Белорусский Государственный Университет
Информатики и Радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра ЭВМ

Лабораторная работа №3

Тема «Кластерный анализ»

Выполнил:

Студент группы 7М2432

Пашковский А.А.

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 квадратом расстояний от каждого объекта наблюдения до центра соответствующего кластера.

Данные для моделирования представлены в таблице 1, 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 a1 и корреляционной матрицей R1, а векторы, относящиеся ко второму классу – по гауссовскому закону с математическим ожиданием a2 и корреляционной матрицей. R2.

Таблица 1 - Исходные данные:

Вариант	n_1	a_1	R_1	n_2	a_1	R_2
3	1000	$\begin{pmatrix}-1\\0\end{pmatrix}$	$\begin{pmatrix}1 & 1\\1 & 2\end{pmatrix}$	2000	$\begin{pmatrix}-4\\3\end{pmatrix}$	$\begin{pmatrix}1 & -1\\-1 & 2\end{pmatrix}$

Реальные статистические данные из заданного набора (выдаются преподавателем).

26. Parkinsons Disease Data Set

Название файла: `26-parkinsons.txt`

Ссылка: http://archive.ics.uci.edu/ml/datasets/Parkinsons

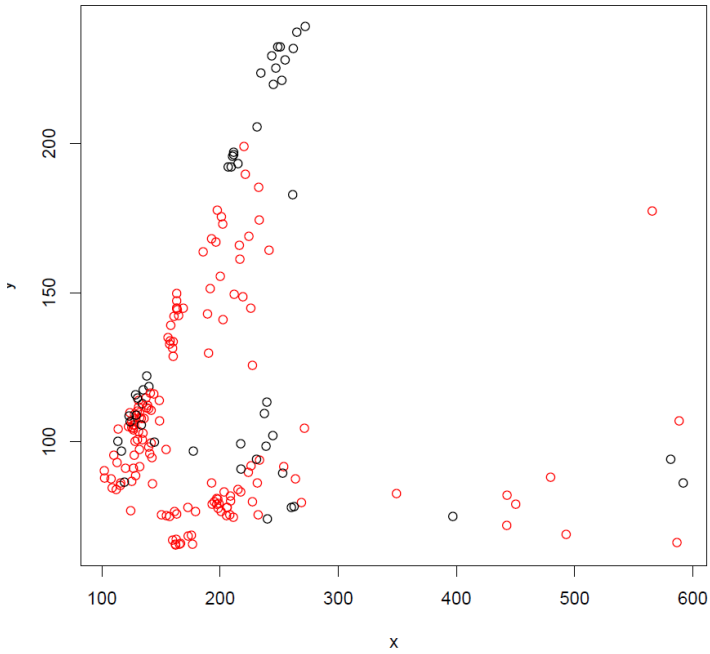
Первый признак: MDVP:Fhi(Hz) (столбец № 3)

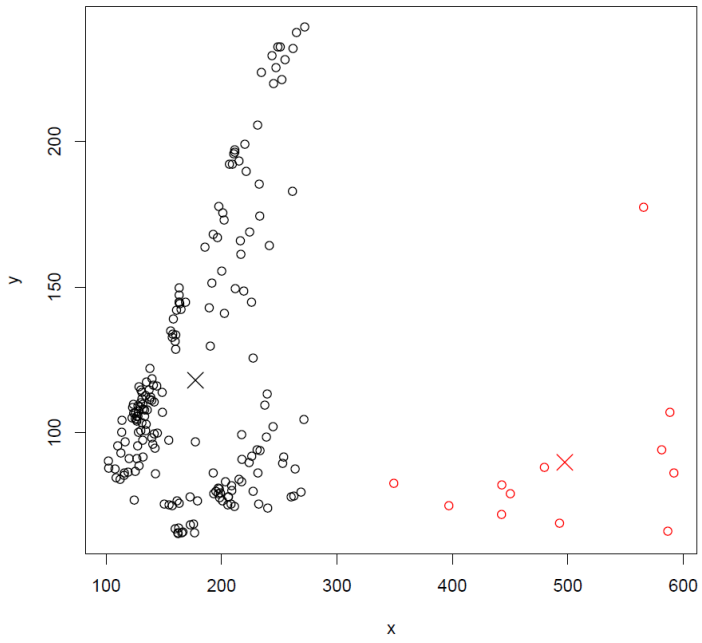
Второй признак: MDVP:Flo(Hz) (столбец № 4)

Класс: status (столбец № 18)

Результаты:

