Scientific Programming Assignment 3

MPhil in Computational Biology

November 5, 2021

If there are errors found, I will update the assignment.

Due date: TBC

Please submit your report to the Moodle website as a single PDF. Name your file `spa3_XXX.pdf`, where XXX is your six digit coursework number.

Your report must be a maximum of ten pages, excluding the appendix. (List your code in the appendix.) This course work will consist of 40% towards your overall mark for this module. Items marked **(Advanced:)** should be attempted if you have time and are confident with your work on the rest of the assignment.

1 Clustering methods [40 marks]

Your task is to implement and study the performance of different clustering algorithms on the classic 'iris' dataset available in R (see `help(iris)` for details)

Choose three of the following five methods for cluster analysis, described at: `https://www.digitalvidya.com/blog/the-top-5-clustering-algorithms-data-scientists-should-know/`

- K means
- Mean-Shift Clustering Algorithm
- DBSCAN – Density-Based Spatial Clustering of Applications with Noise
- EM using GMM – Expectation-Maximization (EM) Clustering using Gaussian Mixture Models (GMM)
- Agglomerative Hierarchical Clustering

Write a report that compares the three methods, describing the key decisions that you had to make when implementing each algorithm. Compare the three algorithms that you chose in terms of performance, efficiency and ease of understanding. (Your write-up should include references to where the methods are defined, and provide (in under 1 page) a summary of how each method works.) [32 marks]

(Advanced:) Find a larger biological dataset to test your results on, so that you can investigate how your methods work with larger data. Describe your reasons for selecting the larger dataset and your findings. [8 marks]

Remember: format this as a report, with references/bibliography and using Tables/Figures with appropriate captions/legends.