Task 1. ML for Finance. Argimiro Arratia. BSE- 2024

Published: 24/04/2024
Hand in: 8/05/2024:: 23:59

- Do not hand in R/Python output, use cut/paste to write a report (LaTeX or Word), and it is not necessary to keep all decimal digits in reported results.
- Put team members names in the report. Teams should be of at most 3 students.
- Submit your report in PDF format plus R/Python code, all in a .zip folder via the Classroom with subject:
  "BSE Task 1 - Team `- members names -`". Only one member of team makes the submission for all.

1. (10%) Show that the EWMA-based variance σ^2_{ewma} can be obtained also from the following recursion: $\sigma^2_{ewma}(t) = \lambda\sigma^2_{ewma}(t-1) + (1-\lambda)r^2_{t-1}$. which is easily derived from the EWMA equation (see Lecture 2).

   This recurrence has the computational advantage of only needing to keep in memory the previous day return r_{t-1} and the previous estimate of σ^2_{ewma}. In fact, the function EMA from the R package TTR implements this recursion[1]. Use EMA to compute $\sigma^2_{ewma}(t)$ with $\lambda = 0.94$ for some market index (e.g take it from Prob. 2). Obtain the volatility time series estimation of the market index from this EMA estimation of variance and compare it to a regular (historical) volatility estimation (e.g. by cumulative sum of square returns or Parkinson, or Garman-Klass estimates). Report what you observe.

2. (35%) Consider the dataset WorldMarkts99_20.RDS containing price history from 1999-01-01 to 2020-04-30 of 11 market indices worldwide plus VLIC and VIX. The script `HW1_markets.R` will help you retrieve data (optionally) and organise it for the rest of this exercise. For the epoch assigned to your team in the Google spreadsheet, the corresponding team must do: A full causality analysis for the first four lags of the **returns** time series of these 11 market indices, and a full causality analysis for the first four lags of the volatilities series of these market return indices (this is known as *volatility spill-over*). Do this analysis sampling the series first **weekly** and then **monthly** periods. Estimate volatility using

[1] Read the Help on EMA: The function EMA in TTR package is defined as the recursion $EMA(t) = K \cdot input + (1-K) \cdot EMA(t-1)$, where $K = 2/(n+1)$ and n =number of periods to average over. Note terms are different than equation given in class so you must adapt accordingly

EMA (see Prob. 1). Tabulate **and comment** your results (four tables, one for returns, other for volatility and each for weekly and for monthly sampling), so that *cause* $\rightarrow$ *effect* goes from row to column, and each entry a 4-vector of $\{0,1\}$ indicating causality (1) or not (0) for each lag (coordinate). Example

	India	Brazil	UK	...
USA	$(1,0,1,1)$	$(0,1,1,0)$	$(1,1,1,0)$	...
Brazil	$(0,0,1,1)$		$(1,0,0,0)$	...
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

so, first entry means $US \rightarrow India$ at lags 1,3,4.

Causality should be considered significant at the 5% level. Comment your results. Among which countries there is true causality, or contemporaneous correlation? Can you device a network of causalities?

3. (35%) **Neural Networks and Gaussian process**. Predict the SP500 with the financial indicators assigned to your team in the google spreadsheet (ep, dp, de, dy, dfy, bm, svar, ntis, infl, tbl , see RLab3_2_GWcausalSP500.R), some lagged series of these indicators and lags of the target using a Neural Network and a GP regression with your desired kernel. Predict return, or price, or direction (up or down). For which target works best? Do some feature selection to disregard some variables, select appropriate lags: causality, (distance) correlation, VAR-test, Lasso ... (The script RLab5_GausProc.R can be of help. The dataset is `goyal-welch2022Monthly.csv` and work within the period 1927/2021.)

4. (20%) This is for doing the extra (and necessary) work for good data analysis in problems 2 and 3:

   - Do proper preprocessing and analysis of data: check for missing values, impute values if necessary, outliers, corrupted data, etc.

   - To be more rigorous in our causality analysis a rolling windows methodology should be applied (e.g. take a window of width a year and roll it weekly, monthly).

   - Do feature selection (as proposed at end of prob. 3)

   - Tune parameters of Nnet (size of hidden layer, decay), and GP (kernels)

   - Compare against ARMA(p,q) fit (as baseline model

   - Can the stationary bootstrap be of help in either of these problems to strengthen the conclusions? Apply it if you think it helps.

5. Write each authors contribution. E.g. AA: worked the proof in Prob 1,BB wrote the report; AA & BB clean data for prob 2, 3, BB wrote code and set experiments, AA made analysis and wrote report; or "All authors contributed equally".