1. Министерство науки и высшего образования Российской Федерации
2. Санкт-Петербургский Политехнический Университет Петра Великого
3. —

Институт компьютерных наук и технологий

Высшая школа искусственного интеллекта

Отчёт по лабораторной работе «Кластеризация»

по дисциплине «Машинное обучение»

Выполнила

студентка гр. 3540201/20302 Обидина А.И.

Проверил

профессор Уткин Л.В.

Санкт-Петербург

2022

**Задание**

1. Разбейте множество объектов из набора данных pluton в пакете «cluster» на 3 кластера методом центров тяжести (kmeans). Сравните качество разбиения в зависимости от максимального числа итераций алгоритма.
2. Сгенерируйте набор данных в двумерном пространстве, состоящий из 3 кластеров, каждый из которых сильно “вытянут” вдоль одной из осей. Исследуйте качество кластеризации методом clara в зависимости от 1) использования стандартизации; 2) типа метрики. Объясните полученные результаты.
3. Постройте дендрограмму для набора данных votes.repub в пакете «cluster» (число голосов, поданных за республиканцев на выборах с 1856 по 1976 год). Строки представляют 50 штатов, а столбцы - годы выборов (31). Проинтерпретируйте полученный результат.
4. Постройте дендрограмму для набора данных animals в пакете «cluster». Данные содержат 6 двоичных признаков для 20 животных. Переменные - [ , 1] war теплокровные; [ , 2] fly летающие; [ , 3] ver позвоночные; [ , 4] end вымирающие; [ , 5] gro живущие в группе; [ , 6] hai имеющие волосяной покров. Проинтерпретируйте полученный результат.
5. Рассмотрите данные из файла seeds\_dataset.txt, который содержит описание зерен трех сортов пшеницы: Kama, Rosa and Canadian. Признаки: 1. область A, 2. периметр P, 3. компактность C = 4\*pi\*A/P^2, 4. длина зерна, 5. ширина зерна, 6. коэффициент ассиметрии, 7. длина колоска.

# Ход работы

## **Пункт 1**

Множество объектов из набора данных pluton (45 объектов) в пакете «cluster» было разбито на 3 кластера методом центров тяжести (kmeans). Качество разбиения в зависимости от максимального числа итераций алгоритма было определено с помощью графиков (приведены на рисунках 1 - 6), где каждая ячейка показывает кластеризацию по двум различным признакам. Видно, что модель сходится к оптимальному разбиению при увеличении максимального числа итераций, но при всё большем увеличении количества итераций сходимость происходит медленнее.

