Problem Set #2: Clustering pt. 1

Allison Collins

10/16/2019

Computation

1. Calculate Manahattan, Canberra, and Euclidean distances "by hand" (i.e., create the data, program each line, and make the calculations). What are the values for each measure?

First, create the vectors

```
p <- c(1,2)
q <- c(3,4)
```

For the Manhattan distance, calculating it by hand I obtain 4 as per below.

```
#Create the points
p <- c(1,2)
q <- c(3,4)
#Calc the distance
abs(q[1] - p[1])+ abs(q[2] - p[2])
```

```
## [1] 4
```

For the Canberra distance, calculating it by hand I obtain .8333333 as per below.

```
abs(q[1] - p[1])/(q[1] +p [1]) + abs(q[2] - p[2])/(q[2] +p [2])
```

```
## [1] 0.8333333
```

For the Euclidean distance, calculating it by hand I obtain 2.82847 as per below.

```
((q[1] - p[1])**2 + (q[2] - p[2])**2)**(0.5)
```

```
## [1] 2.828427
```

2 Use the dist() function in R to check your work. Were you right or wrong? (be honest in your reporting). If wrong, after debugging, where and why did you go wrong?

Leveraging the formulas, as seen below the distances I calculated by hand match R's distance formula and were thus correct.

```
#Bind the points
vector <- rbind(p,q)

#Calcualte all of the distances
dist(vector,"manhattan")
```

```
##   p
## q 4
```

```
dist(vector,"canberra")
```

```
##           p
## q 0.8333333
```

```
dist(vector,"euclidean")
```

```
##          p
## q 2.828427
```

3 What are the key differences between these measures, and why does it matter? How might you see these differences "in action" with these fictitious data?

These measures take different approaches to calculating distance.

The Manhattan distance elongates the distance as each step must be traversed as in a city block – you cannot cut diagonally. Canberra in turn weights the Manhattan distance. Euclidean distance draws a straight line between the two points.

These differences are important to note, because they have implications for the resulting distance calculations as we can see here with this fictitious data, where the values range from < 1 to 4. When we think about this in the context of our clustering analysis, it thus has implications for where different points will end up being clustered.

Old Faithful Data

4. Use some basic EDA techniques to present and discuss the data (e.g., visualize, describe in multiple ways, etc.)

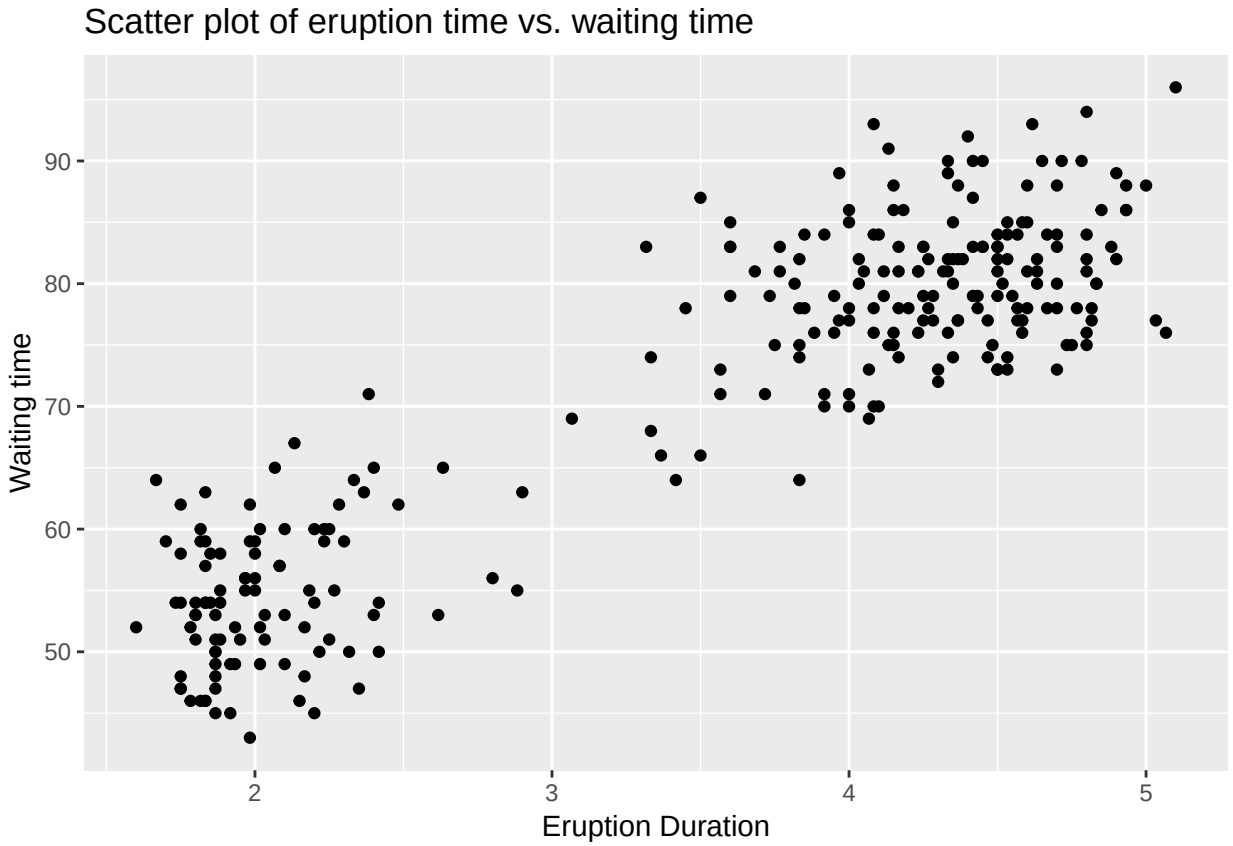
We can leverage the skim package to take a quick look at the old faithful data. We can see that the average time for an eruption is about 3.5 minutes and the average waiting time is about 70 minutes. Further, we can see the minimum – and maximum wait times (would be helpful for trip planning)

