CSC3432

Coursework 2 requirements

Dr. Jaume Bacardit   Dr. Paweł Widera

2022-10-28

Overview

The **goal** of this coursework is to **solidify what you have learned** throughout this module about biomedical data, machine learning algorithms, experiment design, model performance evaluation, and the explainability. It also gives you a **practical experience** with implementation of machine learning **pipelines**, and allows you to **explore** an unknown dataset **on your own**.

> In the report, you will be required to document the process of analysis of **one selected** biomedical dataset. Each dataset provides information in a **different format** and will require slightly **different approach**.

1. **TABULAR** - you will explore **patient records** and **classify** them
2. **IMAGES** - you will build deep learning **classification** models for **MRI images**
3. **TEXT** - you will build deep learning **classification** models for **protein sequences**

The datasets and their description (origin, attributes, related research questions) are provided on Canvas.

What to submit?

- The report (a single `PDF` file, max. 2000 words),
- Python scripts for data processing, visualisation, and classification (`tar.gz` , or `zip` archive file).

The archive **should not include** the original datasets or irrelevant files.

> The deadline for submission is **2022-12-16 16:30**. You must submit the files on **Canvas**.
> Late submissions (up to 1 week) will be penalised with −5 points per day.

Before you start

Make sure to use the recommended project directory layout with `data` , `scripts` , and `results` directories. All scripts should run from the project directory level, for example as `./scripts/task-1.py` .

You don't have to copy the data. Just use the relative paths directly in your scripts (e.g. `data/visits.csv`).

Analysis

Start from reading the data. As these are the real-world dataset (e.g. patient records) they will be a bit messy. You will need to make several adjustments before you can work with them. Read the data description file to understand how these data were collected and what is the meaning of each attribute.

You will have to deal with non-default formatting, and different attribute types (numeric values, text values, dates). Do not forget to make a distinction between nominal and ordinal attributes, and decide how to handle the missing values. You might have to integrate the data from multiple files and folders into a single data frame.

When you are done, turn all the steps into a single `preprocessing.py` script, that generates a new `processed.pkl.gz` file in the project `data` folder. Use this new data file in further analysis.

> There are many possible **approaches** to each problem, and it is unclear which would bring the **best results**. Therefore, you will have to **experiment**!

Explore that new dataset and think about possible angles of analysis. What would be interesting to know about the patients? Maybe you could check how diverse the patient population was, or how the patients progressed over time? Maybe you could cluster the patients, or predict their health status or disease progression from the baseline data? Think about trying different approaches, starting simple and refining it over time (e.g. testing different classification algorithms, performing feature selection, tuning the hyperparameters, changing the model architecture, or using different model interpretation techniques).

Make sure you use a robust **evaluation process** (e.g. 5-fold cross-validation procedure) and appropriate **measure of success** (e.g. F1-score), and that you report and compare useful statistics (e.g. mean value and the 95% confidence interval).

> All scripts must be **reproducible**, so if randomness is involved, fix the **seed** of the random generator.

Did all the methods behave as expected or did the results surprise you? Was prediction as good on the test folds as on the training folds? Please discuss in the report.

Description

For each piece of analysis, start from explaining **why** you decided to do it (what you are hoping to achieve) and **justify** the choice of methodology. Then, work on the analysis itself and make one or two **figures** representing the results. Make sure that each figure has a **caption** that briefly describes what it shows and explains the meaning of visual elements (colours, shapes, etc.). Finally, try to interpret the figure and **discuss** what **knowledge** your analysis has **revealed**.

Conclusions

Your report should end with a conclusion section, that summarises your results (e.g. show a table comparing all approaches together) and suggests some further improvements (future work).

Marking criteria

We will not judge the quality of your code, but we might run your scripts to see if they work and generate the same results as reported. We will deduct points if your figures / results are not reproducible.

Task	Marks
Exploratory analysis	10%
Data pre-processing	10%
Methods	30%
Results	20%
Discussion/Conclusions	20%
Form (writing style, use of figures, etc.)	10%

Before submission

Add a final **Comments** section to your report. Use it to answer the following questions:

- If there are known problems with your code, please list them, and explain how you might have fixed them if you had more time. We are likely to give you partial credit for it, if we see you understand the problems well.
- Did you discuss the final report with other students? If so, let us know who you talked to and how you helped each other.
- Which of the recommended additional readings you found to be most helpful throughout the course, and in completing the report? Please make a list.
- Do you have any comments about this assignment that you would like to share? Was it too long or too hard? Were all the requirements clear? Did you have fun working on it, or did you hate it? Do you think you learned something while doing it, or was it a waste of time? Constructive feedback will be highly appreciated! Be as open as you want, it will not affect your mark.

Report structure

To format the report, use a structure shown on the right.

```
1. Exploratory analysis
2. Pre-processing
3. Machine learning
 3.1. Method A
 3.2. Method B
 ...
4. Conclusions
5. Comments
```

Focus on the **core tasks** of applying machine learning methods to analyse the data. Describe your initial approach to the problem and all further refinements. Provide justification for the choices made (e.g., how to handle missing data, or which clustering algorithm to use). Finally, do not only report the results but also **interpret** them, that is, describe what your analysis has revealed about the data or methods.

The report should not be longer than 2k words (excluding figure/table captions, references, and comments) with 10% tolerance.