COMP6560

Computational Intelligence in Business, Economics and Finance Assignment

Predicting Price Movements with Genetic Algorithms

Deliverables: Zip file with project and other files (input data, implementation, and report). The implementation can be in Java or another language of your preference.

Deadline: **23:55 on Tuesday, November 29 2022 (Week 18)**

Your task is to implement a Genetic Algorithm (GA) to create a predicting algorithm to identify price movements over a specific time window. This will be divided into 3 parts:

- **Modelling**: decide what will be in input values to the GA.
- **Implementation**: decide on the individual representation and the fitness function; and implement a GA that given a set of data, optimizes a forecasting rule.
- **Report**: complete a set of experiments and compile the results.

For this assessment, you should use the PriceData.csv file as the source of prices – available on Moodle. The file has 2 columns:

Price	***Increase in 14 days***
357.54	0
356.44	1
358.10	1
352.03	0

The first column contains the daily closing prices and the second column indicates if the price increased 14 days later (1) or not (0).

Note: Since we are using a time window, some rows will not have an associated value in the "Increase in 14 days" column. These rows should be removed once you finish preparing your data.

Part A: Modelling [20 marks]

For this part, you need to decide on the inputs that will be used for predicting price movements. The input will take form of a technical indicator – for example:

- **SMA_X**: *X*-days simple moving average
- **EMA_X**: *X*-days exponential moving average
- **TBR_X**: *X*-days trade break out
- **VOL_X**: *X*-days volatility
- **MOM_X**: *X*-days momentum

where **X** represents the number of days that is used to calculate the indicator.

Your task is to create 6 different technical indicator inputs using a variation of SMA, EMA, TBR, VOL and MOM. At the end of this part, your data will have the following format:

Price	*SMA_X*	*EMA_X*	*TBR_X*	*VOL_X*	*MOM_X*	*SMA_Y*	*Increase in 14 days*
357.54	362.00	356.91	0.01	0.02	13.24	372.53	0
356.44	363.84	357.78	0.01	0.03	19.32	374.25	1
358.10	365.96	358.82	0.00	0.03	24.28	375.67	1
352.03	368.03	360.20	-0.02	0.04	27.03	376.94	0

After you decide on your input values, each input should be used to generate a an output to indicate whether the price will go up 1 (up) or 0 (down). For example, you could assess whether SMA_X > SMA_Y (where X and Y represent different number of days). When the test is true, you could generate a up value (1); otherwise, a down value (0). For MOM_X, you can create a test with a specific threshold: if MOM_X > 0 then up (1), otherwise down (0).

At the end of this part, your data will look like:

SMA_X > SMA_Y	…	…	…	…	*MOM_X > 0*	*Increase in 14 days*
0					0	0
0					1	1
1					0	1
1					1	0

Note: You must have at least **6 input** (columns) values in additional to the target (last) column.

PART B: Implementation [60 marks]

Now that you have your input data, your task is to implement a Genetic Algorithm (GA) to optimize how the indicator values can be used to predict whether the price will increase in 14 days or not. In other words, you need to define a way to combine the inputs and, depending on the result of this combination, predict whether the "Increase in 14 days" value will be 0 or 1. The input used must not include the value of the "Increase in 14 days" column itself.

For your GA, you need to complete:

- **individual representation and initialisation [10%]**: define a representation; the population should be randomly initialised.

- **fitness function [10%]**: evaluation of candidate solutions. This will consist in matching the value predicted by an individual with the corresponding "Increase in 14 days" value. The more values that are matched, the best the fitness of the individual is.

- **selection method [5%]**: the GA should use tournament selection.

- **genetic operators [10%]**: the GA should use one mutation and one crossover operator of your choice.

- **termination criteria [5%]**: a maximum number of generations should indicate the end of the evolutionary process.

Evaluation [20%]
Improve the evaluation of your GA by dividing your input data into training and testing – e.g., 70% of the data for training, the remaining 30% for testing. You GA should be "trained" first and the best individual found should be evaluated on the test data.

PART C: Report [20 marks]

Prepare a 2-pages report detailing your implementation and results. Topics you should cover:

- How much you managed to achieve in terms of the given tasks.
- Difficulties during your implementation: what has gone well, what has gone wrong.
- Report on your experimental results, including summary statistics from multiple runs. If you haven't obtained any results yet, that is fine, but you should still mention that you don't have any results.
- Report on different GA parameters that you might have used, and how/if they affected the performance of the algorithm. If you've tried different parameters, you should also present statistical analysis to support your argument.
- Anything else you consider useful to mention, e.g., any additional methods implementation you decided to implement.
- Would you do anything differently if you had to re-do the assignment.

