Белорусский Государственный Университет
Информатики и Радиоэлектроники

Факультет компьютерных систем и сетей

Кафедра ЭВМ

Лабораторная работа №4

Тема «Классификация»

Выполнил:

Студент группы 7М2432

Пашковский А.А.

Проверил:

Марченко В.В.

Минск, 2017

Задание:

Требуется исследовать работу алгоритма классификации объектов по ближайшему соседу. Для каждого набора данных требуется выполнить следующие задания:

1. Случайным образом разделить имеющуюся выборку примерно пополам на обучающую выборку и контрольную выборку.
2. Произвести классификацию объектов контрольной выборки, используя данные о классах объектов из обучающей выборки, с помощью алгоритма классификации по ближайшему соседу.
3. Изобразить объекты графически в трёхмерном пространстве. Для объектов разных классов и разных выборок следует использовать разные обозначения. Отдельно представить графики, на одном из которых объекты из контрольной выборки имеют свои настоящие классы, а на другом – классы, к которым их отнес классификатор.
4. Оценить вероятность ошибочной классификации.

Все описанные задания выполнить для двух наборов данных.

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	$\begin{pmatrix}5\\7\\3\end{pmatrix}$	$\begin{pmatrix}2 & -1 & 0.1\\-1 & 4 & -1\\0.1 & -1 & 2\end{pmatrix}$	2000	$\begin{pmatrix}9\\11\\7\end{pmatrix}$	$\begin{pmatrix}4 & 1 & -1\\1 & 2 & 1\\-1 & 1 & 2\end{pmatrix}$

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)

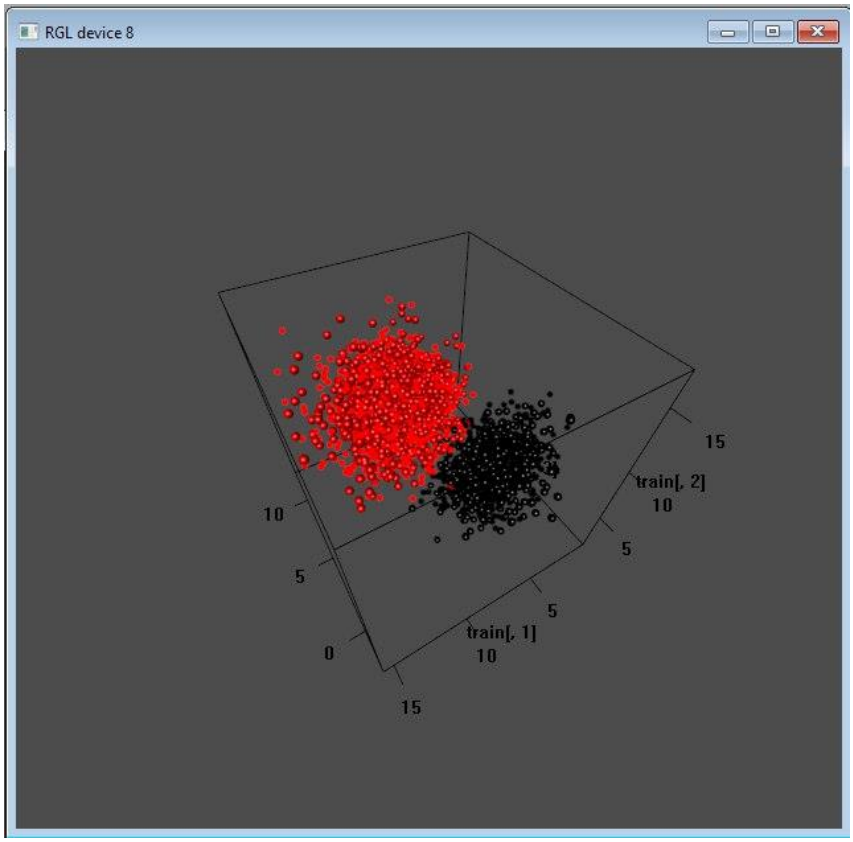
Второй признак: MDVP:Flo(Hz) (столбец № 4)

Третий признак: DFA (столбец № 20)

Класс: status (столбец № 18)