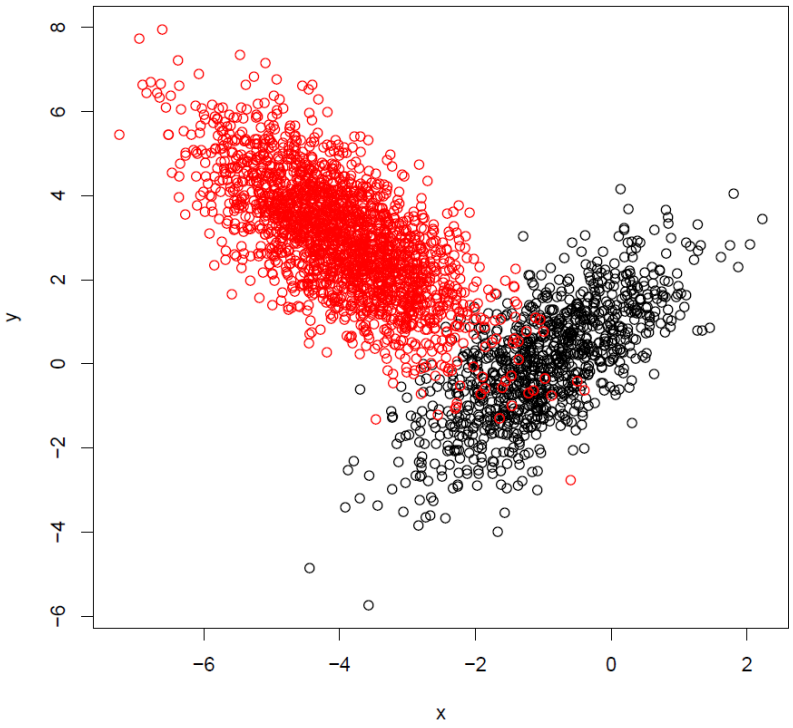
1. Реальные данные:

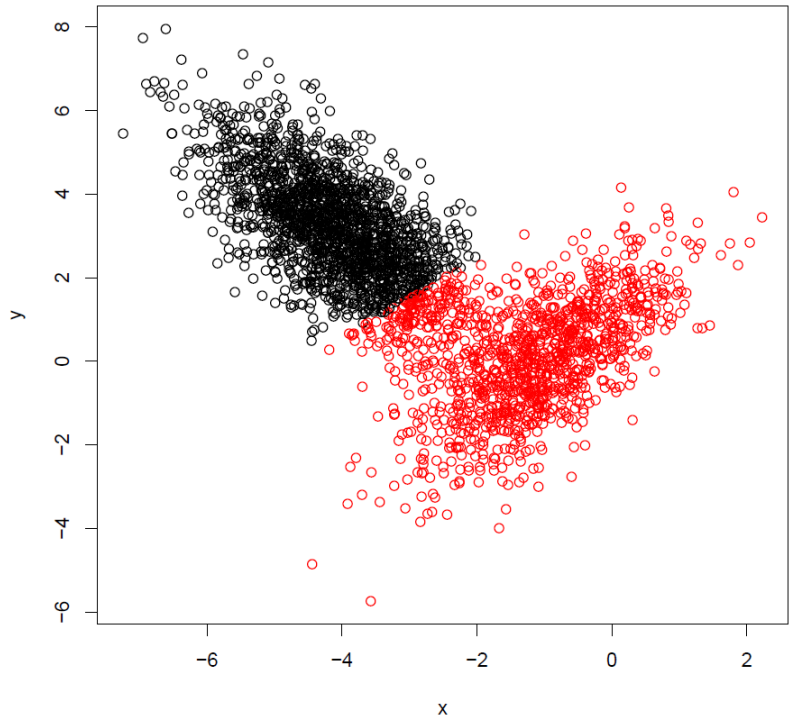




Значение суммы квардартов расстояний от каждого объекта наблюдения до центра соответствующего кластера 1991374

2. Смоделированные данные:





Значение суммы квардартов расстояний от каждого объекта наблюдения до центра соответствующего кластера - 20707.21

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("parkinsons.data.txt", sep=",")
analyse_clust(dat$V3, dat$V4, as.factor(dat$V18))

n1 <- 1000
a1 <- c(-1, 0)
r1 <- cbind(c(1, 1), c(1, 2))
n2 <- 2000
a2 <- c(-4, 3)
r2 <- cbind(c(1, -1), c(-1, 2))
dat <- rbind(mvrnorm(n1, a1, r1), mvrnorm(n2, a2, r2))
analyse_clust(dat[,1], dat[,2], c(rep(1, n1), rep(2, n2)))
```