Note: The page limit is **2 pages**.

Submission

Electronic version (zip file) of project and other files (input data, implementation, and report). The zip file should be submitted via Moodle before the deadline, no later than **23:55 on Tuesday, November 29 2022 (Week 18)**. Any other form of submission will not be accepted. Note that the submission link will not be available after the deadline.

Reference: Technical Indicators

The **Exponential Moving Average (EMA)** is a type of moving average that puts more emphasis on recent price data, while the standard moving average uses the same weight for all prices observations. The formula to calculate the EMA is given below:

$$EMA(L,t) = \left(P_t * \left(\frac{smoothing}{1+L}\right)\right) + \left(EMA(L,t-1) * \left(1 - \frac{smoothing}{1+L}\right)\right)$$

where P_t is the price at the day t, L is the period length (number of days) and the $smoothing$ factor is equal to 2. In other words, the EMA for the t-th day is the price at t multiplied by the smoothing term plus the EMA value of the previous day (t – 1) multiplied by the smoothing term. To calculate the EMA for the first day after the interval L (e.g., EMA at the 12th day for an 12-day interval), we use the value of the Simple Moving Average (SMA) as the EMA:

$$SMA(L,\ t) = \frac{\sum_{i=1}^{L} P_{t-i}}{L}$$

For example, if we are interested in calculating the 12-day EMA, we will first calculate the 12-day SMA as the EMA on the 12th day; then use this value to calculate the EMA for the 13th day and so forth.

The **trade break out rule (TBR)** is given the formula below.

$$TBR(L,t) = \frac{P_t - \max\{P_{t-1}, \dots, P_{t-L}\}}{\max\{P_{t-1}, \dots, P_{t-L}\}}$$

where P_t is the price at the day t, L is the period length (e.g., 24 days), and max{…} returns the maximum price observed in the days $\{t-1, \dots, t-L\}$ – i.e., the highest price in the 24 past days.

The **volatility (VOL)** is given the formula below.

$$Vol(L,t) = \frac{\sigma(P_{t-1}, \dots, P_{t-L})}{SMA(L,t)}$$

where σ is the standard deviation for the prices in the given range, P_t is the price at the day t and L is the period length (e.g., 29 days).

The **momentum (MOM)** is given the formula below.

$$MOM(x,\ t) = P_t - P_{t-x}$$

where P_t is the price at the day t and P_{t-x} is the price at previous x-th day (e.g., price at the $t-x$ day).

Plagiarism and duplication of material

• Late or non submission of coursework

The penalty for late or non submission of coursework is normally that a mark of zero is awarded for the missing piece of work and the final mark for the module is calculated accordingly.

• Plagiarism and Duplication of Material

Senate has agreed the following definition of plagiarism: "Plagiarism is the act of repeating the ideas or discoveries of another as one's own. To copy sentences, phrases or even striking expressions without acknowledgement in a manner that may deceive the reader as to the source is plagiarism; to paraphrase in a manner that may deceive the reader is likewise plagiarism. Where such copying or close paraphrase has occurred the mere mention of the source in a bibliography will not be deemed sufficient acknowledgement; in each such instance it must be referred specifically to its source. Verbatim quotations must be directly acknowledged either in inverted commas or by indenting." The work you submit must be your own, except where its original author is clearly referenced. We reserve the right to run checks on all submitted work in an effort to identify possible plagiarism, and take disciplinary action against anyone found to have committed plagiarism. When you use other peoples' material, you must clearly indicate the source of the material using the Harvard style
(see http://www.kent.ac.uk/uelt/ai/styleguides.html).

In addition, substantial amounts of verbatim or near verbatim cut-and-paste from web-based sources, course material and other resources will not be considered as evidence of your own understanding of the topics being examined.

The School publishes an on-line Plagiarism and Collaboration Frequently Asked Questions (FAQ) which is available at:
http://www.cs.kent.ac.uk/teaching/student/assessment/plagiarism.local

Work may be submitted to Turnitin for the identification of possible plagiarism. You can find out more about Turnitin at the following page:
http://www.kent.ac.uk/uelt/ai/students/usingturnitinsts.html#whatisTurnitin