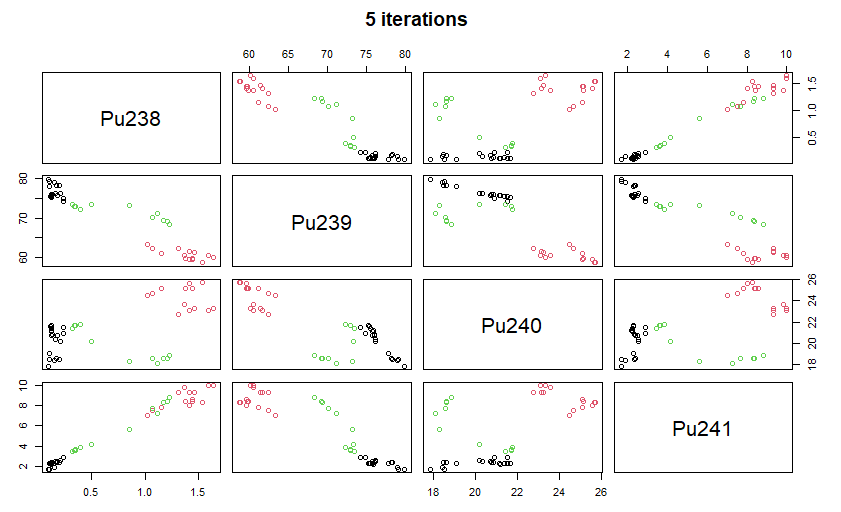


Рисунок 1 – Кластеризация при максимальном количестве итераций = 5

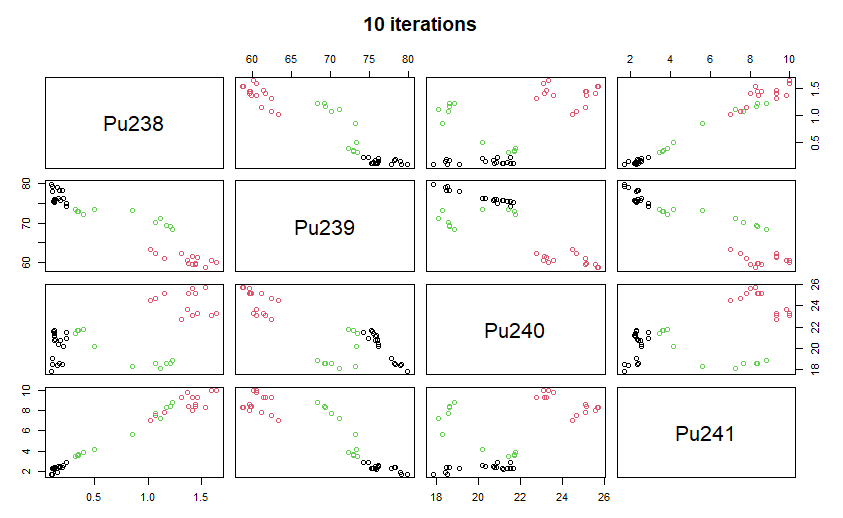


Рисунок 2 – Кластеризация при максимальном количестве итераций = 10

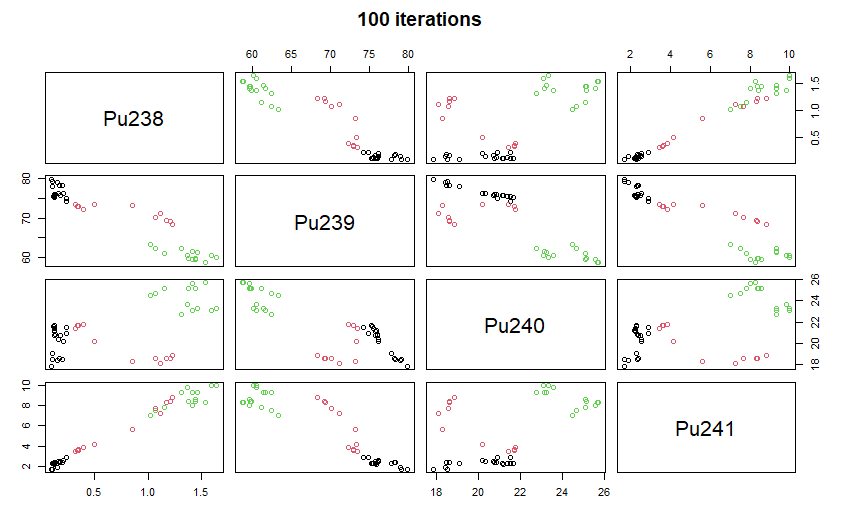


Рисунок 3 – Кластеризация при максимальном количестве итераций = 100

