Fundamentals of Machine Learning: Assignment 1 brief

Description

You are applying for a job with a data analytics company. As part of the recruitment process, they are asking you to use a Multi-Layer Perceptron to model some data. They are providing you with one dataset (see instructions below regarding downloading the data). The dataset contains $1,000$ samples. The input space is 5-dimensional (first 5 columns of the data). The output (last column) is 1D and continuous-valued. You are obviously very keen to know something about the data (so that you can use prior knowledge!) but, sadly, they are being extremely secretive about it. Instead, they are asking you to produce code which they will use to test how well your model does on unseen data. They are also asking you to provide a brief (max 2 pages, see instructions below regarding submission) report describing/justifying your workflow.

Accessing the dataset

Each of you will have a different dataset identified by your **candidate number**. To access your dataset please use the following URL: http://users.sussex.ac.uk/%7Elb203/teaching/FML/dataxxxxxx.csv where xxxxxx is your candidate number. All datasets have been pre-generated using the list of candidate numbers provided by the school office. If the URL returns an error, please make sure you are using your **candidate number** (usually 5 or 6 digits) and not your registration number (usually 8 digits). If you still get an error, please, email me the URL you are using.

Important notes

1. Since being able to implement a MLP is not a learning outcome of this module, you are **not** expected to write your own code for it and can make use of any toolbox you wish (you should properly acknowledge it obviously). However, should you feel confident enough to write your own code, this will be reflected in your mark (see Section Marking criteria).

2. You are free to decide which programming language to use, with the following caveats:

   - If your language is interpreted (e.g., Matlab, Python, Julia), please make sure to include all scripts necessary to execute your code. Clearly identify which file I am supposed to use in order to get predictions for my unseen data (see below regarding the specification of the unseen data).
   - If your chosen language involves compilation (e.g., Java, C/C++) your submission must include source code **as well as** detailed instructions for compilation. Marks will be deducted if I am unable to compile and execute your code within minutes. For this reason, I recommend that you **also** include an executable (which should run on the University machines) so that I can test the code on unseen data (see below regarding the specification of the unseen data).

3. The unseen data (which won't be made available to you) will be stored in a file called `testdata.csv`. This file will have 5 columns (the dimensionality of the input as indicated above). It will have an unspecified number of rows so please do not expect it to be $1,000$. Concretely, this means your code should be written in such a way that it can read in the content of a file named `testdata.csv` (comma separated values) and produce your predictions (as many as there are data points in the test data!). Please note: (1) I should not have to edit any code to test the unseen data; (2) Generating predictions should not involve any (re)training. Getting access to the testing dataset should not have any impact on how your model operates. This means that you will need to save whatever parameters/quantities are needed to load and run your **best model**. To help you ensure your code is submission-ready, you can download a dummy testdata.csv at the following URL: http://users.sussex.ac.uk/%7Elb203/teaching/FML/testdata.csv. Your code should return 5 numbers.

Submission format

Submission is through the e-submission system on Study Direct. Please submit a single file (.zip, .tgz or any other standard compressed format). This should contain:

- source code used to produce your best model.
- source code and any data/file needed to deploy your best model on the unseen data.
- PDF file documenting your approach to the problem. This document is not about describing the detail of the methods. It is about describing and justifying your choices and how you came to your best model. This document should not exceed 2 pages in length (with reasonable margins and fonts, i.e., fonts not smaller than 10pt, margins not less than 2cm). The information that will be expected should include:
  - Any pre/post-processing techniques you may have used along and why. If you decided no pre/post processing was needed, say so and explain why.
  - What were the free parameters and how you went about choosing / validating them.
  - Any creative optimisation of the process you may have deployed (e.g., comparing different algorithms for gradient descent and/or error criterion). However, please remember that it is **not** acceptable to use a learning model other than a Multi-Layer Perceptron.
  - A brief critical evaluation of your work. In particular, you must indicate how confident you are about the predictive power of your model and on what basis you make your assessment. In what ways could you have improved the model?

Marking criteria

NB: Since writing your own code for the Multi-Layer Perceptron is **not** a requirement, doing so will be considered an *extension* and will only attract a nominal bonus of 10% (with the overall mark capped at 100%). Please be strategic about it. If writing your own code impinges on your ability to deliver on what is being assessed (including good predictions), you are much better off sticking with toolboxes.

- 70%-100% **Excellent**: Accurate predictions (generalisation) underpinned by an excellent workflow, code appropriately commented, evidence of critical assessment of the model. The work shows very good understanding supported by evidence that you have extrapolated from what was taught, through extra study or creative thought.
- 60%-69% **Good**: Reasonable predictions (generalisation) underpinned by a solid workflow, code appropriately commented, some attempt to critically assess the model. The work will be very competent in all respects. Work will evidence substantially correct and complete knowledge, typically not going beyond what was taught.
- 50%-59% **Satisfactory**: Reasonable predictions (generalisation) underpinned by a reasonable workflow, little or no critical assessment. The work will be competent in most respects but there may be minor conceptual errors or methodological oversights.
- 40%-49% **Borderline**: Average predictions (generalisation) underpinned by the most basic workflow, no critical assessment of the work. The work will show the most basic understanding of fundamental concepts acceptable at this level.
- 30%-39% **Fail**: Poor predictions underpinned by inadequate workflow, flawed implementation, no evidence of critical assessment of the model. The work will show inadequate knowledge of the subject, is seriously flawed and displaying major lack of understanding.
- 0%-30% **Unacceptable or not submitted**: Work is either not submitted or, if submitted, very seriously flawed.