CS5100 Assignment 3

This assignment must be submitted by Monday 16 December 2019, 17:00. Feedback will be provided by Wednesday 22 January 2020.

Learning outcomes assessed

The learning outcomes assessed are:

- develop, evaluate, and use effectively unsupervised learning models
- apply methods and techniques such as K-means and hierarchical clustering
- understand and experiment with different linkage functions
- implement unsupervised learning algorithms in R

Instructions

In order to submit, copy all your submission files into a directory, e.g., `DA`, on the server `linux` and run the script

```
submitCoursework DA
```

from the parent directory. Choose `CS5100` and Assignment (=exercise) `3` when prompted by the script. You will receive an automatically generated e-mail with a confirmation of your submission. Please keep the e-mail as a submission receipt in case of a dispute; it is your proof of submission. No complaints will be accepted later without a submission receipt. If you have not received a confirmation e-mail, please contact IT Support team.

You should submit files with the following names:

- `HC.R` should contain the source code for your hierarchical clustering;
- optionally, `other.R`, which are all other auxiliary files with scripts and functions (please avoid submitting unnecessary files such as old versions, back-up copies made by the editor, etc.);
- `report.pdf` should contain the numerical results and discussion.

The files you submit cannot be overwritten by anyone else, and they cannot be read by any other student. You can, however, overwrite your submission as often as you like, by resubmitting, though only the last version submitted will be kept. Submissions after the deadline will be accepted but they will automatically be recorded as being late and are subject to College Regulations on late submissions.

The deadline for submission is **Monday, 16 December 2019, 17:00**. An extension can only be given by the academic advisor (and in some cases by the office, but not by the lecturer).

Note: All the work you submit should be solely your own work. Coursework submissions are routinely checked for this.

Tasks

1. Implement the hierarchical agglomerative clustering with the following linkage: single, complete, average and centroid (as described in the lectures, Chapter 7, or in [2], Section 10.3.2, or in [1], Section 14.3.12). Your program should be written in R. *You are not allowed to use any existing implementations of hierarchical agglomerative clustering in R or any other language, and should code the hierarchical agglomerative clustering from first principles.* However, you can use built-in R functions to visualise your results.

2. Apply your program to the `NCI microarray` data set which can be downloaded from the course's Moodle page. This dataset has 64 columns and 6830 rows, where each column is an observation (a cell line) and each row represents a feature (a gene); see [2], pp. 4-5, for further information. (Therefore, the data set is represented via its *transposed* data matrix.) The label of each example can be found at the course's Moodle page. Preprocess the data set as appropriate.

3. Discuss the performance of hierarchical agglomerative clustering when using different linkage functions.

4. Apply the R function `kmeans()` to the above `NCI microarray` data set with different K and discuss its performance.

5. Compare and contrast the performance of K-means and hierarchical agglomerative clustering.

6. **Optional:** Discuss how to choose the number of clusters in the K-means and hierarchical agglomerative clustering.

Feel free to include in your report anything else that find interesting.

Marking criteria

To be awarded full marks you need both to submit correct code and to obtain correct results on the given data set. Even if your results are not correct, marks will be awarded for correct or partially correct code (up to a maximum of 70%). Correctly implementing the hierarchical agglomerative clustering (Task 1) will give you at least 50%.

Extra marks

It is possible to get extra marks (at most 10%) that will be added to your overall mark (the sum will be truncated to 100% if necessary). Extra marks will be given for doing the optional task and for any interesting observations about the dataset and methods (discuss these in your report).

References

[1] Trevor Hastie, Robert Tibshirani and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Springer, New York,second edition, 2009.

[2] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*, Springer, New York, 2013.