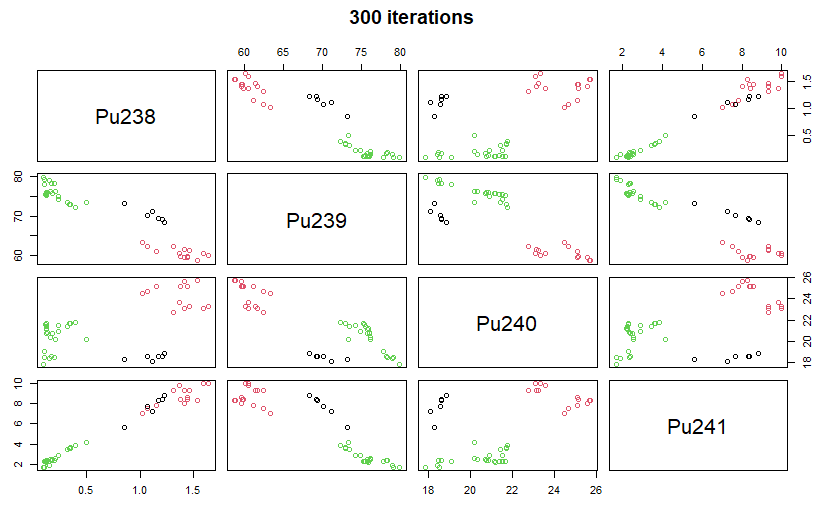


Рисунок 4 – Кластеризация при максимальном количестве итераций = 300

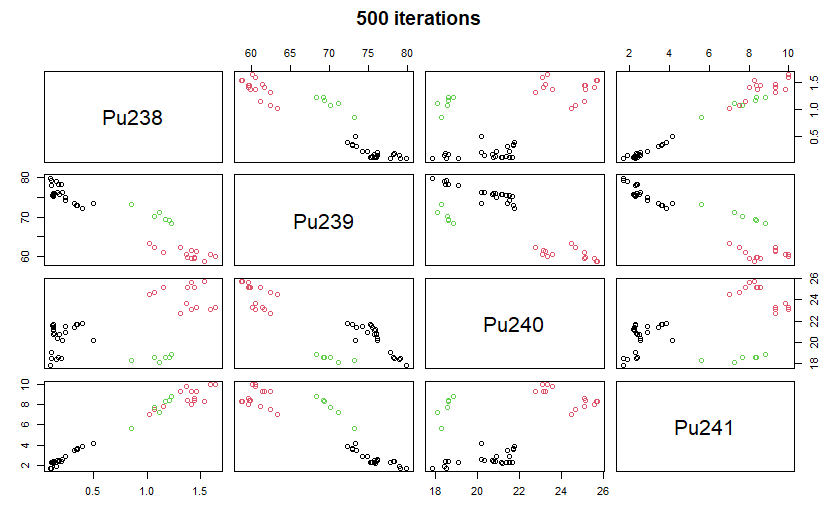


Рисунок 5 – Кластеризация при максимальном количестве итераций = 500

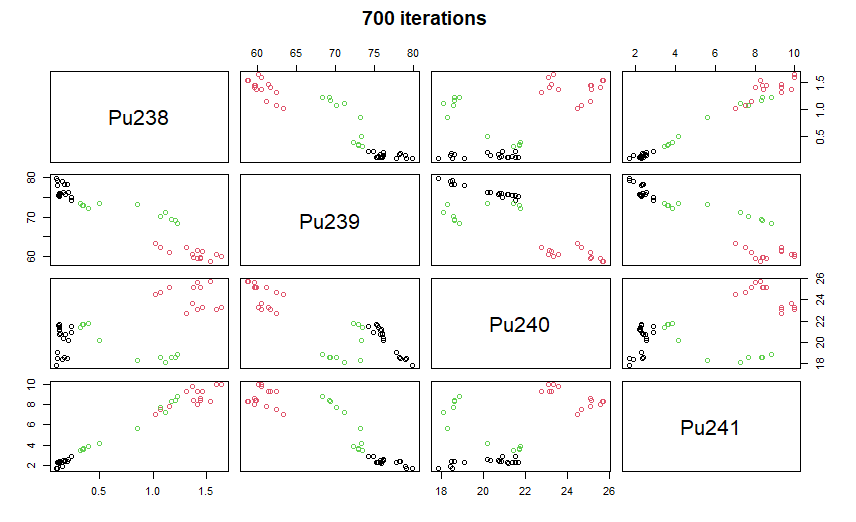


Рисунок 6 – Кластеризация при максимальном количестве итераций = 700

**Код:**

library(cluster)

