Assignment II

Time series forecasting

Loosely following arxiv.org/abs/1703.04385

Grade 4

- Load in the S&P 500 companies stock prices & index using `yfinance` (see Jupyter notebook for first steps)
- Reduce dimensionality of the 500 time series using PCA
  - Change to log daily returns instead of closing time prices
  - Apply augmented Dickey-Fuller Test to test if non-stationarity is removed. Plot the distribution of p-values across your 500 time series in a histogram & discuss it.
  - Reduce dimensionality to retain 99% of the explained variance

Grade 3

Write a forecast model to predict future index data of the S&P 500 stock market index (the actual index and not log daily returns).

- Use the PCA reduced log daily returns as additional time series when modelling the S&P 500 stock market index. Consider that S&P 500 stock market index not only depends on its own past values but also on the past values of the other series. (See e.g., https://joaquinamatrodrigo.github.io/skforecast/0.7.0/user_guides/multivariate-forecasting.html)
- Train & predict the model and measure its performance
  - Plot predicted vs true values
  - Backtest the model
- Tune hyper parameters of the model (briefly outline the procedure)

Grade 2

Analyse the persistent homology — “shape of data” — of the 500 stock prices with `scikit-tda`.

- Use PCA reduced log daily return (top k components that explain 99% of the variation) as input.
- Take a window size of 50 days* and create a data set in the following way:
  - Interpret k time series as a data set with 50 data points and k dimensions.
- Study topological features of this data set data using persistent homology
  - Plot persistence diagram** (PD) using `Ripser` & interpret it
  - Discuss (significant) number of connected components, 1D or "circular" holes, and 2D “voids”

* The windows could be [1, 50] or [2 to 51] or [3 to 52], …

** Limit order of homology to 2 (Betti number ≤ 2): `maxdim=2`

Grade 1

- Use adjacent 20-day periods* (similarly to Sect. 4 in arxiv.org/abs/1703.04385) in the dataset and compute the following:
  - Compute PDs for all time windows and:
    - Determine the persistent entropy of individual PDs using `persim`.
    - Measure difference of PDs between adjacent time windows* using the Wasserstein distance with `persim`.
  - Time shift the entropy and Wasserstein difference curves accordingly (since they looks 20 days into future). Plot persistent entropy and Wasserstein difference curves across time and contrast it to S&P 500 stock index values. Discuss the correlations.

* [1, 20] and [2 to 21], [2 to 21] and [3 to 22], …

Grade 1

- Take the previous forecasting model and add the persistent entropy and Wasserstein distance time series as additional variables*.
- Does the predictive power of the model change?
  - Use the same validation procedure and plot the forecasted vs. actual stock index trend.
  - Discuss/interpret your findings.
    - Does financial data have shape & do shape changes have predictive power?

* Be aware of the time shift, both time series look 20 days into the future. You need to remove the first N days in your log returns and stock index data, where N is the window size of the Wasserstein differences.