Lab 10 - Hierarchical clustering

[Sicily Xie #54385315]

2022-12-01

Brief description

In the final lectures of this course, a variety of unsupervised learning techniques were introduced. Particularly, methods for clustering, principal components analysis and high-dimensional scaling were emphasized. The previous lab covered two common clustering algorithms: K-means and model-based clustering using Gaussian mixtures.

In this lab, we will continue the discussion of clustering algorithms by applying hierarchical clustering to a gene expression data set.

Questions

Unsupervised techniques are often used in the analysis of genomic data. In particular, hierarchical clustering is a popular popular tool. We illustrate this technique on the `NCI60` cancer cell line microarray data, which consists of 6,830 gene expression measurements on 64 cancer cell lines.

Each cell line is labeled with a cancer type. We do not make use of the cancer types in performing hierarchical clustering, as this is an unsupervised technique. But after performing hierarchical clustering, we will check to see the extent to which these cancer types agree with the results of this unsupervised technique.

We will perform hierarchical clustering of the observations using complete, average, and single linkage. Euclidean distance is used as the dissimilarity measure. Note that the linkage criterion in hierarchical clustering determines the distance between sets of observations as a function of the pairwise distances between observations. The linkage criteria between two sets of observations A and B used for each method is:

1. Complete-linkage clustering: $\max\{d(a,b) : a \in A, b \in B\}$
2. Average-linkage clustering: $\frac{1}{|A||B|}\sum_{a\in A}\sum_{b\in B} d(a,b)$
3. Single-linkage clustering: $\min\{d(a,b) : a \in A, b \in B\}$

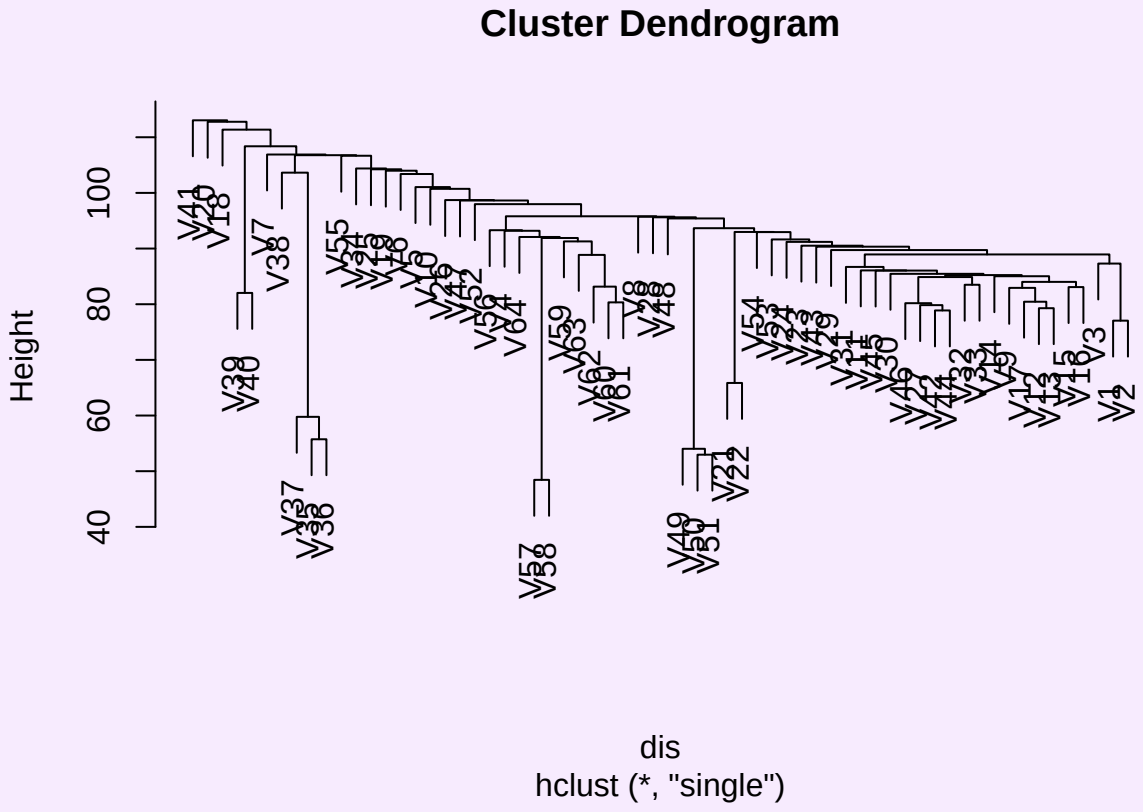
Load the library `ISLR2` in `R`, and then use the following commands to store the gene expressions and the cancer type of each cell:

```
library(ISLR2)
nci_labs <- NCI60$labs
nci_data <- scale(NCI60$data)
```

