TUTORIAL 8

In this tutorial we will cover:

1. Calculating distances in R
2. k-means clustering in R
3. Hierarchical clustering in R

Mode of study:

You can work through this tutorial in your own time, or in Jan Mouton 2017, where help will be available. (Note that although Jan Mouton 2015 will be available for you to work in during the tutorial, no in-person assistance will be available in this venue on Friday.)

Solutions will be made available in Week 11.

Before you start

If you have not worked through Practical 8 yet, you should do so before starting this tutorial. You should also have completed the formative assessment on distance calculations.

DATASET

In Chapter 6 of the textbook, a whiskey analytics example is discussed. The accompanying dataset consists of 109 single malt Scotch whiskeys which have been described in terms of five general attributes:

- **Colour:** 14 different colour descriptions
- **Nose:** 12 different aroma descriptions
- **Body:** 8 possible values
- **Palate:** 15 possible values
- **Finish:** 19 possible values

These attributes are not mutually exclusive; in other words, a single observation can for example have more than one palate description. Consequently, there are 68 binary attributes captured for each whiskey in the dataset, corresponding to each of the 68 possible descriptors in the 5 general attribute categories. In addition, two region attributes are also included in the dataset:

- **Region:** Highlands, Lowlands, and Islay (corresponding to the division of Scotland into three Scotch-producing regions)
- **District:**
  - Northern, Western, Eastern, Speyside, Midlands, Mull, Skye, Jura, Orkney (all under the Highlands region);
  - Central, West, East, Northwest, Borders, Campbeltown (all under the Lowlands region);
  - South shore, North shore, Loch Indaal (all under the Islay region).

The dataset has been uploaded onto SUNLearn and is called `whiskey.csv`. An additional document describing the different attributes has also been made available and is named `whiskey_attributes.pdf`.

QUESTIONS

1. **Import the dataset into RStudio and store it in an object called `whiskey`. View the structure of the dataset and also view the first few rows of the dataset.**
   Note that there is no target variable.

2. Before attempting to cluster the whiskey data, you should revisit the calculation of distance measures. As described in the textbook, this could be useful to help find Scotch whiskeys similar to another. In the textbook (page 146) they identify the five whiskeys that are most similar to a whiskey named Bunnahabhain as:
   - Glenglassaugh
   - Tullbardine
   - Ardbeg
   - Bruichladdich
   - Glenmorangie

   Since the dataset contains attributes (characteristics) of different whiskeys, it seems sensible to use the Jaccard distance to find similar whiskeys, since the Jaccard distance between two whiskeys will be the proportion of all the characteristics shared by the two particular whiskeys. As discussed in the textbook (bottom of page 159), the Jaccard distance is appropriate for problems where the possession of a common characteristic between two items is important but the common absence of a characteristic is not. For example, when wanting to find whiskeys that are similar it may be considered significant if both whiskeys are peaty, but not that they are both not salty.

   Print the list of attributes for the six whiskeys listed above on screen (i.e. Bunnahabhain and its five most similar whiskeys).

The Jaccard distance between Bunnahabhain and Glenglassaugh was calculated in the formative assessment on distance calculations.

You should now use R to calculate the distances between Bunnahabhain and the five whiskeys and use this output to check your manual calculation from the formative assessment. Note that you should not include the `REGION` and `DISTRICT` attributes in your distance calculations.

Remember that the `dist()` function can be used to calculate the distance between attributes; specifying `method = "binary"` will return the Jaccard distance.

You can also attempt to write the code to calculate Jaccard distance yourself! While this is not strictly within the scope of this module, you should have the required R coding skills to do so, and it is a good way of practicing these skills.

3. The published paper by Lapointe and Legendre (referred to in the textbook) used 12 clusters for the whiskey data.

3.1 **Therefore, fit a *k*-means clustering model to the whiskey data, with $k = 12$. For reproducibility, use the command `set.seed(1)` before you run the clustering, and also use `nstart = 50` (in other words, try 50 different starting configurations). Store the output of this model in an object named `km.whiskey`.**

Remember that you should not include the NAME column in the cluster model. You should also omit the REGION and DISTRICT attributes.

3.2 **View the cluster assignments of the model fit in Question 3.1.**

Which whiskeys are in the same cluster as Bunnahabhain?

Also examine the cluster containing Aberfeldy, and the cluster containing Bruichladdich. Compare this to the textbook results on page 179.

4.

4.1 **Now fit a hierarchical clustering model to the whiskey data, and store this in an object named `hc.whiskey`.**

Use complete linkage, and also specify `method = "binary"` as part of your distance function, so that Jaccard distances will be used (in line with the textbook example). For reproducibility, use `set.seed(2)` before you fit the model.

4.2 **Plot the dendrogram for the model fit in Question 4.1.**

4.3 **View the cluster assignments for the model fit in Question 4.1 when there are 12 clusters.**

Remember that you can use the `cutree()` function to do this.

Check how the cluster assignments of Bunnahabhain, Aberfeldy and Bruichladdich have changed compared to the *k*-means clustering.

After completing this week's practical and tutorial you should be able to fit basic clustering models in R.

You should also now read the relevant pages in the textbook, where the whiskey example from this tutorial is discussed. This is on pages 145 – 146, 164, 169 – 170 and 178 – 183 in the textbook.