data(pluton)

clust<-kmeans(pluton,3,iter.max = 5)

plot(pluton,col=clust$cluster,main="5 iterations")

clust<-kmeans(pluton,3,iter.max = 10)

plot(pluton,col=clust$cluster,main="10 iterations")

clust<-kmeans(pluton,3,iter.max = 100)

plot(pluton,col=clust$cluster,main="100 iterations")

clust<-kmeans(pluton,3,iter.max = 300)

plot(pluton,col=clust$cluster,main="300 iterations")

clust<-kmeans(pluton,3,iter.max = 500)

plot(pluton,col=clust$cluster,main="500 iterations")

clust<-kmeans(pluton,3,iter.max = 700)

plot(pluton,col=clust$cluster,main="700 iterations")

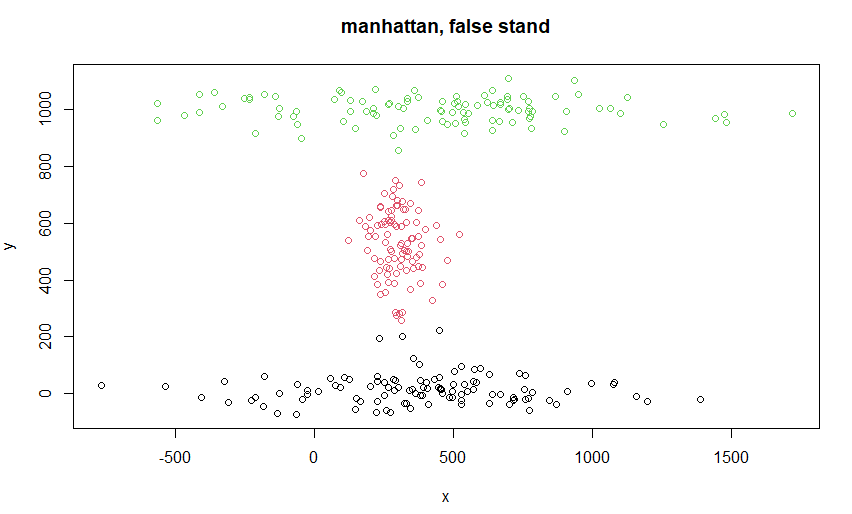
## **Пункт 2**

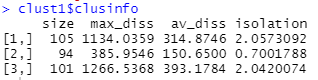
Был сгенерирован набор данных в двумерном пространстве, состоящий из 3 кластеров, каждый из которых сильно “вытянут” вдоль одной из осей (каждый набор данных – нормально распределенные случайные величины). Каждый набор состоит из 100 точек.

Далее было исследовано качество кластеризации методом clara в зависимости от использования стандартизации и типа метрики.

