VRIJE UNIVERSITEIT AMSTERDAM

Assignment
Machine Learning for the Quantified Self
Collecting your own data and developing a model

Mark Hoogendoorn

Intermediate deadline 1: 08/06/2025 23:59
Intermediate deadline 2: 15/06/2025 23:59
Final deadline: 22/06/2025, 23:59

1 THE ASSIGNMENT

In the practical assignment of the course you will focus on collecting your own dataset and developing machine learning models for that dataset following the approaches that will be explained during the lectures. Hereby, these models will be based on a more classical feature engineering approach as well as on a deep learning approach. There are a couple of criteria for selection of your data:

1. it needs to cover sensory data, sampled at a relative high frequency.
2. it should be a dataset where the target values (e.g. labels) are available for more or less each time point you collect data.
3. it can be a dataset covering data of a single user, but also of multiple users (if you want, you can make it a joint effort between teams).

Before starting to collect your data, make an overview of the data you plan to collect (what sensors, what frequency, what time period, etc.) and also what target you want to set for your

machine learning problem (e.g. predict the activity). Also include your approach to evaluate the performance of your algorithm. Hereby select an evaluation approach which measures the generalization well (e.g. on another person, on a new recording of the same person, but not randomly sampling timepoints from a single dataset as train and testset). Discuss this with your TA during one of the practical sessions in the first week.

Once discussed and approved, start collecting the data (e.g. using apps such as phybox/phyphox) and follow the full machine learning for the quantified self cycle as we have seen in the book except for the reinforcement learning techniques. On top, you are asked to apply state-of-the-art deep learning techniques. More specifically, we want to ask you to do and report the following:

1. Define a clear research question and the data (what measurements do you want to use, what do you want to predict) you will use to answer the question.

2. Collect the data and describe a summary of the data you have collected (an exploratory data analysis), similar to Chapter 2 of the book. Discuss choices you make (e.g. resolution with which you process your data).

3. Remove noise and handle missing values using an appropriate technique (i.e. those discussed in Chapter 3 of the book). Again, discuss the choices you make and provide a good rationale.

4. Engineer features (cf. Chapter 4), again describe your choices for setting, and analyse the usefulness of the resulting set of features.

5. Define an appropriate train/test setup and apply classical machine learning techniques (cf. Chapter 7) on the resulting dataset from (4). Describe your rationale, the results, and how you optimized the hyperparameters.

6. Apply a deep learning approach which embeds the temporal dimension that was not explicitly discussed in Chapter 8 (e.g. an LSTM, TCN, or some other state-of-the-art approach) in the same setting you have used in (5). Provide a rationale on the choice of your algorithm, hyperparameter settings, discuss your results, and compare your results to those found under (5).

7. A general conclusion and critical reflection of your findings.

Write everything down clearly and usage of references to support your choices is highly encouraged. Draw inspiration from how things on these steps have been described in the book.

As this course is intensive and we want to give you regular feedback on your progress, two intermediate deadlines are set in which you are required to submit part of your report. These will only be graded with "pass" or "fail" and the main focus is on the feedback provided to you. In order to pass the practical you need to have passed all intermediate deadlines and have a sufficient grade for the final report and the exam.

- Deadline 1 (end of week 1, pass/fail only): Points 1-3 listed in the overview. Describe these in max. 5 pages.
- Deadline 2 (end of week 2, pass/fail only): Points 4-5 listed in the overview. Describe these in max. 5 pages.
- Final deadline (end of week 3, final grade provided that you passed the previous 2): Points 1-7. Your report has a maximum length of 14 pages, excluding references, but including everything else.

2 CODING

You can use source code that has been developed for the course as a basis for your assignment if you like. It can be downloaded from GitHub

`https://github.com/mhoogen/ML4QS/`

Note that we will use the Python 3 code (not the R or Python 2 version). You can find all installation instructions in that directory of the GitHub. If you want to use the code in combination with data you generate yourself, please use the format used in the Tables posted on page 17 of the book for your data, otherwise you will need to make changes to the code (which should not be too difficult). **Warning**: when you use Excel with the time stamps in milliseconds since the start of UNIX time, Excel will round these numbers, which you do not want.

Running the algorithms can be time consuming, be aware of this and start early. You can change the step size to reduce the computation time required if you think you will run out of time. Furthermore, you can also work with a subset of your data (e.g. the first half hour or full hour).

3 SUBMISSION REQUIREMENTS

You should work in groups of three students. You should write a report describing the steps you have performed and the results you obtained. The report should follow the Springer Lecture Notes in Computer Science template (you can easily find it if you Google it, there are both LateX and Word templates) and should be at most 14 pages. You can submit via Canvas.

4 PRESENTATION

For the final lecture (which is mandatory for all) a number of groups that have done outstanding work will be asked to present their work. For those groups, discuss the setting you have selected and present highlights of the results.

5 GRADING

The following criteria are used for grading:

- data collection and dataset description (5%)
- techniques selected (25%)
- quality of evaluation (20%)
- rationale provided (25%)
- overall writing style (15%)
- creativity (10%)