Comprehension Check: Clustering

These exercises will work with the **`tissue_gene_expression`** dataset, which is part of the **`dslabs`** package.

Q1

1/1 point (graded)

Load the `tissue_gene_expression` dataset. Remove the row means and compute the distance between each observation. Store the result in `d`.

Which of the following lines of code correctly does this computation?

- d <- dist(tissue_gene_expression$x)
- d <- dist(rowMeans(tissue_gene_expression$x))
- d <- dist(rowMeans(tissue_gene_expression$y))
- **d <- dist(tissue_gene_expression$x - rowMeans(tissue_gene_expression$x))** ✔

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Q2

1/1 point (graded)

Make a hierarchical clustering plot and add the tissue types as labels.

You will observe multiple branches.

Which tissue type is in the branch farthest to the left?

- cerebellum
- colon
- endometrium
- hippocampus
- kidney
- **liver**

placenta

✔

Explanation
The plot can be made using the following code:

```
h <- hclust(d)
plot(h)
```

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Q3

1/1 point (graded)
Run a k-means clustering on the data with $K = 7$. Make a table comparing the identified clusters to the actual tissue types. Run the algorithm several times to see how the answer changes.

What do you observe for the clustering of the **liver** tissue?

- Liver is always classified in a single cluster.
- Liver is never classified in a single cluster.
- Liver is classified in a single cluster roughly 20% of the time and in more than one cluster roughly 80% of the time.
- Liver is classified in a single cluster roughly 80% of the time and in more than one cluster roughly 20% of the time.

✔

Explanation
The clustering and the table can be generated using the following code:

```
cl <- kmeans(tissue_gene_expression$x, centers = 7)
table(cl$cluster, tissue_gene_expression$y)
```

Liver is split into two clusters (one large and one small) about 60% of the time. The other 40% of the time it is either in a single cluster or in three clusters at roughly equal frequency.

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Q4

1/1 point (graded)
Select the 50 most variable genes. Make sure the observations show up in the columns, that the predictor are centered, and add a color bar to show the different tissue types. Hint: use the `ColSideColors` argument to assign colors. Also, use `col = RColorBrewer::brewer.pal(11, "RdBu")` for a better use of colors.

Part of the code is provided for you here:

```
library(RColorBrewer)
sds <- matrixStats::colSds(tissue_gene_expression$x)
ind <- order(sds, decreasing = TRUE)[1:50]
colors <- brewer.pal(7, "Dark2")[as.numeric(tissue_gene_expression$y)]
#BLANK
```

Which line of code should replace #BLANK in the code above?

- (selected) `heatmap(t(tissue_gene_expression$x[,ind]), col = brewer.pal(11, "RdBu"), scale = "row", ColSideColors = colors)`
- `heatmap(t(tissue_gene_expression$x[,ind]), col = brewer.pal(11, "RdBu"), scale = "row", ColSideColors = rev(colors))`
- `heatmap(t(tissue_gene_expression$x[,ind]), col = brewer.pal(11, "RdBu"), scale = "row", ColSideColors = sample(colors))`
- `heatmap(t(tissue_gene_expression$x[,ind]), col = brewer.pal(11, "RdBu"), scale = "row", ColSideColors = sample(colors))`

✔

Submit You have used 1 of 2 attempts

Answers are displayed within the problem

Ask your questions or make your comments about Clustering here! **Remember, one of the best ways to reinforce your own learning is by explaining something to someone else, so we encourage you to answer each other's questions (without giving away the answers, of course).**

Some reminders:

- Search the discussion board before posting to see if someone else has asked the same thing before asking a new question.
- Please be specific in the title and body of your post regarding which question you're asking about to facilitate answering your question.
- Posting snippets of code is okay, but posting full code solutions is not.
- If you do post snippets of code, please format it as code for readability. If you're not sure how to do this, there are instructions in a pinned post in the "general" discussion forum.

Discussion: Clustering

Show Discussion

Topic: Section 6: Model fitting and recommendation systems / 6.3.3: Clustering