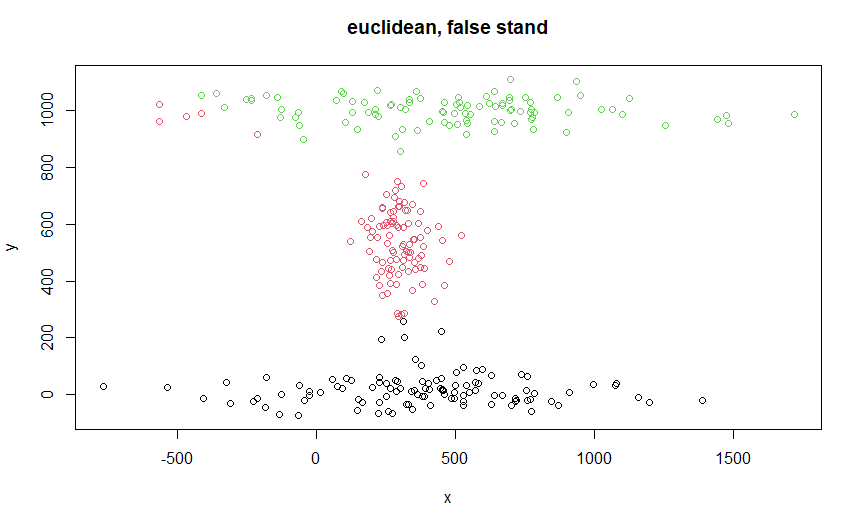
1. Нестандартизованные данные

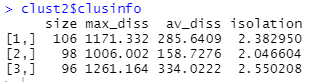
Манхэттенская метрика:





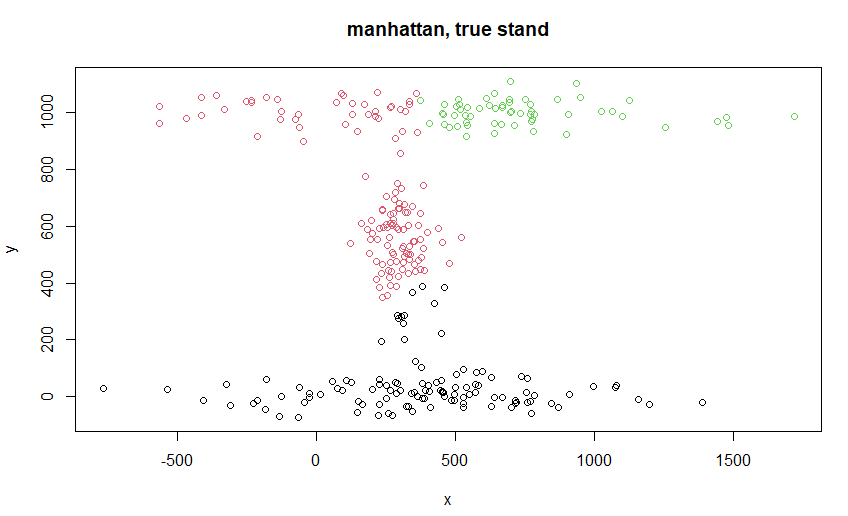
Евклидово расстояние:

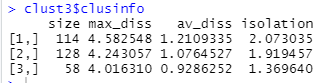




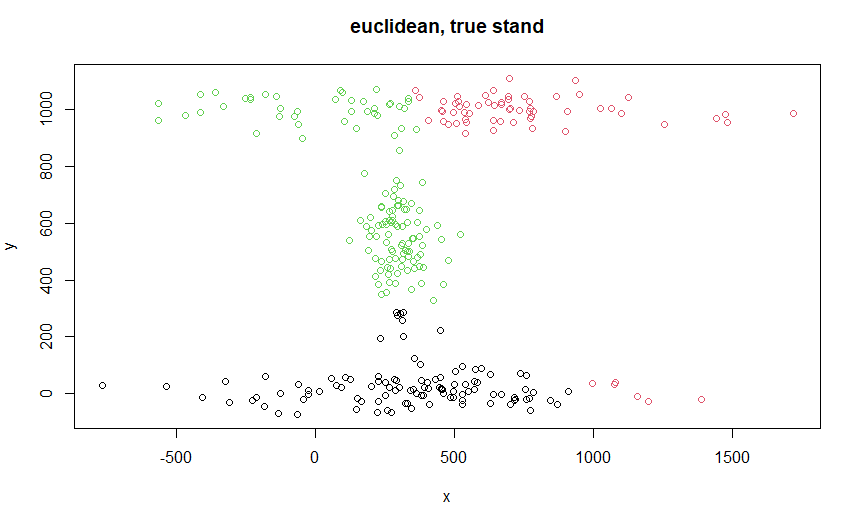
1. Стандартизованные данные

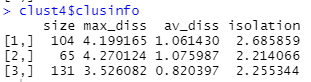
Манхэттенская метрика:





