[K - MEANS 기법을 통한 군집분석]

(iris 데이터셋을 활용하여)

안동대학교 창의융합학부 20191362 양윤태

[1] 군집 분석에 대해 학습한 내용

군집 분석은 비지도 학습, 즉 컴퓨터가 자율적으로 학습하게 하여
변수들을 직접 클러스터링하여 분석하는 기법이다.
변수들을 컴퓨터가 직접 클러스터링하기 때문에 변수 설정을 올바른 방향성을 갖고 해야 하고 변수의 오류를 피하기 위해 변수의 단위 등을 통일시켜 주어야 한다.
군집 간의 데이터 차이는 최대화시켜 차이를 두고 군집 내의 데이터 차이는 최소화하여 차이를 줄인다.

[2] K - MEANS에 대해서

군집 분석 중에서도 분할적 군집에 속하며 프로토타입 기반으로 한 분석 방법 중 하나이다.
평균을 기반으로 하는 군집 분석에서 가장 대표적인 분석 방법이라고 할 수 있다.
K - MEANS에서 K가 의미하는 바는 군집의 수를 나타낸다고 볼 수 있다.
분석을 할 때 컴퓨터에게 설정을 해주는 것인데
만약 K = 2 이면 군집을 두 집단으로 K = 3 이면 군집을 세 집단으로 나눈다고 보면 된다.
군집으로 클러스터링 할 때 유사도라는 개념이 필요한데 K - MEANS에서는 유클리드 거리를 활용하여 유사도를 측정하여 클러스터링에 활용한다.

[3] iris 데이터셋을 이용한 K – MEANS 분석

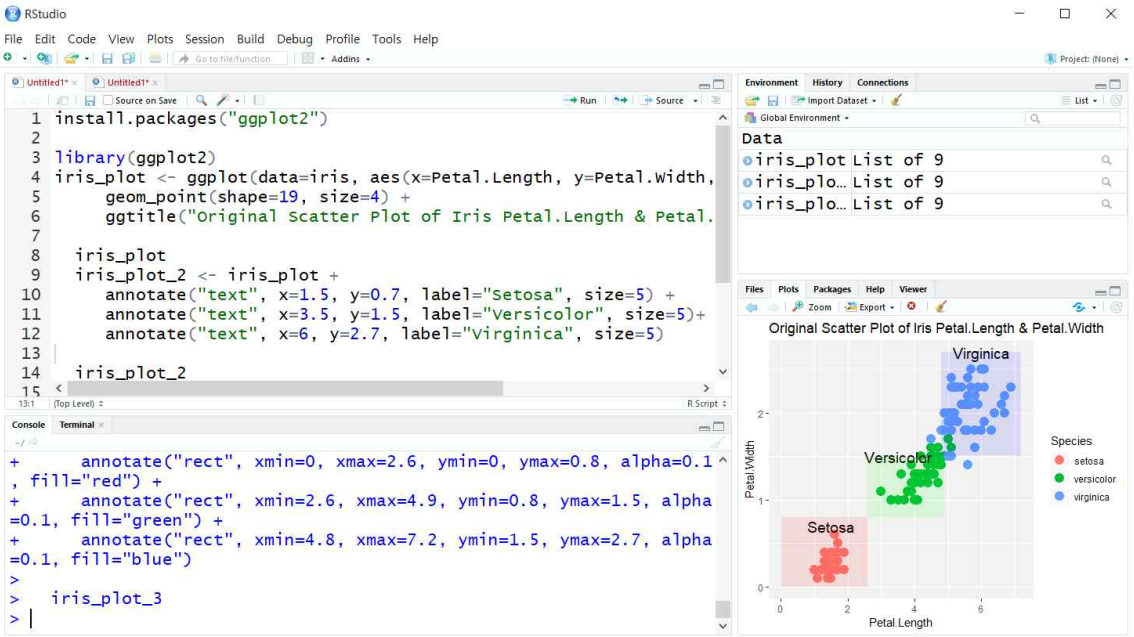
–1) 데이터셋의 구조 확인

```
> 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)
  Sepal.Length Sepal.Width Petal.Length
1          5.1         3.5          1.4
2          4.9         3.0          1.4
3          4.7         3.2          1.3
4          4.6         3.1          1.5
5          5.0         3.6          1.4
6          5.4         3.9          1.7
  Petal.Width Species
1         0.2  setosa
2         0.2  setosa
3         0.2  setosa
4         0.2  setosa
5         0.2  setosa
6         0.4  setosa
> colSums(is.na(iris))
Sepal.Length  Sepal.Width Petal.Length
           0           0            0
 Petal.Width      Species
           0           0
```

먼저 iris 데이터셋을 활용하기 전 str() 함수를 통해서 iris가 가지고 있는 데이터 프레임을 확인한다.

iris 데이터셋은 Sepal.Length, Sepal.Width, Petal.Length, Petal.Width으로 총 4개의 데이터 프레임을 가지고 있고 150개의 관측 결과를 가지고 있음을 확인할 수 있다.

-2) iris 데이터셋을 이용하여 산점도 그리기 - ggplot2 활용



〉install.packages("ggplot2") # ggplot2 패키지를 외부에서 가져온다.
〉library(ggplot2) # ggplot2을 불러온다.