The histograms show us that actually for both, there are more values clustered at the high/low end vs. around the mean.

```
summary(faithful)
```

```
##    eruptions        waiting
##  Min.   :1.600   Min.   :43.0
##  1st Qu.:2.163   1st Qu.:58.0
##  Median :4.000   Median :76.0
##  Mean   :3.488   Mean   :70.9
##  3rd Qu.:4.454   3rd Qu.:82.0
##  Max.   :5.100   Max.   :96.0
```

```
faithful %>%
ggplot(aes(x = eruptions, y= waiting)) +
  geom_point() +
  labs(x = "Eruption Duration",
       y = "Waiting time",
       title = "Scatter plot of eruption time vs. waiting time")
```



Per above, we can see that it does look like there is a positive association between eruption duration and corresponding wait times, and again, there is clustering at both ends, e.g. short eruption + wait time plus long eruption and wait time vs concentration around the mean.

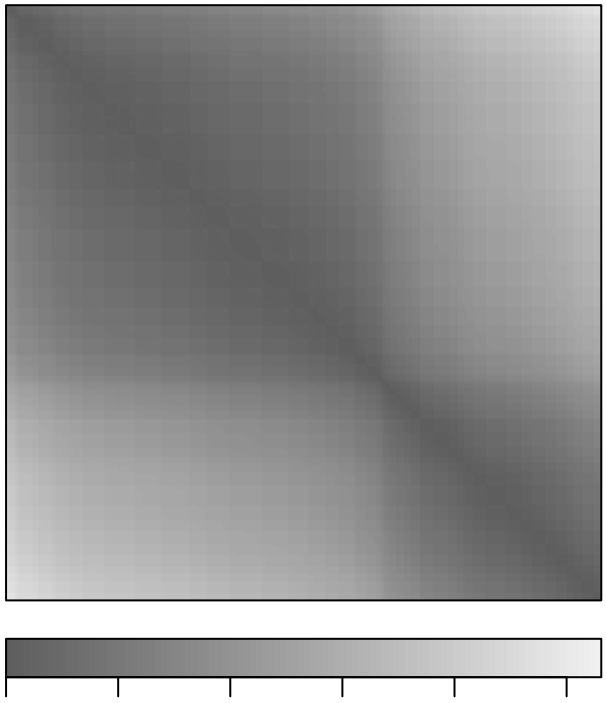
5.Calculate a dissimilarity matrix of these data.

