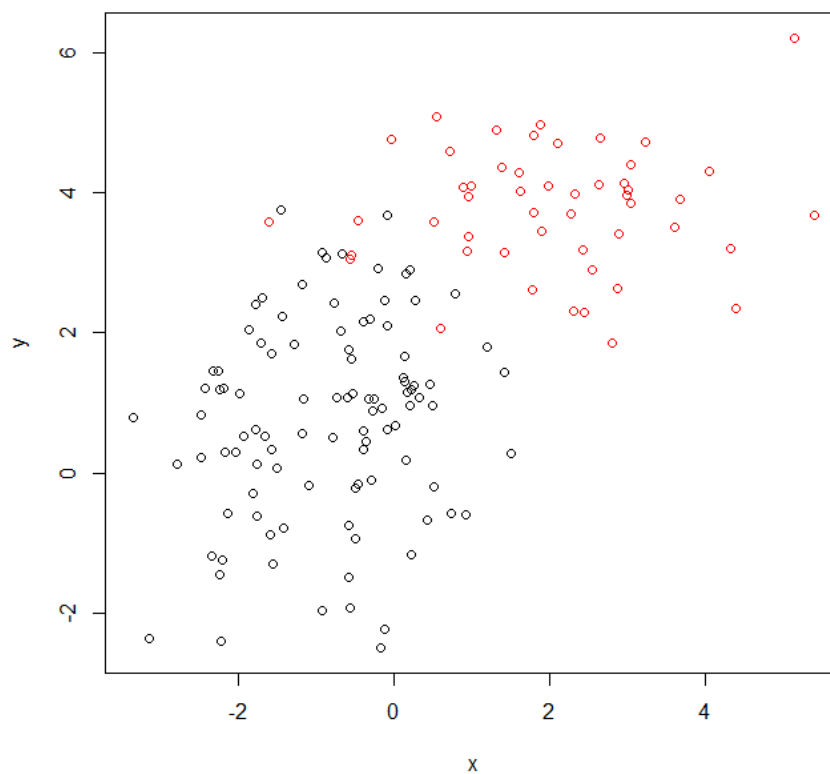
Результаты:

1. Реальные данные:



Значение суммы квадратов расстояний от каждого объекта наблюдения до центра соответствующего кластера - 1067.932.

2. Смоделированные данные:



Значение суммы квадратов расстояний от каждого объекта наблюдения до центра соответствующего кластера - 1028.226.

Листинг программы:

```
require(MASS)
analyse_clust <- function(x, y, clazz) {
  k <- length(unique(clazz))
  clust <- kmeans(cbind(x, y), k)
  print(clust$totss)
  dev.new()
  plot(x, y, col=as.factor(clazz))
  dev.new()
  plot(x, y, col=as.factor(clust$cluster))
  points(clust$centers, col=1:length(clust$centers), pch=4, cex=2)
}

dat <- read.table("wine.csv", sep=",")
analyse_clust(dat$V2, dat$V11, as.factor(dat$V1))

n1 <- 100
a1 <- c(-1, 1)
r1 <- cbind(c(1, 0.1), c(0.1, 2))
n2 <- 50
a2 <- c(2, 4)
r2 <- cbind(c(2, 0.1), c(0.1, 1))
dat <- rbind(mvrnorm(n1, a1, r1), mvrnorm(n2, a2, r2))
analyse_clust(dat[,1], dat[,2], c(rep(1, n1), rep(2, n2)))
```