Encoder Decoder LSTM 구현 및 실험 결과 보고

Tae-Won Lee,
leetee95@gmail.com
2020. 05. 10.

Introduction

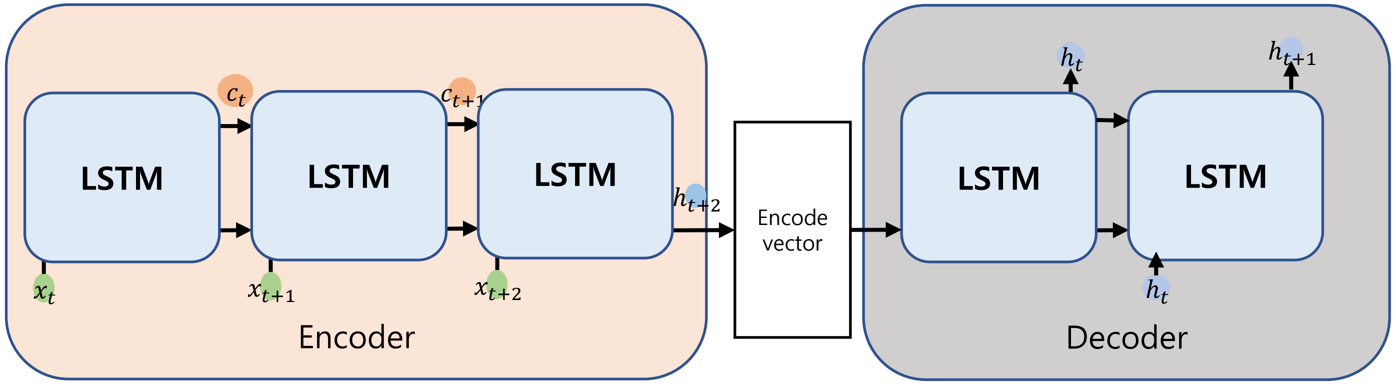
- 기존의 LSTM등 RNNs의 모형에서 input length와 output length가 다른 데이터를 다룰 수 없는 문제점이 있음.

- 이를 해결하기 위해 Encoder-Decoder 구조를 LSTM에 적용함.

- input length와 output length가 다른 길이가 되어도 출력이 가능하게 됨.

Related works

- Encoder-Decoder LSTM



  - 두개의 LSTM 아키텍쳐가 결합된 구조.

  - Encoder아키텍쳐에서 셀의 마지막 시점의 hidden state를 Decoder로 넘겨줌

  - 받은 hidden state를 Decoder에서 최초의 hidden state로 사용함.

Related works

- Stock data Normalization

학회 명	논문 명	발간 년도	인용 수	normalization 방식	normalization 범위	실제 내용
Expert Systems With Applications	ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module	2018	80	min-max normalization	언급 x	where X is original value at a data point, X max and X min are maximum and minimum value of a specific feature, respectively. Consequently, every feature change could be captured in the range of [0,1].
Expert Systems with Applications	Stock price prediction using deep learning and frequency decomposition	2020	1	min-max normalization	언급 x	After decomposition of indices into various frequency spectra, it is required to convert the original data to normal data before training. Normalization leads to the adjustment of the data noise effect and implementation of neural networks with high efficiency and speed (Makridakis, 1993; Diehl, et al., 2015). To this end, we use the following equation.
IEEE ACCESS	Multivariate Financial Time-Series Prediction With Certified Robustness	2020	2	min-max normalization	언급 x	To detect the agricultural commodity futures price pattern, it is necessary to normalize the agricultural commodity futures price data. Since the LSTM neural network requires the agricultural commodity futures patterns during training, we use the ``min-max'' normalization method to reform dataset, which keeps the pattern of the data, as follow:
Quantitative Bio-Science	Forecasting the KOSPI 200 Stock Index Based on LSTM Autoencoder	2020	0	batch normalization	언급 x	batch normalization was added to the model, and early stopping in Keras's callback functions was utilized. For the loss function and the optimization function, both the autoencoder and the predictive model used MSE (Mean-Squared-Error) and Adam Optimizer.
International Journal of Machine Learning and Cybernetics	Study on the prediction of stock price based on the associated network model of LSTM	2020	20	min-max normalization	언급 x	Due to the different measurement unit of different stock index data, for avoiding the impact of different measurement unit, all the attribute data are normalized to fall within a same range. In this paper, the min–max normalization method is used. The normalization function is shown in Eq. 11