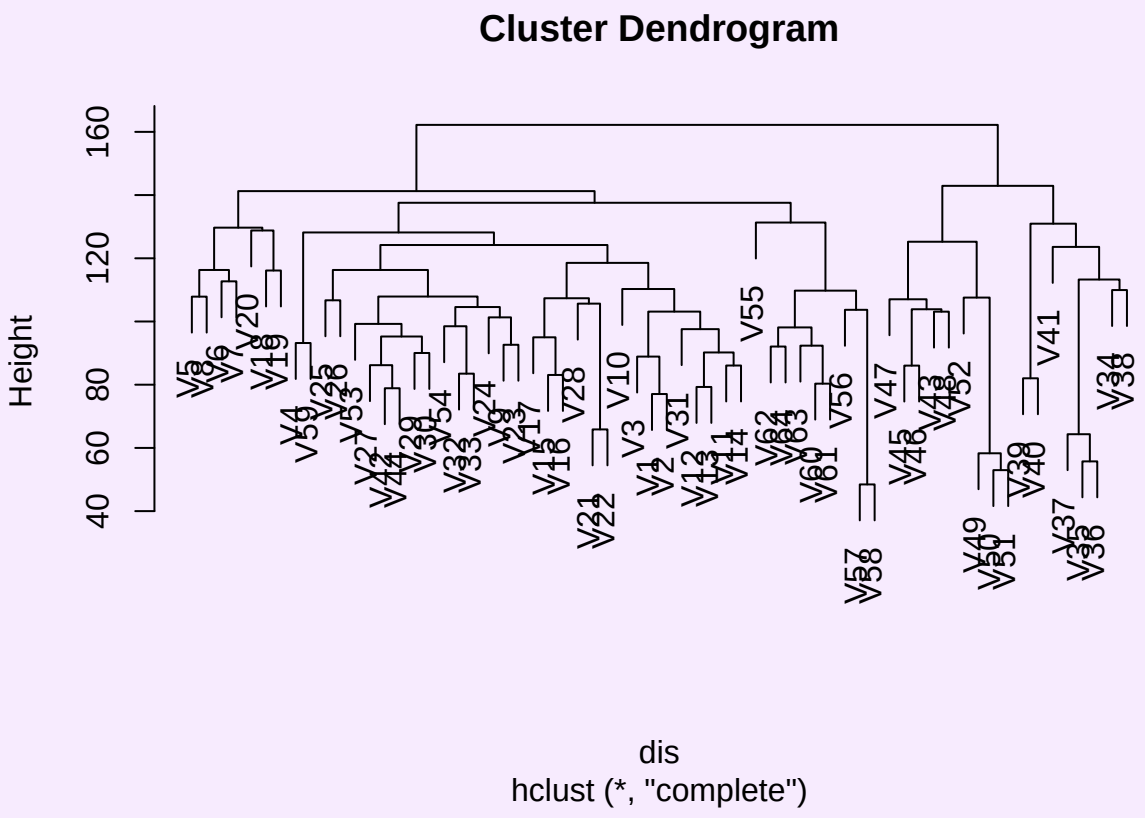
Note that we make sure to standardize the variables to have mean zero and standard deviation one.

Q1 We now proceed to hierarchically cluster the cell lines in the `NCI60` data, with the goal of finding out whether or not the observations cluster into distinct types of cancer. Using the `hclust` function, plot the dendogram complete-linkage, average-linkage and single-linkage. Does the choice of linkage affect the results obtained?

