Task 1. ML for Finance. Argimiro Arratia. BGSE- 2022

Published: 25/04/2021
Hand in: 10/05/2021:: 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 2 students.
- Submit your report in PDF format plus R/Python code, all in a .zip folder via the Classroom with subject:
  "BGSE Task 1 - Team - `members names` -".

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. Use EMA to compute $\sigma^2_{ewma}(t)$ with $\lambda = 0.94$ for some market index (e.g take it from Prob. 3). Obtain standard deviation of the market index from this EMA estimation of variance and compare to regular std.

2. (15%) Give a proof of the fact that the best estimator for X_{t+h} with information set $Z = (X_t, X_{t-1}, \ldots, X_{t-p})$ is a linear regression on Z, when all variables follow a normal distribution (see Lecture 1).
   (Hint: Recall that X and Z have joint Gaussian distribution, this means

   $$\begin{bmatrix} X \\ Z \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_X \\ \mu_Z \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XZ} \\ \Sigma_{ZX} & \Sigma_{ZZ} \end{bmatrix}\right)$$

   from this obtain the conditional distribution of X given Z.)

3. (25%) 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 `HW2_markets.R` will help you retrieve data (optionally) and organise it for the rest of this exercise. For the epoch selected by 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 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

4. (10%) For each of the four kernels below, and the suggested values of their parameters, sample and plot a set of 5 different latent functions f from the Gaussian process prior: $f \sim N(0, K)$, where K is the covariance matrix. The kernels to try out are:

   Square exponential SE:

$$k_{se}(\boldsymbol{x}, \boldsymbol{x}') = h^2 \exp\left(\frac{-(\boldsymbol{x}-\boldsymbol{x}')^2}{\lambda^2}\right), \quad \text{for } h = 1;\ \lambda = 0.1, 1, 10$$

   Rational quadratic RQ:

$$k_{rq}(\boldsymbol{x}, \boldsymbol{x}') = h^2 \left(1 + \frac{(\boldsymbol{x}-\boldsymbol{x}')^2}{\alpha\lambda^2}\right)^{-\alpha}, \quad \text{for } h = 1;\ \lambda = 0.1, 1, 10; \alpha = 1, 5, 10;$$

   Observe (show): when $\alpha \rightarrow \infty$ RQ kernel reduces to the SE kernel with length scale λ

$$k_3(\boldsymbol{x}, \boldsymbol{x}') = 2\exp\left(\frac{-\sin(\pi(\boldsymbol{x}-\boldsymbol{x}')/3)^2}{2\lambda^2}\right), \text{ for } \lambda = 0.1, 1.5, 5$$

$$k_4(\boldsymbol{x}, \boldsymbol{x}') = 2\exp\left(\frac{-(\boldsymbol{x}-\boldsymbol{x}')^2}{2\lambda^2}\right) + 1.5\boldsymbol{x}\boldsymbol{x}', \text{ for } \lambda = 0.1, 1.5, 5$$

   What do you observe? Comment on the plots.

5. (30%) **Gaussian process**. Predict the SP500 with the financial indicators selected by your team in the google spreadsheet (ep, dp, de, dy, dfy, bm, svar, ntis, infl, tbl , see RLab3_GWcausalSP500.R), some lagged series of these indicators and lags of the target using a GP regression with your desired kernel. Predict return, or price, or trend (for which target works best?) select appropriate kernel and justify its use.

   Do some feature selection to disregard some variables, select appropriate lags: causality,

(distance) correlation, VaR-test, Lasso ... (The script RLab3_GPlab.R can be of help. The dataset is `GoyalMonthly2005.csv` and work within the period 1927/2005.)

6. (10%) **GP and AR**(1). Consider the Ornstein-Uhlenbeck process $U(t)$, which is a GP with kernel $k(\boldsymbol{x}_i, \boldsymbol{x}_j) = k(|i-j|) = \dfrac{\sigma^2}{2\gamma}\exp(-\gamma|i-j|)$ (see Lec. 3); that is, for $t \in (0, +\infty)$,

$$U(t) = \sigma e^{-\gamma t}\int_0^t e^{\gamma s} dB(s)$$

and sample this process at equally spaced times: $\{i\tau : i = 0, \ldots, n\}$, $\tau > 0$.

Show that the series $U_i = U(i\tau)$ obeys an $AR(1)$ model. (Hint: Develop the expression $U_{i+1} = \sigma\displaystyle\int_0^{(i+1)\tau} e^{-\gamma((i+1)\tau - s)} dB(s)$)