Proposed method

- Stock data Normalization

학회 명	논문 명	발간 년도	인용 수	normalization 방식	normalization 범위	실제 내용
IEEE ACCESS	Feature Engineering for Mid-Price Prediction With Deep Learning	2019	18	min-max normalization	전체 데이터에 적용	The next step of the pre-processing step is data normalization. We perform a filtering and a normalization method during the feature extraction process and training. The first one is a statistical filtering method , while the second one is based on MinMax. More specifically, we perform the filtering methodology first and apply it directly on the raw MB data. The main idea of the methodology is to identify and eliminate any observation that does not reflect market activity.
Neural Computing and Applications	Stock price prediction based on deep neural networks	2019	43	min-max normalization	언급 x	To facilitate better observation and comparison of the experimental results, PSR, de-noising and normalization were first performed in the model parameter experiment.
Proceedings of Machine Learning Research	Stock Price Prediction Using Attention-based Multi-Input LSTM	2018	33	min-max normalization	언급 x	Notice that the MSE we calculated are derived from normalized data. That's because there exists huge value gap among different stocks. If we use original stock price to evaluate error, the error of high price stocks would probably be much more larger than low price ones, which implies models perform better on high price stocks would very likely to have better overall performance. Thus the performance on low price stocks would become dispensable. To avoid the bias caused by the aforementioned problem we evaluate the error with normalized stock price ranged from -1 to 1.
IEEE	Hybrid Deep Learning Model for Stock Price Prediction	2018	18	min-max normalization	전체 데이터에 적용	we present the normalized training data form 1950 to 2002. The curve shows the randomness of stock price data. Our main goal is to predict the closing price of next day. We put all the training data into a one dimensional vector and pass it to the network with a fixed learning rate(.001). Figure 6 shows the normalized test dataset representation, which is also showing the randomness of stock price.

Proposed method

- Stock data Normalization

학회 명	논문 명	발간 년도	인용 수	normalization 방식	normalization 범위	실제 내용
Physica A: Statistical Mechanics and its Applications	Financial time series forecasting model based on CEEMDAN and LSTM	2018	111	min-max normalization	전체 데이터에 적용	Where minx(t) and maxx(t) are the minimum and maximum value of the whole time series respectively.
International Conference on Advances in Computing, Communications and Informatics (ICACCI)	STOCK PRICE PREDICTION USING LSTM,RNN AND CNN-SLIDING WINDOW MODEL	2017	235	min-max normalization	언급 x	The data varies with in a range of 2000 to 4000 for Infosys and TCS and for Cipla it is 400 to 700. To unify the data range, it was subjected to normalization and was mapped to a range of 0 to 1. This normalized data was given to the network for training.
IEEE International Conference On Recent Trends in Electronics Information & Communication Technology (RTEICT)	Short Term Stock Price Prediction Using Deep Learning	2017	41	min-max normalization	언급 x	The data of each stock consists of approximately 85000 - 90000 points. The data was normalized into the range of [0, 1] using MinMaxScaler(2), since neural networks are known to be sensitive to unnormalized data, which is then fed to the prediction model.
IEEE International Conference on Big Data	A LSTM-based method for stock returns prediction : A case study of China stock market	2015	286	mean and standard deviation	각각의 seqence vector에 적용	normalization was applied for each vector of a sequence by using the mean and standard deviation of each stock computed from the training set.
International Conference on Business Analytics and Intelligence	Stock Price Prediction Using Machine Learning and Deep Learning Frameworks	2018	19	min-max normalization	전체 데이터에 적용	The raw data is normalized using the min-max normalization
Expert Systems With Applications	ModAugNet: A new forecasting framework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM module	2018	80	min-max normalization	전체 데이터에 적용	X max and X min are max- imum and minimum value of a specific feature,

Experimental setting

- Dataset

stocks	N	Train	val	test
KOSPI	5156	3210	979	967
KOSDAQ(2013-03-03 ~)	1898	938	474	486
NASDAQ	5265	3263	1002	1000
S&P 500	5265	3263	1002	1000
다우존스	5265	3263	1002	1000
홍콩 항생	5151	3196	976	979
일본 니케이	5124	3182	971	971
독일 DAX	5306	3300	1005	1001
프랑스 CAC	5343	3313	1015	1015
스페인 IBEX	5314	3283	1017	1014
대만 기권	5144	3206	971	967
네덜란드 AEX	5347	3316	1016	1015
인도 선섹스	5151	3204	970	977
브라질 보베스파	5265	3263	1002	1000
Gold price	5050	3067	994	989
Oil price(2000-08-23 ~)	5060	3076	994	990
10년 만기 미국 국채	3264	3264	1027	688
10년 만기 한국 국채				
비트코인 암호화폐(2014-09-17 ~)	2275	1560	359	356
이더리움 암호화폐(2015-08-07 ~)	1951	1236	359	356