Евклидово расстояние:





**Сравнение нестандартизованных и стандартизованных данных:**

Лучше разбились на кластеры нестандартизованные данные (преимущественно с манхэттенской метрикой, с евклидовым расстоянием данные разбиты не согласно нашему представлению). Последний столбец матрицы clusinfo - максимальное различие между объектами в кластере и медианой кластера, поделенное на минимальное различие между медианой кластера и медианами других кластеров – в этом случае имеет минимальное значение, значит, кластеры хорошо изолированы друг от друга.

**Сравнение манхэттенской метрики и евклидова расстояния:**

Судя по матрице clusinfo, разбиение с использованием манхэттенской метрики точнее.

Проанализировав полученные наблюдения, можно сказать, что результат зависит от данных и представления о том, как они должны быть разбиты.

**Код:**

n = 100

X1 = rnorm(n, mean = 400, sd = 400)

Y1 = rnorm(n, mean = 0, sd = 40)

X2 = rnorm(n, mean = 300, sd = 80)

Y2 = rnorm(n, mean = 500, sd = 130)

X3 = rnorm(n, mean = 550, sd = 450)

Y3 = rnorm(n, mean = 1000, sd = 40)

data <- cbind(as.matrix(c(X1,X2,X3)),as.matrix(c(Y1,Y2,Y3)))

colnames(data) <- c("x", "y")

data

clust1 <- clara(data, 3, stand = FALSE, metric = "manhattan")

plot(data, col = clust1$clustering, xlab = "x", ylab = "y",main="manhattan, false stand")

clust1$clusinfo

clust2 <- clara(data, 3, stand = FALSE, metric = "euclidean")

plot(data, col = clust2$clustering, xlab = "x", ylab = "y",main="euclidean, false stand")

clust2$clusinfo

clust3 <- clara(data, 3, stand = TRUE, metric = "manhattan")

plot(data, col = clust3$clustering, xlab = "x", ylab = "y",main="manhattan, true stand")

clust3$clusinfo

clust4 <- clara(data, 3, stand = TRUE, metric = "euclidean")

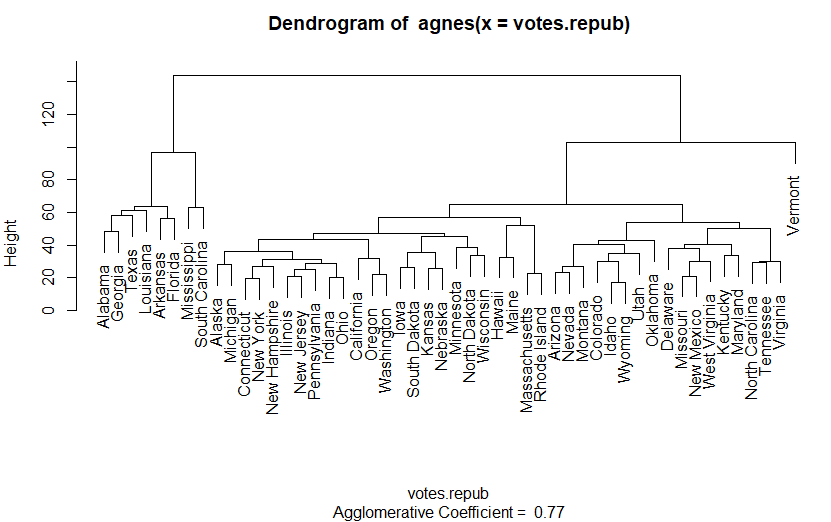
plot(data, col = clust4$clustering, xlab = "x", ylab = "y",main="euclidean, true stand")

clust4$clusinfo

## **Пункт 3**

Была построена дендрограмма для набора данных votes.repub в пакете «cluster» (число голосов, поданных за республиканцев на выборах с 1856 по 1976 год). Строки представляют 50 штатов, а столбцы - годы выборов (31).

