ch4-2

Harry Woo

2020 6 16

rm(list = ls())  
library(dplyr)  
library(pls)  
library(ggplot2)  
library(GGally)  
library(gt)

data(iris)  
str(iris)

## 'data.frame': 150 obs. of 5 variables:  
## $ Sepal.Length: num 5.1 4.9 4.7 4.6 5 5.4 4.6 5 4.4 4.9 ...  
## $ Sepal.Width : num 3.5 3 3.2 3.1 3.6 3.9 3.4 3.4 2.9 3.1 ...  
## $ Petal.Length: num 1.4 1.4 1.3 1.5 1.4 1.7 1.4 1.5 1.4 1.5 ...  
## $ Petal.Width : num 0.2 0.2 0.2 0.2 0.2 0.4 0.3 0.2 0.2 0.1 ...  
## $ Species : Factor w/ 3 levels "setosa","versicolor",..: 1 1 1 1 1 1 1 1 1 1 ...

head(iris) %>% gt()

Sepal.Length

Sepal.Width

Petal.Length

Petal.Width

Species

5.1

3.5

1.4

0.2

setosa

4.9

3.0

1.4

0.2

setosa

4.7

3.2

1.3

0.2

setosa

4.6

3.1

1.5

0.2

setosa

5.0

3.6

1.4

0.2

setosa

5.4

3.9

1.7

0.4

setosa

# 각 변수의 표준화한 값

z\_iris <- cbind(stdize(as.matrix(iris[-5])), iris[5])  
head(z\_iris) %>% gt()

Sepal.Length

Sepal.Width

Petal.Length

Petal.Width

Species

-0.8976739

1.01560199

-1.335752

-1.311052

setosa

-1.1392005

-0.13153881

-1.335752

-1.311052

setosa

-1.3807271

0.32731751

-1.392399

-1.311052

setosa

-1.5014904

0.09788935

-1.279104

-1.311052

setosa

-1.0184372

1.24503015

-1.335752

-1.311052

setosa

-0.5353840

1.93331463

-1.165809

-1.048667

setosa

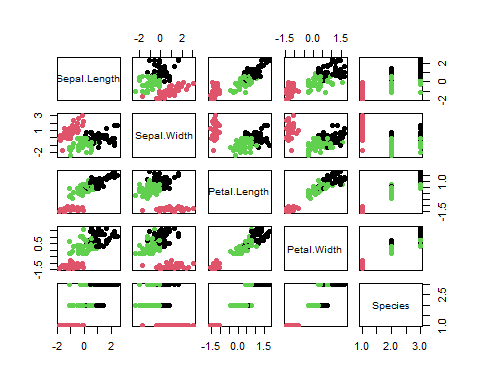
# K-평균 군집분석방법을 이용하여 3개 군집에 대한 군집분석 실시

set.seed(2992)  
km\_cluster <- kmeans(z\_iris[-5], 3)  
km\_cluster

## K-means clustering with 3 clusters of sizes 47, 50, 53  
##   
## Cluster means:  
## Sepal.Length Sepal.Width Petal.Length Petal.Width  
## 1 1.13217737 0.08812645 0.9928284 1.0141287  
## 2 -1.01119138 0.85041372 -1.3006301 -1.2507035  
## 3 -0.05005221 -0.88042696 0.3465767 0.2805873  
##   
## Clustering vector:  
## [1] 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2 2  
## [38] 2 2 2 2 2 2 2 2 2 2 2 2 2 1 1 1 3 3 3 1 3 3 3 3 3 3 3 3 1 3 3 3 3 1 3 3 3  
## [75] 3 1 1 1 3 3 3 3 3 3 3 1 1 3 3 3 3 3 3 3 3 3 3 3 3 3 1 3 1 1 1 1 3 1 1 1 1  
## [112] 1 1 3 3 1 1 1 1 3 1 3 1 3 1 1 3 1 1 1 1 1 1 3 3 1 1 1 3 1 1 1 3 1 1 1 3 1  
## [149] 1 3  
##   
## Within cluster sum of squares by cluster:  
## [1] 47.45019 47.35062 44.08754  
## (between\_SS / total\_SS = 76.7 %)  
##   
## Available components:  
##   
## [1] "cluster" "centers" "totss" "withinss" "tot.withinss"  
## [6] "betweenss" "size" "iter" "ifault"

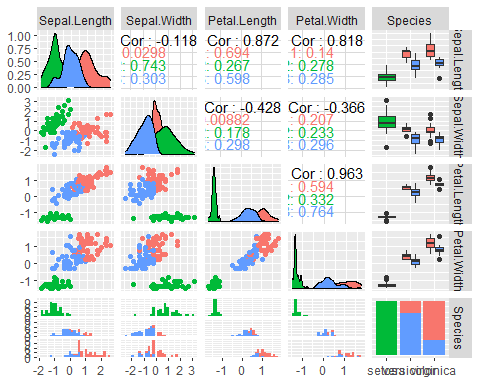
# 군집 결과를 소속군집 산점도로 표현

pairs(z\_iris, col = km\_cluster$cluster, pch = 16)



ggpairs(z\_iris, aes(colour = as.factor(km\_cluster$cluster)))

## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.  
## `stat\_bin()` using `bins = 30`. Pick better value with `binwidth`.



# K-평균 군집분석의 군집결과와 붓꽃 데이터에 주어져 있는 종류와의 분할표 통해 군집분석의 성능 평가

z\_iris2 <- cbind(z\_iris, fit = km\_cluster$cluster)  
confu\_mat <- table(z\_iris2$Species, z\_iris2$fit)  
confu\_mat

##   
## 1 2 3  
## setosa 0 50 0  
## versicolor 11 0 39  
## virginica 36 0 14

# 성능 평가를 위하여 fit 결과를 분류

z\_iris3 <- z\_iris2 %>%   
 mutate(fit = ifelse(fit == 2, "setosa", ifelse(fit == 3, "versicolor", "virginica")))  
confu\_mat2 <- table(z\_iris3$Species, z\_iris3$fit)  
confu\_mat2

##   
## setosa versicolor virginica  
## setosa 50 0 0  
## versicolor 0 39 11  
## virginica 0 14 36

# 오분류율 계산  
error <- 1 - sum(diag(confu\_mat2))/sum(confu\_mat2)  
error

## [1] 0.1666667

1 - error

## [1] 0.8333333

오분류율 16.7%, 정분류율은 83.3%, 특히 Setosa는 완벽히 분류되었음