Белорусский Государственный Университет

Информатики и Радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра ЭВМ

Лабораторная работа №3

Тема «Кластерный анализ»

Выполнил: Проверил:

Студент группы 7М2431 Марченко В.В.

Зайцев Ю.В.

Минск, 2017

Задание:

Входные данные: n объектов, каждый из которых характеризуется

двумя числовыми признаками: и , а также номером класса

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 квадратом расстояний от каждого объекта наблюдения до центра соответствующего кластера.

Исходные данные:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| n1 | a1 | R1 | n2 | a1 | R2 |
| 100 |  |  | 50 |  |  |

Где независимые случайные векторы (X, Y), n1 из которых относятся к первому классу, а n2 – ко второму классу. Векторы, относящиеся к первому классу, распределены по гауссовскому закону с математическим ожиданием а1 и корреляционной матрицей R1, а векторы, относящиеся ко второму классу – по гауссовскому закону с математическим ожиданием а2 и корреляционной матрицей. R2.

Название файла: wine.csv

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

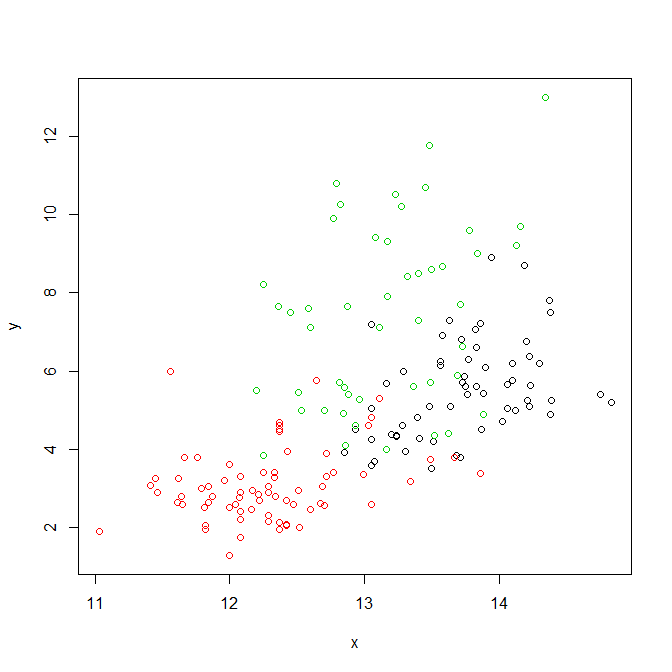
Первый признак: alcohol (столбец № 2)

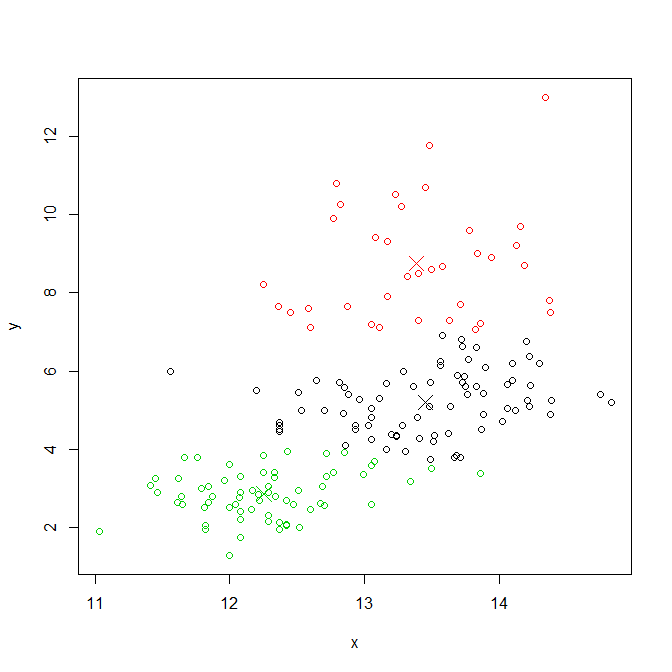
Второй признак: color-intensity (столбец № 11)

Класс cultivar (столбец №1)

Результаты:

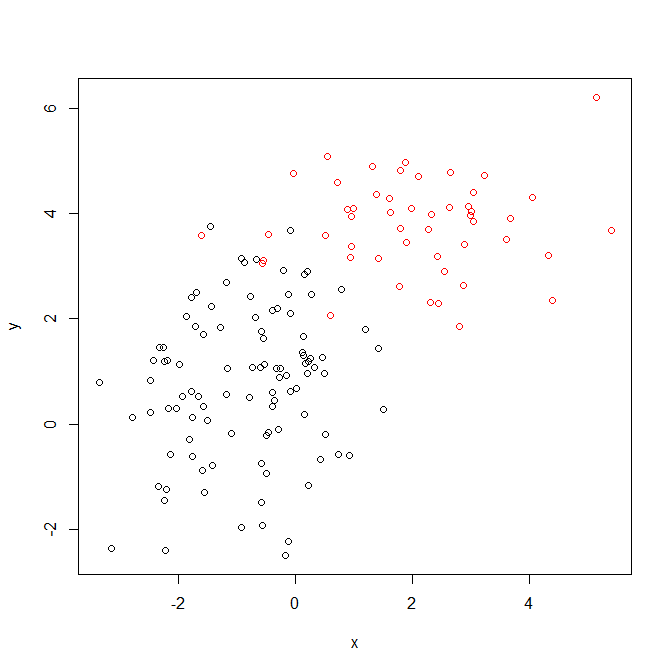
1. Реальные данные:

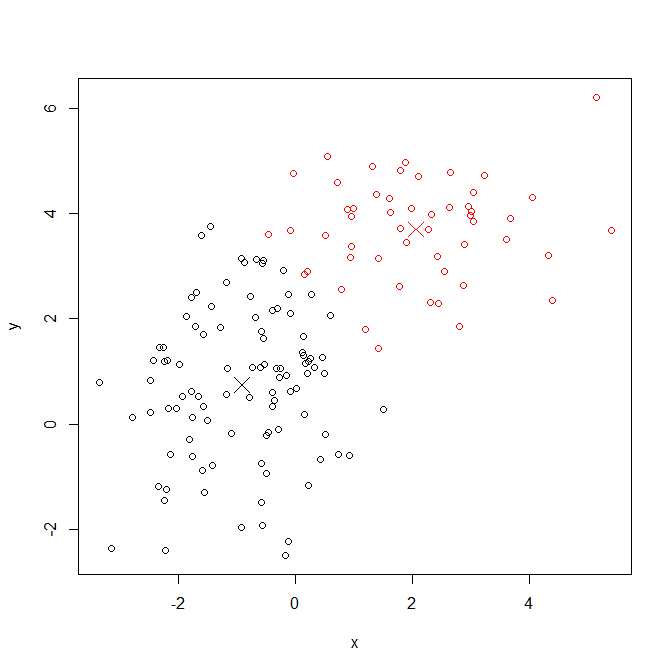




Значение суммы квардартов расстояний от каждого объекта наблюдения до центра соответствующего кластера - 1067.932.

1. Смоделированные данные:





Значение суммы квардартов 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)))