Построенная дендрограмма:



Штаты разделены на 2 больших кластера, замечена следующая закономерность: чем меньше высота дендрограммы (измеряется евклидово расстояние) между двумя штатами, тем больше похожи между собой количества голосов, поданных за республиканцев, для этих двух штатов по всем годам. На основании этой закономерности штаты объединяются.

**Код:**

library(cluster)

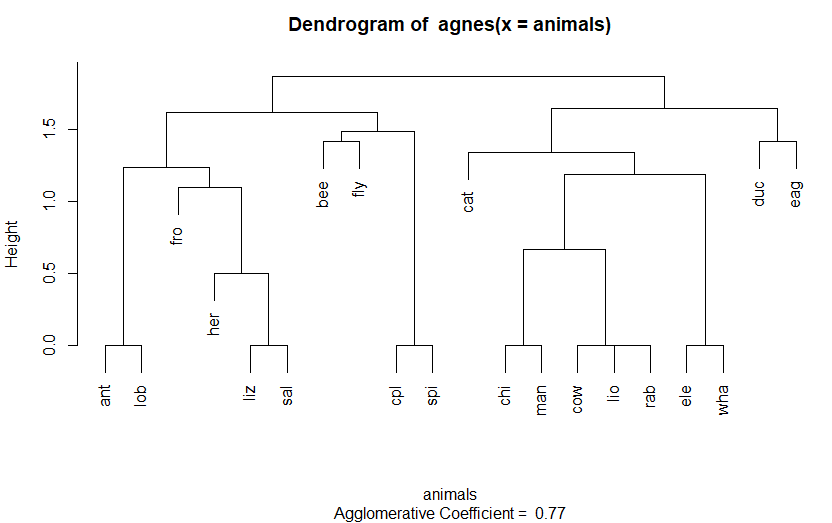
data(votes.repub)

plot(agnes(votes.repub))

## **Пункт 4**

Была построена дендрограмма для набора данных animals в пакете «cluster». Данные содержат 6 двоичных признаков для 20 животных. Переменные - [ , 1] war теплокровные; [ , 2] fly летающие; [ , 3] ver позвоночные; [ , 4] end вымирающие; [ , 5] gro живущие в группе; [ , 6] hai имеющие волосяной покров.

Построенная дендрограмма:



Животные разделены на 2 больших кластера, замечена следующая закономерность: чем меньше высота дендрограммы (измеряется евклидово расстояние) между двумя штатами, тем больше похожи между собой животные по 6 признакам. Например, можно сказать, что животные, объединенные на нижнем уровне дендрограммы, почти идентичны (значения признаков не различаются).

**Код:**

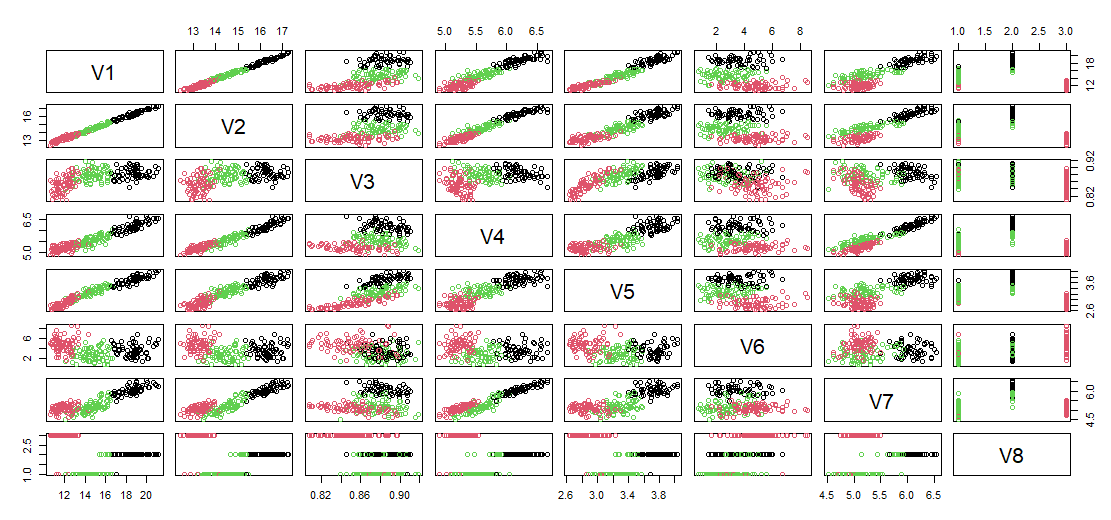
library(cluster)

