STU33011 MLA Assessment

This assignment is worth 20% of your mark for STU33011, and will consist of a report with a maximum length of ten pages. Assessments should be submitted via Turnitin by 10am Monday 11th November 2019. If you have any questions, contact me on arwhite@tcd.ie.

I would like you to analyse the Rituximab data set. This data set is available to download on blackboard and on the class website: https://www.scss.tcd.ie/~arwhite/Teaching/STU33011/rituximab.csv.

The Rituximab dataset is a flow cytometry dataset from a drug-screening project to identify agents that would enhance the anti-lymphoma activity of Rituximab, a therapeutic monoclonal antibody. The data consist of nine variables: two scattering parameters (FSC.H, SSC.H), five fluorescence parameters (FL1.H, FL2.H, FL3.H, FL1.A, FL2.W), a time variable (Time) and gating variable (Gate). I do not recommend using the Time variable in your analysis.

A common goal of flow cytometry analysis is to perform “gating”, an exercise which aims to identify distinct populations of cells within a data set. The Gate variable records the results of one such gating analysis. This variable can take values {-1, 1, 2}. In this analysis, two gates have been identified, and data points assigned to these gates have been labelled 1 and 2 respectively. Entries with a value of -1 represent “noisy” data points or outliers, i.e., they do not have any cohesive or useful structure from the perspective of an analyst.

Your analysis should have two key objectives:

- Use unsupervised learning methods to identify subsets of the data which are similar to the identified gates?
- Using supervised learning methods, can you accurately predict which data are assigned to the identified gates?

Your report of this analysis should consist of no more than ten A4 pages (please note that this is a limit, not a target). Please do not include any code that was used in the analysis in your report; graphs and summary output are fine. Your report should be well written and understandable to somebody unfamiliar with multivariate analysis; for example, it should be accessible to a researcher in the School of Immunology.

The report will be marked subject to the following criteria. Creativity will be rewarded:

- Data description and visualisation [20%]
- Appropriate use of unsupervised learning methods [30%]
- Appropriate use of supervised learning methods [30%]
- Clarity of writing and exposition [20%]