ST443: Group Project

Deadline: **4pm, 11 December, 2024**

The group project involves applying various statistical learning methods to binary classification on real-world data and evaluating their performance. It consists of two parts: the first addresses a binary classification problem using an RNA dataset, while the second focuses on a binary classification problem with an emphasis on feature selection in a drug discovery dataset.

1 Task 1: binary classification

The first part of this project involves applying statistical machine learning techniques to a binary classification problem, using the dataset provided on Moodle as a compressed CSV file, `data1.csv.gz`.

1.1 Data Description

The dataset contains RNA expression level measurements for $p = 4123$ genes across $n = 5471$ cells (a subset of the dataset from Suo et al. (2022)). The CSV file has 5471 rows (excluding the header) and 4124 columns. Each row represents a single cell. The first column specifies the cell type (`TREG` or `CD4+T`), while each of the remaining columns records the logarithmically normalized RNA expression level for a specific gene (with gene names provided in the column headers). Due to the large file size, it is recommended to read the compressed file directly into `R` or `Python`, for example, by using the command `read.csv("data1.csv.gz")`.

For background (not required to complete the project), CD4-positive T cells are an essential class of immune cells in the human body. In 1995, researchers discovered a subclass of these cells, known as T regulatory cells, which play a crucial role in controlling autoimmune diseases. Previously, distinguishing this subclass from other CD4-positive cells was challenging. However, advances in technology now allow for gene expression measurements in individual cells, making it much easier to differentiate T regulatory cells (`TREG`) from conventional CD4-positive T cells (`CD4+T`) using appropriate statistical learning tools.

1.2 Task

This task consists of the following sub-tasks:

T1.1 Explore the data to generate summary statistics and plots that help the reader understand the data, with a focus on information relevant to the classification task.

T1.2 Train and evaluate the following classifiers (covered in the course):

- Linear Discriminant Analysis (LDA)
- Logistic classifier
- Quadratic Discriminant Analysis (QDA)
- Nearest Neighbor Classifier (k-NN)
- Gradient Boosting Decision Trees (GBDT)
- Random Forest
- Support Vector Machine (SVM).

Additionally, train and evaluate these classifiers using PCA with 10 components.

T1.3 Train and evaluate three classifiers of your choice with the goal of improving the F1 score. You may use any classifier or combination of methods covered in the course, including those listed in **T1.2**, as well as methods like bagging, boosting, and regularisation.

T1.4 Choose the best approach among those you have tested and implement your predictor as a function in your code.

1.3 Requirements

- **Label classes**: Define `TREG` as the positive class and `CD4+T` as the negative class.
- **Evaluation metrics:** Use accuracy, balanced accuracy, AUC, F1 score, confusion matrix, ROC. Balanced accuracy is defined as the average of recall obtained on each class:

$$\text{Balanced accuracy} = \frac{1}{2}\frac{\text{TP}}{\text{TP}+\text{FN}} + \frac{1}{2}\frac{\text{TN}}{\text{TN}+\text{FP}}.$$

F1 score is defined as the harmonic mean of the precision and recall:

$$\text{F1} = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

where

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \text{ and } \text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}.$$

2 Task 2: feature selection

The second part of the project focuses on achieving high classification accuracy using a minimal number of features. You will use the dataset provided on Moodle as a compressed CSV file: `data2.csv.gz`.

2.1 Data Description

The dataset contains binary features describing the three-dimensional properties of a molecule in a compound or a random probe for $p = 100000$ (50000 real variables and 50000 random probes, randomly permuted) across $n = 800$ compounds. The CSV file has 800 rows (excluding the header) and 100001 columns, with each row representing a compound. The first column indicates whether the compound bound to a target site on thrombin. The DOROTHEA dataset was used in the NIPS 2003 feature selection challenge Guyon et al. (2004). Due to the large file size, it is recommended to read the compressed file directly into `R` or `Python`, for example, by using the command `read.csv("data2.csv.gz")`.

For background (not required to complete the project), drugs are typically small organic molecules that achieve their desired activity by binding to a target site on a receptor. The first step in discovering a new drug is usually to identify and isolate the receptor to which it should bind, followed by testing numerous small molecules for their ability to bind to the target site. This process leaves researchers with the task of determining what distinguishes active (binding) compounds from inactive (nonbinding) ones. Such insights can then be used in designing new compounds that not only bind but also possess additional properties required for a drug.

2.2 Task

This task consists of the following sub-tasks:

T2.1 Explore the data to generate summary statistics and plots that help the reader understand the data, with a focus on information relevant to the classification task.

T2.2 Train and evaluate feature selection methods with the goal of achieving high balanced accuracy using a minimal number of selected features. Use three different approaches of your choice that were covered in the course.

T2.3 Train and evaluate up to three additional feature selection methods, other than those in **T2.2**, which may include methods not covered in the course, with the goal of achieving high balanced accuracy using a minimal number of selected features. Provide an explanation of each new method.

For **T2.2** and **T2.3**, your evaluation results should include the balanced accuracy and the number of selected features achieved by each method for various numbers of selected features. You may also compare the balanced accuracy achieved by different methods for a similar number of selected features.