〉iris_plot 〈- ggplot(data=iris, aes(x=Petal.Length, y=Petal.Width, colour=Species))
#x축에 Petal.Length, y축에 Petal.Width를 할당한다. 색상은 종에 따라서 나눈다.
+ geom_point(shape=19, size=4) #모양은 19번째 모양으로, 크기는 4로
+ ggtitle("Original Scatter Plot of Iris Petal.Length & Petal.Width") #제목 설정
〉iris_plot #iris_plot 산점도 확인

〉iris_plot_2 〈- iris_plot #기존 iris_plot 에다가 부가 옵션 더해서 iris_plot_2에 할당
+annotate("text", x=1.5, y=0.7, label="Setosa", size=5) #크기5인 Setosa 텍스트를 추가
+annotate("text", x=3.5, y=1.5, label="Versicolor", size=5) #크기5인 Versicolor 텍스트를 추가
+annotate("text", x=6, y=2.7, label="Virginica", size=5) #크기5인 Virginica 텍스트를 추가
〉iris_plot_2 #iris_plot_2 산점도 확인

〉iris_plot_3 〈- iris_plot_2 #기존 iris_plot_2 에다가 부가 옵션 더해서 iris_plot_3에 할당
+annotate("rect", xmin=0, xmax=2.6, ymin=0, ymax=0.8, alpha=0.1, fill="red")
#빨간색 음영 추가
+annotate("rect", xmin=2.6, xmax=4.9, ymin=0.8, ymax=1.5, alpha=0.1, fill="green")
#초록색 음영 추가
+annotate("rect", xmin=4.8, xmax=7.2, ymin=1.5, ymax=2.7, alpha=0.1, fill="blue")
#파란색 음영 추가
〉iris_plot_3 #iris_plot_3 산점도 확인

=〉 군집이 3개로 나누어 분류된 것을 볼 수 있다.

-3) K-MEANS 분석

```
> iris_k_means <- kmeans(iris[,c("Petal.Length", "Petal.Width")], 3)
> iris_k_means
K-means clustering with 3 clusters of sizes 54, 46, 50

Cluster means:
  Petal.Length Petal.Width
1     4.292593    1.359259
2     5.626087    2.047826
3     1.462000    0.246000

Clustering vector:
  [1] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3
 [32] 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 3 1 1 1 1 1 1 1 1 1 1 1
 [63] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 2 1 1 1 1 1 2 1 1 1 1 1 1 1 1
 [94] 1 1 1 1 1 1 1 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2 1 2 2 2 1
[125] 2 2 1 1 2 2 2 2 2 2 2 2 2 2 1 2 2 2 2 2 2 2 2 2 2 2 2

Within cluster sum of squares by cluster:
[1] 14.22741 15.16348  2.02200
 (between_SS / total_SS =  94.3 %)

Available components:

[1] "cluster"      "centers"      "totss"        "withinss"
[5] "tot.withinss" "betweenss"    "size"         "iter"
[9] "ifault"
>
```

=〉 위와 같이 군집의 개수, 군집 별 크기, 군집 내 구성요소의 개수 등 클러스터링 된 내용들을 확인 가능하고 Petal.Length와 Petal.Width의 평균 좌표를 보여준다.

[4] 느낀 점 및 각오

이번 과제는 평소보다 더 얻은 것이 많았던 과제인 것 같다.
느낀 것이 대략 세 가지가 있는데
첫 번째로는 보고서를 작성하는 경험을 했다는 것이 굉장히 뜻 깊었다. 다른 친구들이 정성스럽게 보고서를 한글 문서로 제출하는 것을 보고 여태 과제를 제출할 때 스마트LMS로만 작성하여 교수님께서 얼마나 읽기 힘드셨을지 죄송스러웠다. 또한 보고서 양식을 직접 설정하고 글의 개요를 생각하면서 앞으로 보고서를 쓸 일이 많을 나에게 큰 경험이 된 것 같아 뜻 깊었다.
두 번째로는 스스로 학습하면서 큰 개념과 부가적인 개념들을 모두 학습할 수 있다는 점이었다. 예를 들어가서 군집 분석을 공부하면서 K-MEANS 분석 코드를 알게 되는 것이 큰 개념이지만 K-MEANS를 분석하기 이전에 산점도로 나타내면서 알게 된 annotate라는 함수를 공부할 수 있었다는 점이다. annotate를 통해서 그래프에 텍스트를 추가하고 음영을 추가하여 그래프의 질을 높일 수 있다는 느낌을 받을 수 있었고 기존 ggplot2를 이용해 그래프를 꾸며 보았던 경험이 있었던 나에게 부가적인 함수들이 늘어나면서 점점 R언어에 대해서 전문가가 되고 있다는 느낌을 받게 되었다.
세 번째로는 누구에게 판별 분석과 군집 분석의 기본적인 개념과 예제를 설명할 수 있게 되었다는 점이다. 팀프로젝트부터 시작해서 수업 복습과 개별과제까지 함께 하니 판별 분석과 군집 분석에 대해서 기본적으로 설명이 가능하다는 점을 느끼면서 “아 정말 공부는 반복이고 또 반복이다. 반복의 연속이구나” 라는 생각이 들게 되었다. 나는 앞으로 직접 분석을 공부하고 예제를 찾으면서 직접 적용하고 보고서를 쓰게 되면서 느꼈던 점들과 R언어에 대한 흥미를 살리면서 R언어에 대해 전문가가 되어야 겠다는 생각이 들었다.