I'll use the euclidian distance here to create the dissimilarity matrix

```
faithful_dist <- dist(faithful,
                      method = "euclidean")
```

6. Generate an ODI for the Old Faithful data. What do you see?

```
dissplot(faithful_dist)
```



In the above ODI, we can see (as with the scatterplot) these two clusters (the darker squares) where there is smaller distance (because short wait time + eruption time or long for both). This would suggest that the data is likely clusterable into two "quick" and "longer" clusters of eruption and wait time.

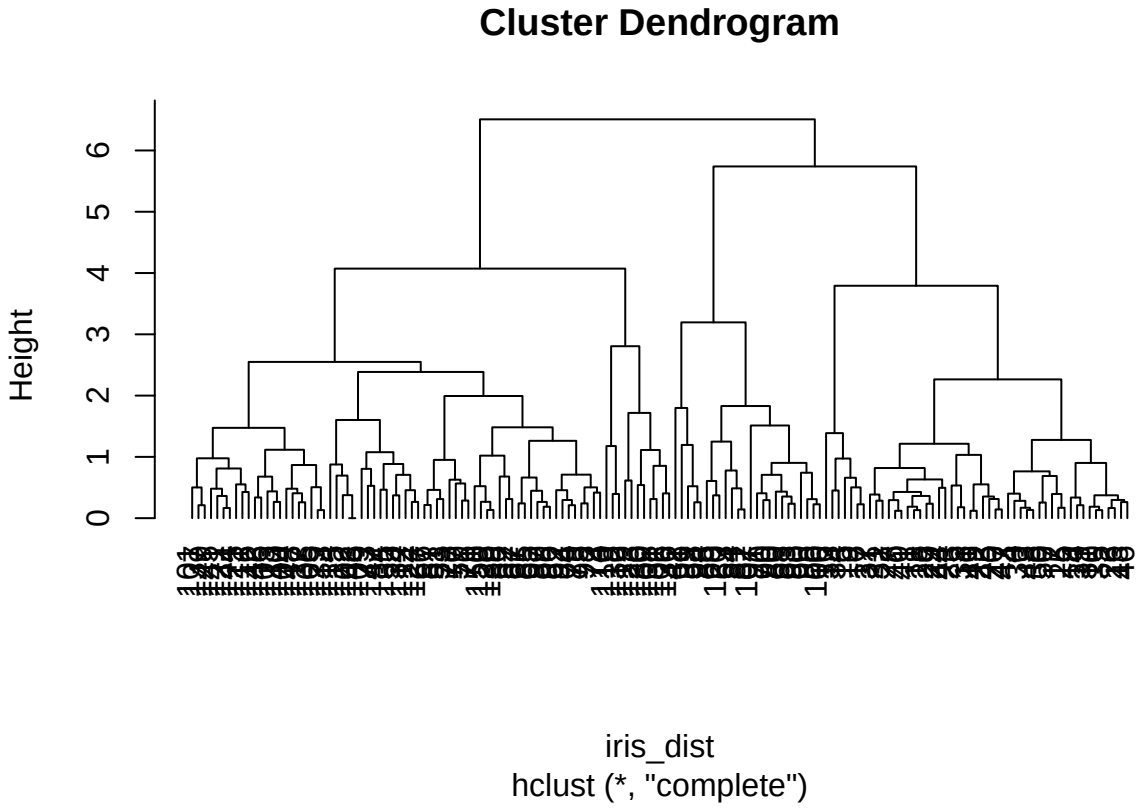
Iris data

7 Using any munging tools you'd like (e.g., dplyr from the Tidyverse), create a subset of the data excluding the species feature, scaling the features, and calculating a dissimilarity matrix (think "pipe" for stacking functions to do this quickly, e.g.)

```
iris_dist <- iris %>%
  select(-Species) %>%
  scale() %>%
  dist()
```

8. Fit an agglomerative hierarchical clustering algorithm using complete linkage on your subset data and render the dendrogram of clustering results. What do you see?

```
iris_complete <- hclust(iris_dist,
                        method = "complete"); plot(iris_complete, hang = -1)
```