3 Submission and assessment

Each group needs to submit **one written report** and **two notebooks** (one for each task).

3.1 Written report

Your written report should be in PDF format and named `report.pdf`. The main text of the submitted written report should be limited to **nine content pages**, including all figures and tables. Additional pages containing references do not count as content pages. The formatting of your report may be similar to that of the NeurIPS conference; e.g. see here: https://media.neurips.cc/Conferences/NeurIPS2023/Styles/neurips_2023.pdf. If you use Latex, you may use the NeurIPS 2024 Latex style: https://media.neurips.cc/Conferences/NeurIPS2024/Styles.zip choosing the preprint option

```
\usepackage[preprint]{neurips_2024}
```

We suggest your report to include:

- An abstract limited to one paragraph.
- A description of the data (summary statistics and plots will be helpful) and the questions you are interested in answering.
- For Task 1:
  - For **T1.2**, show a summary table of the evaluation results achieved and highlight the winning classifiers for different evaluation metrics. For example, your summary table may look similar to Table 3.1.[1]

Table 1: Summary of evaluation results. The numbers highlighted in boldface indicate the best classifiers for each evaluation metric.

	Accuracy	Bal. Acc	AUC	F1
LDA	0.751	0.71	0.791	0.664
Logistic	0.854	0.83	0.822	0.725
QDA	0.682	0.717	0.68	0.726
k-NN	**0.945**	0.75	**0.826**	0.736
GBDT	0.701	0.825	0.76	0.667
RF	0.793	**0.9**	**0.826**	0.662
SVM	0.891	0.74	0.652	**0.756**

  - For **T1.3**, review the approaches you tried. The explanation of your approaches should be sufficiently detailed so that a student who has taken the ST443 course will be able to reproduce your classifier.
  - Summary of the final approach you used and the reason why you chose that approach.
  - Summary and discussion of the results.

[1]The number shown in the table are random and used only for illustrative purposes.

- Task 2:
  - Provide similar information as for Task 1. In the summary table of results, include a column for the number of selected features. You may also summarise your obtained results in a plot with the number of selected features on the x-axis and the balanced accuracy on the y-axis.

3.2 Code

You can implement your statistical learning algorithms in `R`, `Python` or any other programming language (but each group should choose only one language for the entire project). Please submit **two notebooks / source code files, one for each task** of the project, containing all your well-documented source code. Name your files so that it is clear which task they belong to (e.g. for Task 1 your file names should look like `task1.R`, `task1.Rmd`, `task1.py` or `task1.ipynb`).

Your code should reproduce the findings discussed in your report. You may assume that the code will be executed in the same local directory containing the data file, and that all relevant packages have been pre-installed.

For Task 1, your code should **include a function named** `mypredict()`, which takes no input arguments. Your function should read in a compressed csv file named `test.csv.gz` from the working directory, which is in the same format as the provided data (`data1.csv.gz`). Your function should predict the class label of each row and return your prediction labels saved in a plain text file (one prediction label per line, in the same order as in the input file).

3.3 Marking scheme

- Task 1 (total 45%)
  - Illustration of the dataset (10%)
  - Explanation of your approaches (15%)
  - Performance of your final classifier (10%)
  - Quality of coding (10%)
- Task 2 (total 45%)
  - Illustration of the dataset (10%)
  - Explanation of your approaches (15%)
  - Performance of your feature selection methods (10%)
  - Quality of coding (10%)
- Presentation quality of the written report (10%)

All group members are assumed to have made equal contributions. If this is not the case, please indicate at the beginning of your report a breakdown of contributions of all members. The grade for each group member will be a function of the contribution of each group member using the following equation.

$$\text{member grade } = \text{report grade } \times \frac{\text{member contribution}}{\text{maximum contribution}}.$$

For example, for a group with 5 members contributing, 30%, 20%, 20%, 20%, 10% and the report grade is 75 (out of 100), the individual grades are

$$75 \times \frac{30}{30} = 75, \quad 75 \times \frac{20}{30} = 50, \quad 75 \times \frac{20}{30} = 50, \quad 75 \times \frac{20}{30} = 50, \quad 75 \times \frac{10}{30} = 25.$$

4 Suggested timeline

- Week 5–7: Contact group members, decide who contributes to which part, run exploratory analysis on the data, and explore approaches that you may try.
- Week 8–10: Analyse the data, train and evaluate prediction methods.
- Week 10–11: Write and submit the report.

Deadline to submit your coursework report:

- **4:00pm, 11th December 2024, Wednesday of the 11th week.**

References

Guyon, I., Gunn, S., Ben-Hur, A. and Dror, G. (2004) Result Analysis of the NIPS 2003 Feature Selection Challenge. In *Advances in Neural Information Processing Systems* (Saul, L., Weiss, Y. and Bottou, L., eds.), vol. 17, MIT Press.

Suo, C., Dann, E., Goh, I., Jardine, L., Kleshchevnikov, V., Park, J.-E., Botting, R. A., Stephenson, E., Engelbert, J., Tuong, Z. K. et al. (2022) Mapping the developing human immune system across organs. *Science*, eabo0510.