BS\_HW3.R

sanghyun

2025-05-28

library(ISLR2)  
data(NCI60)  
dim(NCI60$data)

## [1] 64 6830

table(NCI60$labs)

##   
## BREAST CNS COLON K562A-repro K562B-repro LEUKEMIA   
## 7 5 7 1 1 6   
## MCF7A-repro MCF7D-repro MELANOMA NSCLC OVARIAN PROSTATE   
## 1 1 8 9 6 2   
## RENAL UNKNOWN   
## 9 1

sel <- c("BREAST", "COLON", "MELANOMA", "NSCLC", "RENAL")  
nci.data <- NCI60$data[(NCI60$labs %in% sel), ]  
nci.labs <- NCI60$labs[(NCI60$labs %in% sel)]  
  
dim(nci.data)

## [1] 40 6830

length(nci.labs)

## [1] 40

table(nci.labs)

## nci.labs  
## BREAST COLON MELANOMA NSCLC RENAL   
## 7 7 8 9 9

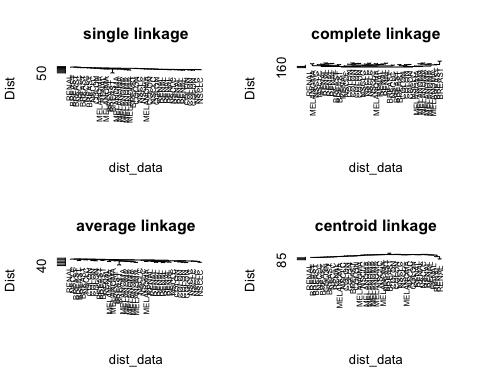
data\_scaled = scale(nci.data)  
  
rownames(data\_scaled) = as.character(nci.labs)  
colnames(data\_scaled) = colnames(nci.data)  
  
data\_scaled[1:5, 1:5]

## 1 2 3 4 5  
## RENAL 0.7727037 -0.2028032 1.7036721 -0.3235800 1.2694549  
## BREAST 1.2519882 -0.3904730 0.9820737 1.2238357 0.2898509  
## BREAST -0.3261439 -0.8808358 -1.1574021 0.4794521 -0.1596145  
## NSCLC 1.1935389 0.1725363 2.8936764 -0.8559271 -0.1826640  
## NSCLC 1.8949309 1.9765873 0.6909025 0.4974977 -1.3812382

# 1 -----------------------------------------------------------------------  
  
ans\_1 = list()  
  
for (subtype in sel){  
   
 sub\_data = data\_scaled[rownames(data\_scaled) == subtype, ]  
   
 dist\_mat = dist(sub\_data, method = "euclidean")  
   
 min\_dist = min(dist\_mat)  
 median\_dist = median(dist\_mat)  
 max\_dist = max(dist\_mat)  
   
 ans\_1[[subtype]] = data.frame(  
 Min\_dist = min\_dist,  
 Median\_dist = median\_dist,  
 Max\_dist = max\_dist  
 )  
}  
  
ans\_1

## $BREAST  
## Min\_dist Median\_dist Max\_dist  
## 1 50.47401 129.3183 149.9546  
##   
## $COLON  
## Min\_dist Median\_dist Max\_dist  
## 1 81.52555 101.2246 110.4879  
##   
## $MELANOMA  
## Min\_dist Median\_dist Max\_dist  
## 1 83.24161 102.0338 121.4439  
##   
## $NSCLC  
## Min\_dist Median\_dist Max\_dist  
## 1 85.85057 110.6411 135.8858  
##   
## $RENAL  
## Min\_dist Median\_dist Max\_dist  
## 1 81.46907 100.3635 140.2839

