Final Take-Home

Carson James

4/28/2022

Due Monday May 9 at 11:59 pm central time

(5 problems). Download the R dataset "ThreeCancers.RData" from Canvas and load it into your R workspace. Any loading method can be used.

Two objects are contained in this dataset.

`GeneExp`: A matrix of gene expression levels. Each row represents a cancer patient.

`CancerType`: A `factor` (i.e. categorical) vector that gives the cancer type of each patient.

The order of patients in `GeneExp` is the same as that in `CancerType`.

We have 3 types of cancer in our dataset: `LUAD` (lung adenocarcinoma), `KIRC` (kidney renal clear cell carcinoma), `BRCA` (breast invasive carcinoma).

Instructions.

1. When reporting numerical values, keep at least two decimal digits. For example, if the true answer is `1.3267123`, then `1.32`, `1.33`, and `1.3267` are all treated as correct but `1` is not.

If you use code to obtain an answer, include that code. If you do not use code to obtain an answer, include an explanation of how you obtained that answer.

Problem 1.

(i) Find the number of subjects and the number of genes for the matrix `GeneExp`.

(ii) Find the number of patients with LUAD cancer using the vector `CancerType`.

(iii) The expression levels of `Gene2` for all subjects can be accessed using `GeneExp[,2]` or `GeneExp$Gene2`. Find its mean expression level.

(iv) Use the `ggplot2` package to make a scatterplot by plotting the expression level of `Gene68` (vertical axis) against that of `Gene45` (horizontal axis).

(v) Compute the correlation coefficient between `Gene45` and `Gene68`.

Problem 2.

(i) Make a box plot of `Gene15` vs `CancerType` and pass it the following arguemnts: `col = c("red", "green", "blue")` and `ylab = "Gene15"`.

(ii) What does the red box represent?

(iii) Which cancer type(s) have outliers?

(iv) Which cancer type has the smallest median expression level?

Problem 3. Run the following code to create a binary response variable y such that $y_i = 1$ if the i-th subject has KIRC cancer and $y_i = 0$ otherwise.

```
y = as.numeric(CancerType == 'KIRC')
```

For all questions below, use `GeneExp$Gene15` as the explanatory variable and `y` as the response.

(i) Fit a simple linear regression model.

(ii) a) What is the estimated coefficient of the `Gene15` variable?
b) What is the p-value associated to the coefficient of `Gene15`?
c) What null hypothesis does this p-value correspond to?
d) What can we conclude?

(iii) Consider the model obtained in (i). The fitted value for the i-th subject can be computed using $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$. Find the values for $\hat{\beta}_0$ and $\hat{\beta}_1$.

(iv)

Problem 4.

(i) Uncomment and complete the following code to perform k-means clustering using the `GeneExp` matrix with three clusters:

```
set.seed(1)
# gene.kmeans = kmeans( , )
```

(ii) Uncomment and complete the following code to obtain a contingency table that has the patient counts for each cluster and cancer type. Make sure you ran the `set.seed` and `kmeans` functions together in the previous part:

```
# table( , CancerType)
```

(iii) According to the above contingency table, which cluster (from the k-means output) best represents the patients with KIRC cancer?

(iv) Argue that the k-means result shows the gene expression profile of a KIRC patient and that of a LUAD patient tends to be very different.

Problem 5. The following code obtains a distance object, distance matrix and performs a hierarchical clustering of the first 10 subjects in `GeneExp`. The dendrogram is shown below. Subject labels are colored according to the cancer type: blue for KIRC (subjects `3`, `6`), green for LUAD (subjects `1`, `10`), red for BRCA (the rest).

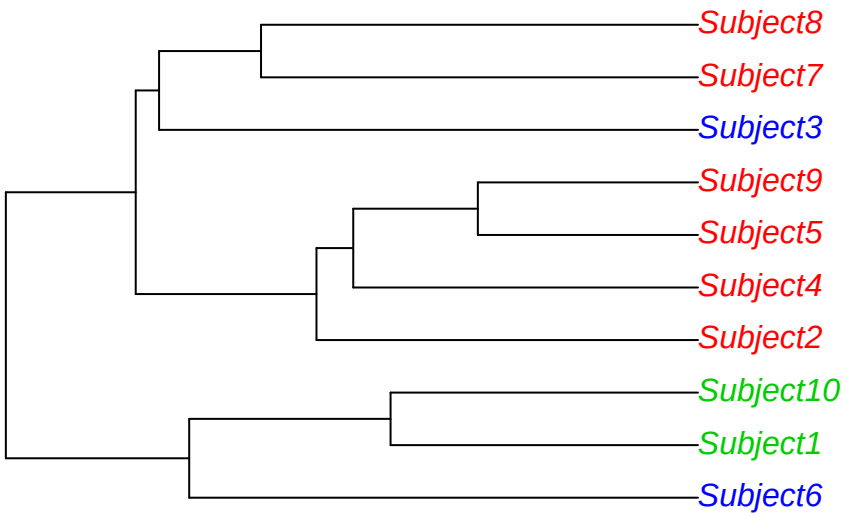
(i) Uncomment and complete the following code by using the `dist` function to get a distance object from the first 10 subjects in `GeneExp` using the Euclidean method.

```
#GeneExp.dist = dist(x=, method=)
```

(ii) Uncomment and complete the following code by using the `hclust` function to perform a hierarchical clustering of the first 10 subjects in `GeneExp` using the complete method.

```
#gene.hclust = hclust(d=, method=)
```

The following is a plot of the dendrogram obtained from hierarchical clustering of the first 10 subjects:



(iii) Create a distance matrix called `GeneExp.dist.mat` from the distance object `GeneExp.dist`

(iv) Find the Euclidean distance between `Subject1` and `Subject2` using `GeneExp.dist.mat`.

(v) Write your own code to compute the Euclidean distance between `Subject1` and `Subject2` using the matrix `GeneExp`. Do not use `dist`.

(vi) Which subject has the smallest distance to `Subject5`?

(vii) List two linkage criteria other than the complete linkage.

(viii) At the bottom of the dendrogram, subjects `1`, `6`, `10` form a cluster. This cluster has a sub-cluster containing subjects `1` and `10`. Below is the distance matrix for these 3 subjects. Using complete linkage, what is the distance between subject `6` and the sub-cluster of subjects `1` and `10`?

```
##           Subject1 Subject6 Subject10
## Subject1      0.00    17.99     10.87
## Subject6     17.99     0.00     16.95
## Subject10    10.87    16.95      0.00
```