9. Try cutting the tree at 2 and 3 branches and show these trees side-by-side. How do they differ?

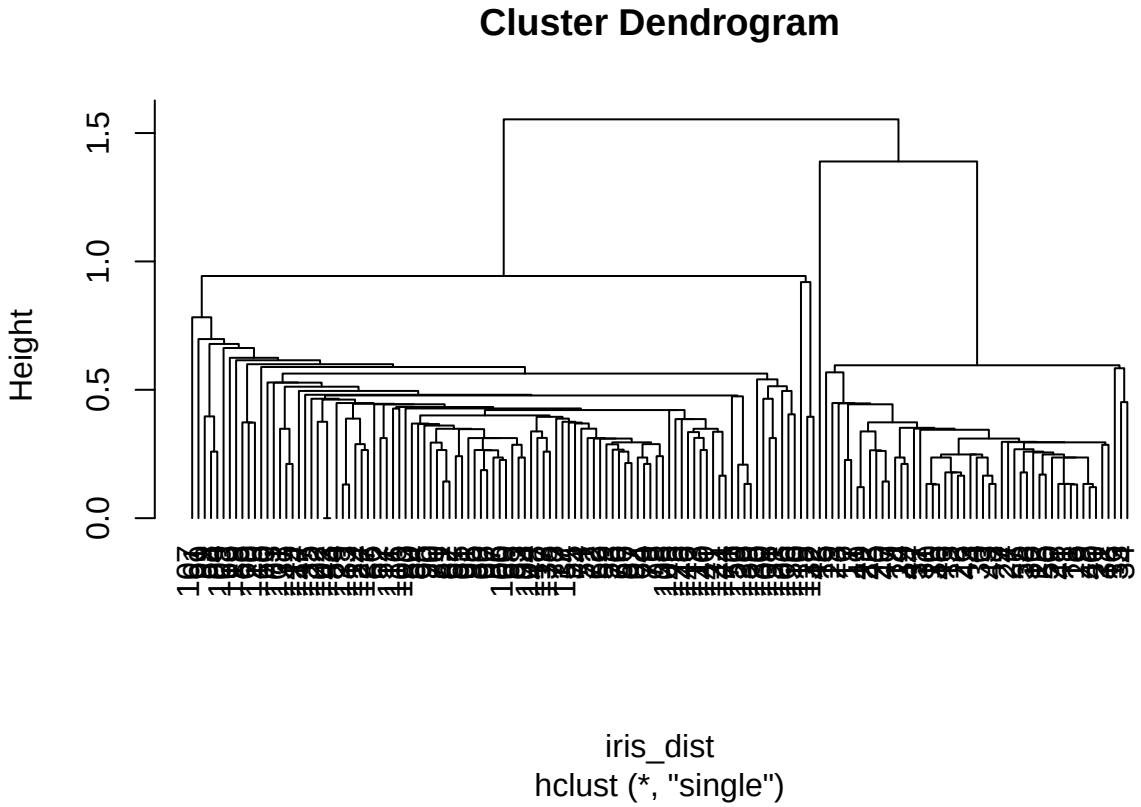
```
cuts <- cutree(iris_complete,
               k = c(2,3))
table(`2 Clusters` = cuts[,1],
      `3 Clusters` = cuts[,2])
```

```
##           3 Clusters
## 2 Clusters  1  2  3
##          1 49 24  0
##          2  0  0 77
```

We can see in moving from 2 to 3 clusters, the first cluster remains unchanged, but the second cluster fractures into two (with most moving to the third) – overall, it doesn't seem like a lot of information is lost as the clusters remain overall the same.

10. Now fit the algorithm using single and complete linkage and present each dendrogram side-by-side. Discuss the differences. What effects can we see in the clustering patterns when using different linkage methods?

```
iris_complete <- hclust(iris_dist,
                        method = "single"); plot(iris_complete, hang = -1)
```



Single linkage results in 2 very different sized clusters at the offset. Additionally, because of the way it fractures off individual, single points, the size of the ensuing clusters as we move down the tree differs between the two. ##Critical Thinking

1. You just assessed the clusterability of some feature space, R". Address the following questions:

a. How would you go about determining whether clustering made sense to consider or not?