# 2 -----------------------------------------------------------------------  
  
dist\_data = dist(data\_scaled, method = "euclidean")  
  
linkage\_types = c("single", "complete", "average", "centroid")   
  
ans\_2 = list()  
  
par(mfrow = c(2,2))  
  
for (type in linkage\_types){  
   
 hclust\_result = hclust(dist\_data, method = type)  
   
 plot(hclust\_result,   
 labels = rownames(data\_scaled),   
 main = paste(type, "linkage"), ylab = "Dist", sub = NA, cex = 0.6)  
   
 clusters = cutree(hclust\_result, k = 5)  
   
 comp\_table = table(clusters, nci.labs)  
 ans\_2[[type]] = comp\_table  
}



ans\_2

## $single  
## nci.labs  
## clusters BREAST COLON MELANOMA NSCLC RENAL  
## 1 5 7 8 8 8  
## 2 1 0 0 0 0  
## 3 0 0 0 0 1  
## 4 1 0 0 0 0  
## 5 0 0 0 1 0  
##   
## $complete  
## nci.labs  
## clusters BREAST COLON MELANOMA NSCLC RENAL  
## 1 2 0 1 2 5  
## 2 1 4 1 6 4  
## 3 2 3 0 0 0  
## 4 0 0 0 1 0  
## 5 2 0 6 0 0  
##   
## $average  
## nci.labs  
## clusters BREAST COLON MELANOMA NSCLC RENAL  
## 1 4 7 8 7 8  
## 2 1 0 0 1 0  
## 3 0 0 0 0 1  
## 4 2 0 0 0 0  
## 5 0 0 0 1 0  
##   
## $centroid  
## nci.labs  
## clusters BREAST COLON MELANOMA NSCLC RENAL  
## 1 5 7 8 8 8  
## 2 1 0 0 0 0  
## 3 0 0 0 0 1  
## 4 1 0 0 0 0  
## 5 0 0 0 1 0

# Explain the reason of your choice

4가지 linkage 방법(single, complete, average, centroid) 중에서 complete linkage 방법을 선택하는 것이 가장 타당하다. single, average, centroid 방법 모두 1번 Cluster에 거의 모든 subtype이 몰려 있다. single linkage의 1번 Vluster에는 BREAST 5개, COLON 7개, MELANOMA 8개, NSCLC 8개, RENAL 8개로 모든 subtype이 한 Cluster에 섞여 있는 나쁜 결과를 보인다. 하지만 상대적으로 complete linkage는 비교적 subtype별로 분리되는 경향을 보인다 Cluster 3: MELANOMA 3개 -> MELANOMA 중심 Cluster 5: NSCLC 6개 -> NSCLC 중 Cluster 1: RENAL 5개 -> RENAL 중심 나머지 Cluster들도 특정 subtype 위주로 분포되어 있어 상대적으로 해석이 가능해진다. 정리하면 다른 방법들은 한 Cluster에 모든 subtype이 뒤섞이는 반면, complete linkage는 subtype 간 분리가 상대적으로 잘 되어 있다.

# 3 -----------------------------------------------------------------------  
  
sel = c("BREAST", "COLON", "MELANOMA", "NSCLC", "RENAL")  
  
ans\_3 = matrix(1, nrow=5, ncol=5, dimnames = list(sel, sel))  
  
for (i in 1:5){  
 for (j in 1:5){  
  
 A = data\_scaled[rownames(data\_scaled) == sel[i], ]  
 B = data\_scaled[rownames(data\_scaled) == sel[j], ]  
   
 #' cor-> 열 별  
 #' 우리는 샘플 cor니깐 transpose rr  
 pcc\_mat = cor(t(A), t(B))  
   
 max\_abs\_corr = pcc\_mat[which.max(abs(pcc\_mat))]  
   
 ans\_3[i,j] = max\_abs\_corr  
 }  
}  
  
ans\_3

## BREAST COLON MELANOMA NSCLC RENAL  
## BREAST 1.0000000 -0.3064158 0.3248212 -0.2141222 0.3094733  
## COLON -0.3064158 1.0000000 -0.2449891 0.2353319 -0.2641208  
## MELANOMA 0.3248212 -0.2449891 1.0000000 -0.2265809 0.3825929  
## NSCLC -0.2141222 0.2353319 -0.2265809 1.0000000 0.2097277  
## RENAL 0.3094733 -0.2641208 0.3825929 0.2097277 1.0000000

