Computer Assignment 5 - Hierarchical Clustering

Machine Learning, Spring 2020

YOUR NAME

```r
# If necessary, make sure to install.packages("gplots") so that the following command works:
library(gplots)
#Load the data
trial.sample = read.table("TCGA_sample.txt", header = TRUE)
#Store the subtypes of tissue and the gene expression data
Subtypes = trial.sample[,1]
Gene.Expression = as.matrix(trial.sample[,2:2001])
```

Hierarchical Clustering - TCGA Data

Because k-means works by finding the physical mean point in space for each cluster, we give it the raw data as input. However, our next two methods do not need the raw data, only the distances between points. As we are clustering the patients, we first must calculate pairwise distances between the patients. The *dist(X, method)* function can be used to calculate the pairwise distances between the rows of a matrix X using the specified *method* as a distance metric. Calculate the average Euclidean and average absolute pairwise distances between patients (a 217×217 matrix) using the following code:

```r
dist.euclid = dist(Gene.Expression, method = "euclidean")
dist.abs = dist(Gene.Expression, method = "manhattan") #absolute distance
```

There are many ways in **R** to perform hierarchical clustering. We will use the most basic version, *hclust()*. Run hierarchical clustering on the patients using the Euclidean and absolute distance matrix using the following code. Using *plot()* on the output object automatically draws the dendrogram. For a quick look at how well these clusters match the true cancer subtypes, we use *labels = types*.

```r
types = rep(0,217)
types[which(Subtypes == "Basal")] = "B"
types[which(Subtypes == "Normal")] = "N"

#Euclidean distance
hc.euclid = hclust(dist.euclid)
plot(hc.euclid, labels = types, cex = 0.5)

#Absolute distance
hc.abs = hclust(dist.abs)
plot(hc.abs, labels = types, cex = 0.6)
```

Note that the dendrograms are generated based on pairwise distances between the patients. We can visualize this by using *heatmap( )* to plot the original data according to the hierarchical cluster ordering.

```r
heatmap(Gene.Expression, Rowv = as.dendrogram(hc.euclid), Colv = NA,
        main = "Heatmap of TCGA data based on Euclidean Distance",
        xlab = "Genes", ylab = "Patients", col = redgreen(50))
```

Questions

1. Comment on the differences between the two dendrograms. Does one distance appear to cluster differently than the other? Which do you prefer?

YOUR ANSWER HERE

2. What agglomeration (linkage) method was used for the above clusterings? (Hint: check the manual page)

YOUR ANSWER HERE

3. Perform the euclidean distance clustering with `single` and `average` linkage. Print out their respective dendrograms and comment on how they differ. Do they differ at all from the above euclidean distance clustering?

```
YOUR CODE HERE
```

4. Hierarchical clustering does not automatically make a certain number of clusters from the data - this depends on where you "cut" the dendrogram. Draw a line on your plots showing where you cut the dendrogram. How closely do the subtypes appear to cluster in these two groups?

```
YOUR CODE HERE
```

5. The function *cutree(tree, ...)* will produce clusters based on a certain cut of the dendrogram *tree*. We can specify either height (h) or number of clusters (k). Use *cutree()* on the Euclidean distance clustering to assign clusters. What percentage of the each cluster is Normal and Basal? (Hint: Use *?cutree* to figure out exactly how to use this function.)

```
YOUR CODE HERE
```

6. Suppose that you read a scientific paper where the authors use hierarchical clustering on a data set and show a figure similar to what you just created. What kinds of questions might you be inclined to ask the authors regarding their clustering? Does the flexibility of the clustering (the choice of distance, the choice of linkage, etc.) make you more or less confident in the authors' clustering?

YOUR ANSWER HERE

Hierarchical Clustering - Cereals

Now let's practice Hierarchical Clustering (HC) with a different data set. The cereal data (named `cereals.csv`) contains cereal brands, manufacturers (also a variable called group, which is the same info, but group is numeric and manufacturer is categorical), and nutrition information (calories, protein, fat, sodium, fiber, carbs, sugar, potassium) per serving. Do a brief analysis of the data.

```
YOUR ANALYSIS HERE
```

Perform HC using euclidean distance and `average` linkage and all variables *except* `brand`, `manufacturer`, and `group`. Make sure to plot the dendogram, and label each leaf by `brand`. Do you see any interesting clusterings based on cereal brands? Use the `cutree()` function and experiment with h and k to see if you can find any meaningful clusters.

```
YOUR CODE, AND ANALYSIS, HERE
```

Now, change the labels on the dendrogram to `manufacturer` and see if you can find a meaningful k and h. Comment on your findings.

```
YOUR CODE, AND ANALYSIS, HERE
```

Using the k you decided upon using HC with `manufacturer` as your labels, run a k-nearest neighbors clustering on the cereal data. Compare to the clustering you found using HC. This comparison can be done a number of ways: you can project onto the first two PCs and plot, you can print out the cluster labels from each method and calculate how often the matched/didn't match, etc.

> Note: This question is intentionally open ended. Use code from previous assignments, and suggestions online, to provide a FULL hierarchical clustering analysis of this data.