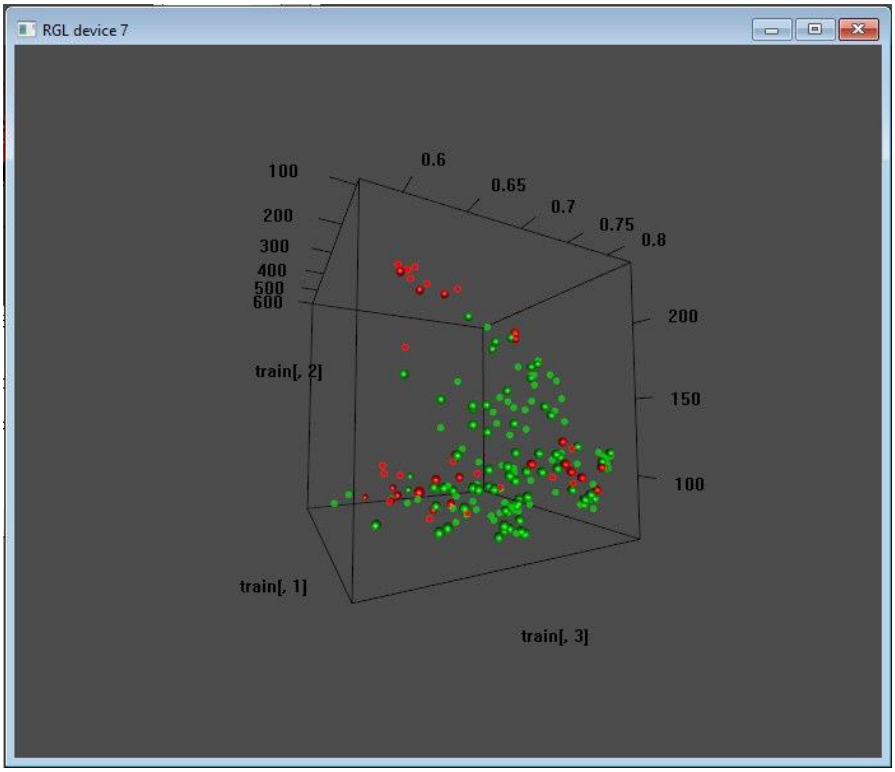
Результаты:

1. Смоделированные данные:



Вероятность ошибочной классификации: 0.004333333

2. Реальные данные:



Вероятность ошибочной классификации: 0.1076923

Листинг программы:

```
install.packages("rgl")

require(class)
require(MASS)
require(rgl)

plot_points <- function(train, test, clazz.train,
                        clazz.test) {
  rgl.open()
  plot3d(train[, 1], train[, 2], train[, 3],
         col=clazz.train, type='p', size=5, add=FALSE)
  plot3d(test[, 1], test[, 2], test[, 3],
         col=clazz.test, type='s', size=1, add=TRUE)
}

analyse_knn <- function(dat, clazz) {
  n <- nrow(dat)
  rnd.num <- sample(1 : n)
  train.num <- rnd.num[1 : (n %/% 2)]
  test.num <- rnd.num[(n %/% 2 + 1) : n]
  train <- dat[train.num,]
  test <- dat[test.num,]
  clazz.train <- clazz[train.num]
  clazz.test <- clazz[test.num]
  clazz.knn <- knn(train, test, clazz.train)
  print(sum(clazz.test != clazz.knn) / n)
  plot_points(train, test, clazz.train, clazz.test)
  plot_points(train, test, clazz.train, clazz.knn)
}

dat <- read.table("parkinsons.data.txt", sep=",")
analyse_knn(cbind(dat$V3, dat$V4, dat$V20), unclass(dat$V18))

n1 <- 1000
a1 <- c(5, 7, 3)
r1 <- cbind(c(2, -1, 0.1), c(-1, 4, -1), c(0.1, -1, 2))
n2 <- 2000
a2 <- c(9, 11, 7)
r2 <- cbind(c(4, 1, -1), c(1, 2, 1), c(-1, 1, 2))

dat <- rbind(mvrnorm(n1, a1, r1), mvrnorm(n2, a2, r2))
analyse_knn(dat, c(rep(1, n1), rep(2, n2)))
```