As we did in class - and above - to first determine whether clustering made sense, I would do some visual exploratory analysis (e.g. scatterplots to look at relationships between the features), followed by scaling the data and creating a dissimilarity matrix/distance plot. This would allow me to see whether there might be natural groupings in the data suited to clustering – as for ex. we saw with the old faithful data, where the distance plot yielded two groupings at high and low times respectively. If all of the distances are uniform between points or totally dispersed on the scatterplots and there is no visible groupings, then it may not make sense.

b. What are techniques you would use, and what might you be looking for from each?

As discussed in (a), I would probably look at scatterplots, which would be an "informal" way firstly to show whether there is an association between features. If the plot showed points having no patterns at all and just spread throughout, the lack of association would make me think that clusters might not be the right approach. I would then look at more formal techniques like the dissimilarity matrix and visualizing the ODI

plot; again, I would want to see a pattern in distances between feature vectors as described above. I could also use other forms of testing, such as the Hopkins statistic, to test the null hypothesis of spatial randomness via sparse sampling and examine the test statistic.

c. How might these techniques work together to motivate clustering or not?

These techniques should be leading you down the same path. If scatterplots demonstrate groupings when we informally examine, the distances between those features when we go to calculate dissimilarity matrix and make ODI plots should be smaller - given those patterns were exhibited by the same underlying data. For instance we could have two groupings centered around 1,1 (e.g. (1.5, 1.5), (1.3, 1.2) and so on) and then similarly around 10,10 - the distance between the points in these groupings (which would visually appear similar in the scatterplot) would also manifest as smaller in the dissimilarity matrix // and patterns in the ODI plot.

d. And ultimately, can/should you proceed if you find little to no support for clusterability? Why or why not?

If we proceed after there is little support, we are not likely to get meaningful clusters. If our goal is to understand patterns and groupings in the data, we likely won't get high quality results back that can drive actionable insights (for example, if we were looking at data to drive a marketing campaign, we may not get groupings that have meaningful similarities to target in the campaign)

2. Locate (and read) a paper that applies the hierarchical agglomerative clustering technique. Address the following questions:

The paper I selected to read is: "Agglomerative Hierarchical Clustering Analysis of co/multi-morbidities"

a. Describe the author(s) process.

The authors have a dataset of a patient population in Texas, and are trying to understand better co-multiple morbities given the increase in moretality that they confer.

They first conduct exploratory data analysis, identifying outliers; doing some imputation; and creating binary variables for conditions. They then cluster (using the bottom-up method) on isolating multimorbidity patients Using Ward's method and Gower's distance matrix like we discussed in class, to highlight linkages in different conditions. The paper then discusses the resulting clusters from the agglomerative hierarchical technique (e.g. what patient demographics are in the given cluster; what illnesses are prevalent in the cluster) and how they believe this will be able to inform programming.

b. Do they go through similar steps as we covered this week both in setting the stage for clustering (e.g., assessing clusterability, calculating distance, etc.), as well as in fitting the algorithm? If not, what did they omit and does this omission impact their findings in your opinion?

The authors create a correlation matrix for the various conditions (using spearman's correlation analysis) with size and color to indicate coefficient, and use k-means clustering to validate before the agglomerative

clustering technique. This does not adhere to all of the precise steps we did in class, but it comprises some, and shows their efforts toward identifying patterns before diving into the agglomerative clustering. Thus it does not impact my interpretation of their final results. However, I would be curious to know a little more about their data cleaning, e.g. confirming they scale all the variables, etc. as this is not explicitly mentioned. In using Gowers' method we discussed in class to account for binary variables vs. continuous (as they are using binary for disease markers), that gives me confidence that will not be distorting clustering.

c. Describe at least one possible extension from the study that could emerge based on their findings.

This study proposed nine clusters, and analyzed the presence of conditions found within each cluster. An extension of the study that could emerge based on findings would be targeting treatment based on commonly associated diseases; for instance, one cluster yielded high prevalence of post-partum depression and addiction. Perhaps an exension could be joint programming to provide support for women after childbirth exhibiting signs of depression, that offers coping mechanisms and/or educates them on the associated risks of using drugs to cope with this.

An extension in terms of additional analysis to complete that could be interesting would be to perhaps isolate female vs. male patients and re-run the agglomerative clustering to see whether different clusters emerge, given that certain conditions are specific to men vs. women (+ maybe this could reveal different associations with other chronic conditions), that would allow a more targeted treatment plan as per above.