Unsupervised learning

Assignment 2

Miguel Rodo

Introduction

This assignment focuses on using dimension reduction methods to analysis

In this assignment, you will use dimension reduction methods to perform two different exercises:

- The first is to visualise flow cytometry data.
- The second is to predict TB progression using protein expression data.

Data availability

Both datasets are available as CSV files in the assignment page.

Question 1: Visualisation of Flow Cytometry Data

Introduction

In this exercise, you will work with a large flow cytometry dataset consisting of the expression levels of approximately 20 markers across different cell types. Your task is to apply various dimensionality reduction techniques to visualise this data in 2D and evaluate the effectiveness of these methods in separating different cell types.

Data

The file name is `flow_cytometry_data.csv`.

A subset of the data is as follows:

major	minor	cd3	cd4	cd8	cd19	cd20	cd16	cd56	cxcr3
b	memory	0.62	2.60	1.4	5.4	4.10	0.73	3.4	2.10
b	memory	2.60	0.98	1.7	4.6	4.70	-0.34	1.9	-0.54
cd4	centralmemory	5.40	4.90	1.4	2.3	2.20	2.80	1.5	2.50
cd4	effectormemory	4.40	4.70	2.5	2.4	0.99	0.55	2.2	3.00
tcrgd	vgamma9vdelta2	4.80	2.30	4.5	2.6	2.60	4.90	1.5	4.80

Each row is a cell, and the columns are as follows:

- `major`: Major cell type (e.g. cat versus elephant)
- `minor`: Cell subtype (e.g. amongst cats, a tiger versus a leopard).
- The remaining columns represent the expression levels of different markers. Note that not all columns were displayed.

Some cell types are common and others are rare.

For example, the major cell types compose the following percentages of the data:

major	percentage
b	8.8
cd4	46.0
cd8	36.0
mait	2.2
nk	5.7
tcrgd	1.2

Within each major cell type, the minor cell types are distributed as follows:

major	minor	percentage
b	memory	1.30
b	naive	6.70
b	plasma	0.74
cd4	centralmemory	9.40
cd4	effectormemory	15.00
cd4	naive	12.00
cd4	th1	3.90
cd4	th2	5.40
cd8	centralmemory	12.00
cd8	cytotoxic	4.10
cd8	effectormemory	13.00

major	minor	percentage
cd8	naive	6.40
mait	cd4+	0.59
mait	cd8+	1.70
nk	cd56+cd16-	0.37
nk	cd56bright	3.00
nk	cd56dim	2.40
tcrgd	vdelta1	0.48
tcrgd	vgamma9vdelta2	0.71

Objective

Your task is to apply a range of dimensionality reduction techniques to create 2D visualisations of this data. You will then critically evaluate the effectiveness of these methods in separating different cell types.

Tasks

1. **Visualisation:**

- Apply the dimensionality reduction method.
- Create 2D plots for each method. Use different colors or markers to distinguish between major and minor cell types.
- For each method, vary the parameters to optimise the visualisation.

3. **Evaluation:**

    - Critically assess the effectiveness of each method in separating different cell types. Consider factors such as separation of cell types, preservation of data structure, and runtime.
    - Discuss the strengths and weaknesses of each method in this context.

Question 2: Predictive Modelling of TB Progression Using Protein Expression Data

Introduction

In this exercise, you will work with protein expression data from a cohort of individuals, some of whom developed tuberculosis (TB) within a year and some who did not. Your goal is to use cross-validation to evaluate how different dimensionality reduction methods affect the predictive accuracy of a logistic regression model with LASSO regularisation.

Data

The file name is `protein_expression_data.csv`.

The data consists of data for 154 individuals, the following numbers of which did or did not develop TB:

tb	count
no	104
yes	50

For each individual, we have measurements of 2604 protein levels. The data is structured as follows:

sample_id	tb	protein	value
sample_1	yes	14-3-3	4.16
sample_1	yes	14-3-3 protein gamma	3.89
sample_1	yes	14-3-3 protein theta	4.16
sample_1	yes	14-3-3 protein zeta/delta	4.74
sample_1	yes	14-3-3E	4.47
sample_1	yes	17-beta-HSD 1	4.29

The columns are as follows:

- `sample_id`: Unique identifier for each individual.
- `tb`: TB status (yes or no).
- `protein`: Protein name.
- `value`: Protein expression level.

Accordingly, each individual will have 2604 rows.

Objective

The objective is to evaluate how different dimensionality reduction methods affect the predictive accuracy of a logistic regression model with LASSO regularisation.

Tasks

1. **Cross-validation**:

- Within each fold of an 80/20 cross-validation split:
  - Apply the dimensionality reduction method
  - Fit a logistic regression model with LASSO regularisation
  - Evaluate the model's performance using the area under the ROC curve (AUC).
- For each method, vary the parameters to optimise the AUC.

2. **Evaluation:**
   - Discuss the strengths and weaknesses of each method.

Submission

1. **Deadline:**
   - The deadline for this assignment will be four weeks from today, with the exact (and possibly updated) date on the assignment tab.
2. **Submission Format:**

- *Documents*: Submit the following:
  - Code: RMarkdown/Quarto document (along with any R scripts)
  - Rendered document: PDF/HTML
  - Plagiarism declaration
- *Code*: I actually prefer including code in the main document, but this is not essential. If you do, only code that is "important" or interesting, such as the implementation and running of an algorithm, but do not show boiler plate code, like reading in data and creating plots.
- *Page limit*:
  - If you include code in the main document, the page limit is 25.
  - If you do not include code, the page limit is 20.
  - This must be strictly adhered to, and I will stop marking after that many pages.
  - An appendix is permitted, but this is merely for supporting information or figures, and the main body of the report should essentially be self-sufficient.

3. **Collaboration**:

- "Light" discussion amongst yourselves is allowed, but the following are automatically plagiarism:

  - Any copying of code or text
  - Two assignments that do not differ significantly in their structure or content

4. **Grading:**

   - The assignment will be graded based on the following criteria:
     - Correctness of the code
     - Clarity and quality of the visualisations
     - Depth and quality of the analysis
     - Overall presentation and clarity of the report
   - The first question counts 20 marks, and the second 30 marks.