data(animals)

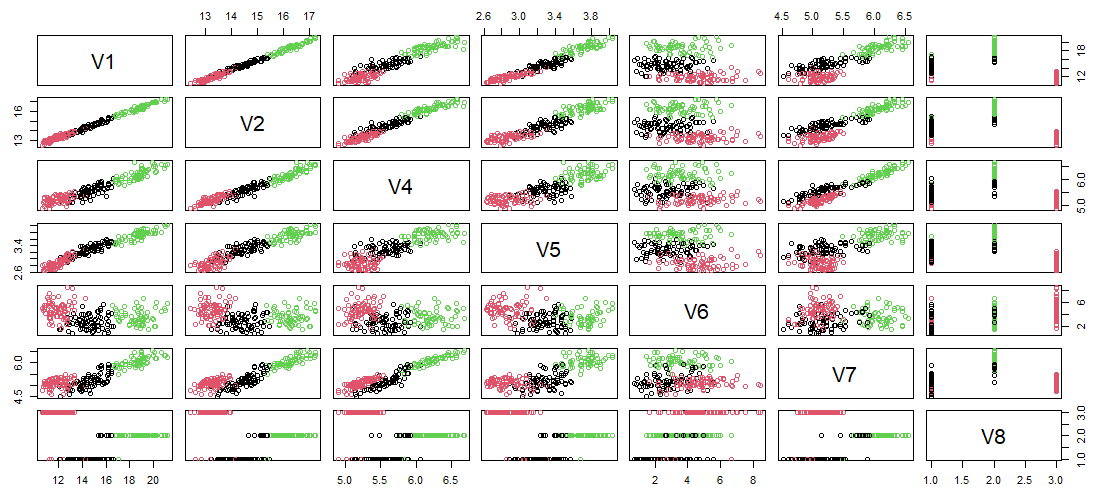
plot(agnes(animals))

## **Пункт 5**

Были рассмотрены данные из файла seeds\_dataset.txt, который содержит описание зерен трех сортов пшеницы: Kama, Rosa and Canadian. Признаки: 1. область A, 2. периметр P, 3. компактность C = 4\*pi\*A/P^2, 4. длина зерна, 5. ширина зерна, 6. коэффициент асимметрии, 7. длина колоска.



Признак 3 коррелирует с признаками 1 и 2, попробуем убрать его и разделить данные на 3 класса с помощью kmeans:



По полученным графикам видно, что разбиение на кластеры достаточно чёткое.

**Код:**

library(cluster)

data <- read.table("D:\\seeds\_dataset.txt")

clust <- kmeans(data, 3)

plot(data, col = clust$cluster)

points(clust$centers, col = 1:2, pch = 8, cex=2)

data

data <- data[,-3]

clust <- kmeans(data, 3)

plot(data, col = clust$cluster)

points(clust$centers, col = 1:2, pch = 8, cex=2)

## **Вывод**

При выполнении данной лабораторной работы был изучен метод кластеризации данных. Этот метод объединяет данные в кластеры, но количество кластеров должно быть известно заранее.

Дендограммы позволяют визуально  
оценить близость объектов вне зависимости от размерности признаков.