# 4 -----------------------------------------------------------------------  
  
factor = as.factor(nci.labs)  
pvals = NULL  
  
#anova(lm(data\_scaled[,1] ~ factor))$Pr[1]  
pvals = apply(data\_scaled, 2, function(data\_) anova(lm(data\_~factor))$Pr[1])  
  
top\_20\_gene = as.integer(names(sort(pvals)[1:20]))  
  
top\_20\_data = data\_scaled[ ,top\_20\_gene]  
  
set.seed(1234)  
  
k\_means\_ = kmeans(top\_20\_data, 5, nstart = 20)  
W = k\_means\_$tot.withinss  
  
W\_j = NULL  
  
for (i in 1:ncol(top\_20\_data)){  
   
 W\_j[i] = kmeans(top\_20\_data[, -i], 5, nstart = 20)$tot.withinss  
   
}  
  
ans\_4 = data.frame(  
 gene\_Num = colnames(top\_20\_data),  
 WW\_j = abs(W-W\_j)  
)  
  
ans\_4\_sorted = ans\_4[order(-ans\_4$WW\_j), ]  
  
ans\_4 = list(  
 List\_of\_WW\_j = ans\_4\_sorted,  
 Most\_sign\_gene = ans\_4\_sorted$gene\_Num[1]  
)  
  
ans\_4

## $List\_of\_WW\_j  
## gene\_Num WW\_j  
## 1 282 18.204978  
## 11 5619 16.935619  
## 19 5988 16.868598  
## 13 4100 16.442232  
## 17 5512 16.302889  
## 8 233 15.608414  
## 9 237 14.267705  
## 3 4099 13.682761  
## 4 2024 13.251694  
## 10 2551 13.203650  
## 18 321 12.993595  
## 16 4298 12.313270  
## 2 262 11.554649  
## 20 4332 11.225235  
## 7 5898 10.890954  
## 6 5955 10.261658  
## 15 4295 9.362451  
## 12 5899 9.298253  
## 5 4236 9.294489  
## 14 4327 4.095231  
##   
## $Most\_sign\_gene  
## [1] "282"

# 5 -----------------------------------------------------------------------  
  
set.seed(1234)  
boot <- apply(matrix(1:40, 40, 1000), 2, sample, replace=TRUE)  
  
pve = numeric(1000)  
for (i in 1:1000){  
   
 index = boot[, i]  
   
 X\_b = data\_scaled[index, ]  
   
 C = cor(t(X\_b)) # 샘플 간 correlation (40 × 40)  
   
 eig = eigen(C)  
   
 values = eig$values  
   
 pve[i] = sum(values[1:2]) / sum(values)   
   
}  
  
CI = quantile(pve, c(0.025, 0.975))  
CI

## 2.5% 97.5%   
## 0.2212480 0.3007429

# 6-------------------------------------------------------------------------  
  
pca = prcomp(data\_scaled, scale=TRUE)  
PC\_scores = pca$x[, 1:20] # 40x20  
  
factor\_ = as.factor(nci.labs)  
   
F\_value = apply(PC\_scores, 2, function(pc) anova(lm(pc ~ factor\_))$F[1])  
  
set.seed(1234)  
perm <- apply(matrix(1:40, 40, 1000), 2, sample, replace=FALSE)  
  
F\_perm = matrix(NA, nrow=20, ncol=1000)  
  
  
for (i in 1:1000) {  
   
 factor\_perm = factor\_[perm[, i]]  
   
 F\_perm[, i] = apply(PC\_scores, 2, function(pc) anova(lm(pc ~ factor\_perm))$F[1])  
}  
  
#(sum(F\_perm[1, ] >= F\_value[1]) + 1) / (1000 + 1)  
perm\_pval = (rowSums(F\_perm >= F\_value) + 1) / (1000 + 1)  
  
  
sig\_idx = which(perm\_pval < 0.05)  
ans\_6 = data.frame(  
 PC = paste0("PC", sig\_idx),  
 perm\_pval = signif(perm\_pval[sig\_idx], 3)  
)  
  
ans\_6

## PC perm\_pval  
## 1 PC1 0.000999  
## 2 PC2 0.000999  
## 3 PC3 0.033000  
## 4 PC5 0.002000  
## 5 PC9 0.004000