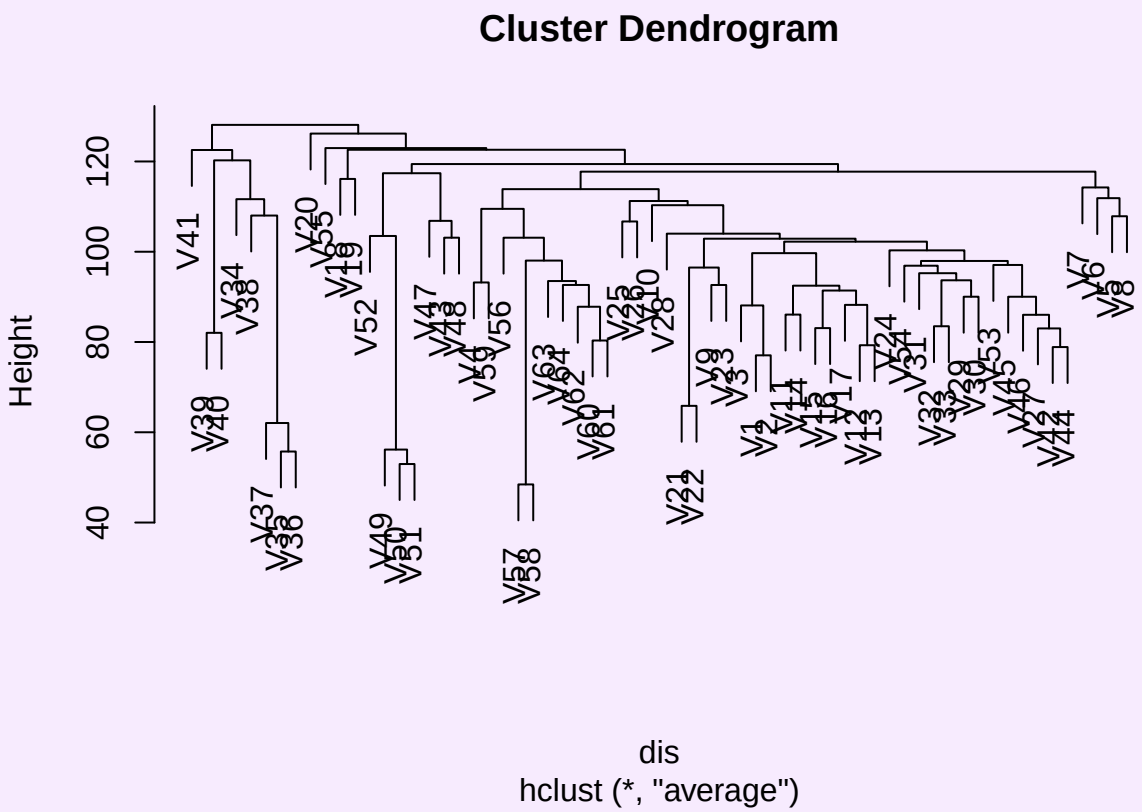
```
# some code
dis <- dist(nci_data)
S.hc <- hclust(dis, method = "single")
C.hc <- hclust(dis, method = "complete")
A.hc <- hclust(dis, method = "average")
plot(S.hc)
```




```
plot(C.hc)
```


Cluster Dendrogram
Height
160
120
80
60
40
dis
hclust (*, "complete")

```
plot(A.hc)
```



Yes, the choice of linkage affect the results obtained.

Q2 Which linkage method(s) tends to yield less balanced dendogram. By less balanced, we mean a dendogram in which cancers of the same type are not grouped as often together.

This Single-Linkage method tends to yield less balanced dendogram.

Q3 We can cut the dendrogram at the height that will yield a particular number of clusters. Using the `cutree()` function, cut the number of clusters to four. For the *LEUKEMIA* cancer type, which linkage method(s) cluster all 6 observations into a single group? (Include your code but hide the output.)

```
# code to make the tables goes here
cutS <- cutree(S.hc, k =4)
cutC <- cutree(C.hc, k =4)
cutA <- cutree(A.hc, k =4)
```

For the *LEUKEMIA* cancer type, both the Complete and Average Linkage method cluster all 6 observations into a single group.

Q4 We will investigate whether K-means clustering and hierarchical clustering with the dendrogram cut to 4 clusters yield different results. Run the K-means algorithm with `nstart=50` and by running `set.seed(123)`

right before `kmeans`. Use the `table` function to compare the cluster output for each method. Only consider hierarchical clustering with complete-linkage for this question. (Be careful again with the label switching. Matched clusters basically mean that, in the table, the rest of a row and column will be 0.)

```
set.seed(123)
k.4 <- kmeans(nci_data, centers = 4, nstart = 50)
cutC <- cutree(C.hc, k =4)
table(cutC,k.4$cluster)
```

```
##
## cutC  1  2  3  4
##    1  0  9 20 11
##    2  0  0  7  0
##    3  8  0  0  0
##    4  0  0  0  9
```

Q5 We will investigate whether model-based clustering with Gaussian mixtures and hierarchical clustering with the dendrogram cut to 4 clusters yield different results. Run `set.seed(123)` right before `Mclust` with 4 groups. Use the `table` function to compare the cluster output for each method. Only consider hierarchical clustering with complete-linkage for this question.

```
library(mclust)
set.seed(123)
k.4 <- kmeans(nci_data, centers = 4, nstart = 50)
mc <- Mclust(nci_data, G = 4)
table(k.4$cluster,mc$classification)
```

```
##
##      1  2  3  4
##   1  0  0  8  0
##   2  0  0  0  9
##   3 27  0  0  0
##   4  7 13  0  0
```