project model 3-Hierarchical clustering

Jingjing Li

2022-12-16

I cannot knit all 3 clustering models at once, so I did them separately.

For Hierarchical clustering, I cannot generate PDF from RMarkdown file. I tried to troubleshoot using this website: https://yihui.org/tinytex/r/#debugging%20for%20debugging%20tip and other websites, such as, stack overflow.

At the end, I found the issue is that there are invalid characters in the code that the LaTex cannot recognize, the character is the function fviz_dend.

Once I deleted the line with fviz_dend, I was able to generate PDF. However I need the function.

At the end, I knit to Word document, then converted the Word file to PDF.

Helper packages

```
library(tidyverse)
```

```
## ── Attaching packages ─────────────────────────────────────── tidyverse 1.
3.2 ──
## ✓ ggplot2 3.4.0      ✓ purrr   0.3.4
## ✓ tibble  3.1.7      ✓ dplyr   1.0.10
## ✓ tidyr   1.2.1      ✓ stringr 1.4.1
## ✓ readr   2.1.2      ✓ forcats 0.5.2
## ── Conflicts ────────────────────────────────────────── tidyverse_conflict
s() ──
## ✕ dplyr::filter() masks stats::filter()
## ✕ dplyr::lag()    masks stats::lag()
```

```
library(dplyr)
library(stringr)
library(gridExtra)
```

```
##
## Attaching package: 'gridExtra'
##
## The following object is masked from 'package:dplyr':
##
##     combine
```

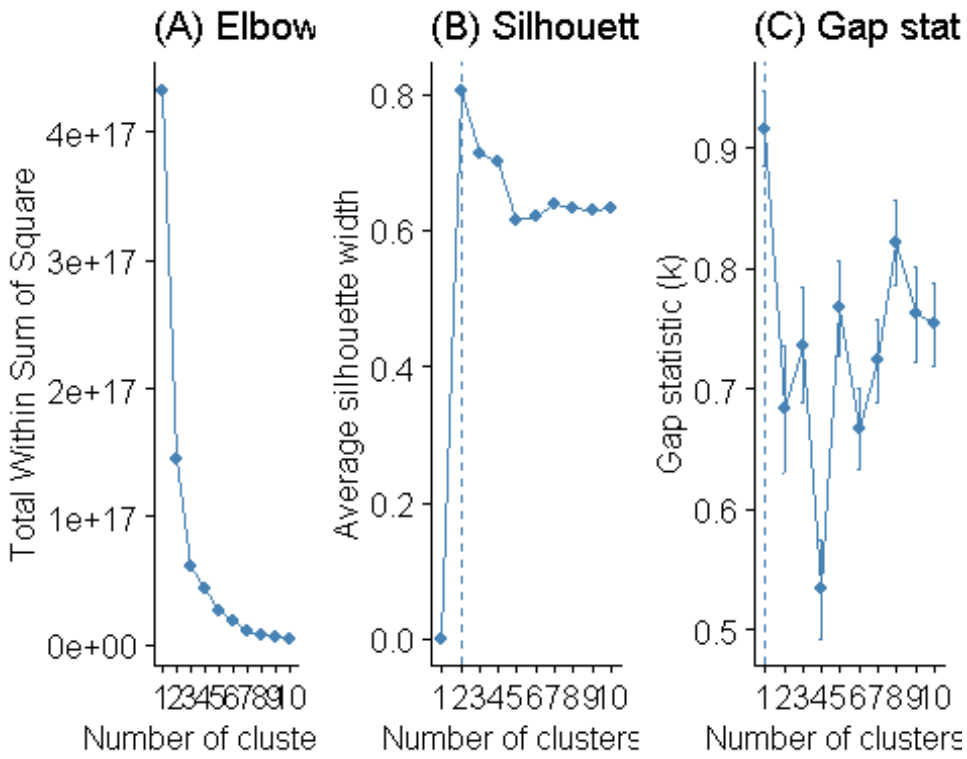
```
library(cluster)
library(factoextra)

## Welcome! Want to learn more? See two factoextra-related books at https://g
oo.gl/ve3WBa

library(mclust)

## Package 'mclust' version 6.0.0
## Type 'citation("mclust")' for citing this R package in publications.
##
## Attaching package: 'mclust'
##
## The following object is masked from 'package:purrr':
##
##     map
```

process the data

```
df <-read.csv("radiomics_completedata.csv")

df <- na.omit(df)

df<-select(df,-c(Institution, Failure.binary))

set.seed(123)
```

Hierarchical clustering

```
# Plot cluster results
p1 <- fviz_nbclust(df, FUN = hcut, method = "wss",
                   k.max = 10) +
  ggtitle("(A) Elbow method")
p2 <- fviz_nbclust(df, FUN = hcut, method = "silhouette",
                   k.max = 10) +
  ggtitle("(B) Silhouette method")
p3 <- fviz_nbclust(df, FUN = hcut, method = "gap_stat",
                   k.max = 10) +
  ggtitle("(C) Gap statistic")

# Display plots side by side
gridExtra::grid.arrange(p1, p2, p3, nrow = 1)
```




```
d<- dist(df, method = "euclidean")
hc5 <- hclust(d, method = "ward.D2" )
dend_plot <- fviz_dend(hc5)#an warning
```

```
## Warning: The `<scale>` argument of `guides()` cannot be `FALSE`. Use "none
" instead as
## of ggplot2 3.3.4.
## ℹ The deprecated feature was likely used in the factoextra package.
##   Please report the issue at <]8;;https://github.com/kassambara/factoextra
/issueshttps://github.com/kassambara/factoextra/issues]8;;>.
```

```
dend_data <- attr(dend_plot, "dendrogram")
dend_cuts <- cut(dend_data, h = 2)

# Ward's method
hc5 <- hclust(d, method = "ward.D2" )

# Cut tree into 4 groups
sub_grp <- cutree(hc5, k = 2)

# Number of members in each cluster
table(sub_grp)
```

```
## sub_grp
##   1   2
## 185  12
```

```r
# Plot full dendogram
fviz_dend(
  hc5,
  k = 2,
  horiz = TRUE,
  rect = TRUE,
  rect_fill = TRUE,
  rect_border = "jco",
  k_colors = "jco",
  cex = 0.1
)
```