Imperial College London
Department of Mathematics
MSc in Mathematics and Finance
Academic year 2022–2023, Autumn term

MATH70116 Deep Learning

Coursework (weight: 10%), 23 November 2022

General rules:

- ⋆ This coursework is to be completed in groups (2-3 students; recommended) or individually.
- ⋆ Present your analysis in a short written report (at most 7 pages). Include your code in an appendix (which may extend beyond 7 pages). Alternatively, the report can be a Jupyter notebook, in that case it should be submitted both as a PDF and as an `.ipynb` file.
- ⋆ You may use any *publicly available* (free or commercial) software and packages. Please indicate in your report which software (and packages) you have used.
- ⋆ There are two deliverables: the **report/notebook** (as discussed above) and a **set of predictions** (details given below). Please send both your report/notebook and your set of predictions by email to `l.gonon@imperial.ac.uk`. Please send one e-mail per group and mention clearly who is part of your group.
- ⋆ **Deadline: Monday, 12 December 2022, 4:00pm UK time.**

In this coursework you will use deep learning to predict high-frequency price changes of a US stock, the identity of which will remain unknown for now. The compressed file `DL-2022-CW-data.zip` contains two CSV files `Data_A.csv` and `Data_B_nolabels.csv`.

Firstly, `Data_A.csv` contains a $100\,000 \times 22$ array with the following information:

- ⋆ Column 1: the **label** — midprice change direction (we define midprice = $\frac{\text{bid price}+\text{ask price}}{2}$) coded as follows: `0` down, `1` up.
- ⋆ Columns 2–22: the **features**, all recorded just prior to the midprice change corresponding to the label.

- Column 2: Sell side, limit order book level 1, Price (in US dollars multiplied by 10000), that is, the **ask price**.
- Column 3: Sell side, limit order book level 1, Volume (in number of shares).
- Column 4: Buy side, limit order book level 1, Price, that is, the **bid price**.
- Column 5: Buy side, limit order book level 1, Volume.
- Column 6: Sell side, limit order book level 2, Price.
- Column 7: Sell side, limit order book level 2, Volume.
- Column 8: Buy side, limit order book level 2, Price.
- Column 9: Buy side, limit order book level 2, Volume.
- Column 10: Sell side, limit order book level 3, Price.
- Column 11: Sell side, limit order book level 3, Volume.
- Column 12: Buy side, limit order book level 3, Price.
- Column 13: Buy side, limit order book level 3, Volume.
- Column 14: Sell side, limit order book level 4, Price.
- Column 15: Sell side, limit order book level 4, Volume.
- Column 16: Buy side, limit order book level 4, Price.
- Column 17: Buy side, limit order book level 4, Volume.
- Columns 18–22: five previous midprice change directions (`0/1`-coded like the labels).

The rows of this file have been *randomly drawn* from a larger data set covering the period 1 August – 27 October 2022, and they can be treated as 100000 independent samples. No *time series structure* can be recovered from the data.

Secondly, `Data_B_nolabels.csv` contains a 10000×21 array with further 10000 samples (drawn similarly as those in `Data_A.csv`) but with *labels omitted*.

In the coursework you are asked to do the following:

(A) Build and train a *binary classifier* that predicts the label in the first column of `Data_A.csv`. Style is free, but your approach should use *neural networks* in a meaningful way. **[5 marks]**

(B) Use the binary classifier created in part **(A)** to predict the labels missing from `Data_B_nolabels.csv`. That is, you are asked to produce 10000 predictions of the form `0/1`. **[5 marks]**

Your solution to part **(B)** (**set of predictions**) should be a text file with 10 000 rows containing `0`s and `1`s. Name this file as "[your CID]_[your surname]`.txt`". For example, a fictionary person *John Smith* with CID *00123456* should name his file as "`00123456_Smith.txt`". **Please adhere to this format carefully, as your solutions will be processed automatically.** If you work in a group, please submit such a file for each group member (i.e., if you work in a group of three you will submit three times the same file, but each of them named differently).

Your solution to **(B)** will be marked based on *accuracy*, defined as

$$\frac{\text{Number of correctly predicted labels}}{\text{Total number of labels}}.$$

Hints: It is a good idea not to use the entire data set in `Data_A.csv` to do the training, but instead split it into training and validation sets. It is also advisable to centre and scale the features — if you do so, remember however to make the same adjustments when predicting labels in part **(B)**.