KOSPI INDEX log difference

2000.01.01 – 2020.12.31

Experimental setting

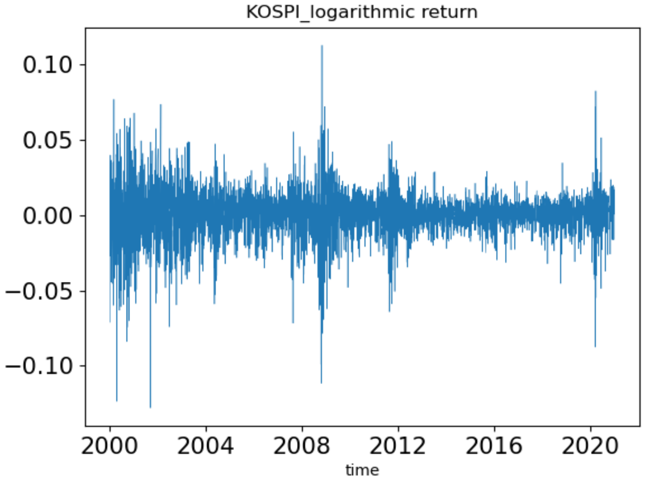
- Statistics

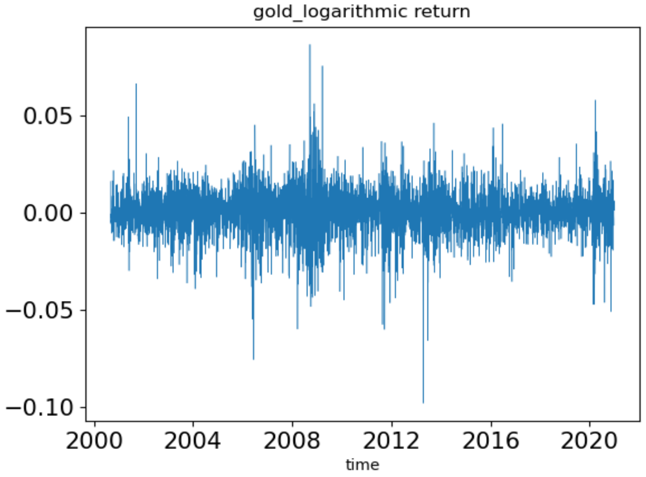
stocks	Mean	Standard deviation	Skewness	Kurtosis(3)	Jerque Bera	ADF
KOSPI	1582.518	587.343	-0.421	1.861	432.61***	-0.438834
KOSDAQ	668.623	103.98	0.46	2.632	78.4663***	-1.173884
NASDAQ	3791.769	2367.556	1.388	4.395	2125.9183***	3.346301
S&P 500	1653.635	674.270	1.033	3.067	941.5057***	1.678640
다우존스	14655.366	5792.134	1.043	2.906	959.6158***	0.870942
홍콩 항생	19919.806	5726.487	-0.18	2.076	211.8634***	-1.579436
일본 니케이	14640.566	4694.21	0.393	2.009	342.692***	-0.491717
독일 DAX	7661.465	3006.491	0.385	1.978	363.2483***	-0.506127
프랑스 CAC	4469.135	918.824	0.153	2.207	161.6942***	-2.233112
스페인 IBEX	9680.838	2017.084	0.667	3.565	466.6305***	-2.262163
대만 기권	7903.82	2055.782	0.243	2.707	69.2831***	-0.264858
네덜란드 AEX	435.655	108.924	0.265	2.222	197.9852***	-1.899507
인도 선섹스	17921.698	11385.273	0.401	2.120	305.4783***	0.679944
브라질 보베스파	3791.769	2367.556	1.388	4.395	2125.9183***	3.346301
Gold price	997.637	479.258	-0.105	1.796	315.6375***	-0.459757
Oil price	61.785	25.912	0.400	2.381	216.5233***	-1.997886
10년 만기 미국 국채	118.257	9.575	-0.285	2.145	219.8195***	-2.216594
10년 만기 한국 국채						
비트코인 암호화폐	4925.441	4830.747	1.056	4.332	595.9625***	1.440318
이더리움 암호화폐	223.111	231.702	1.601	6.123	1642.5896***	-2.082944
KOSPI INDEX log difference	0.000193	0.0149	-0.572	10.03	10943.0983***	-33.339841***

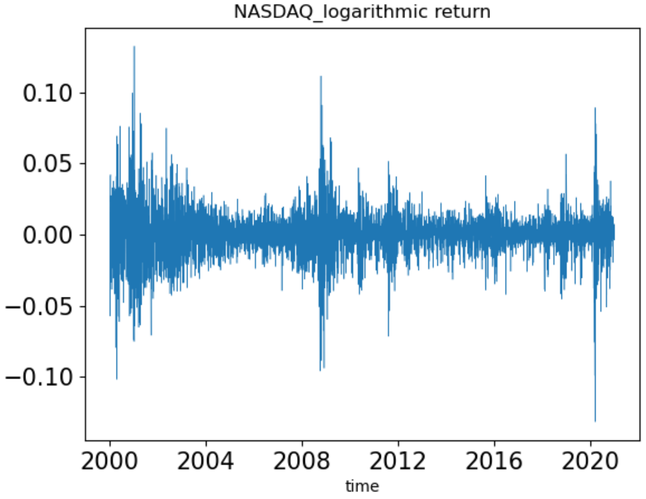
Used Close(종가) variable.

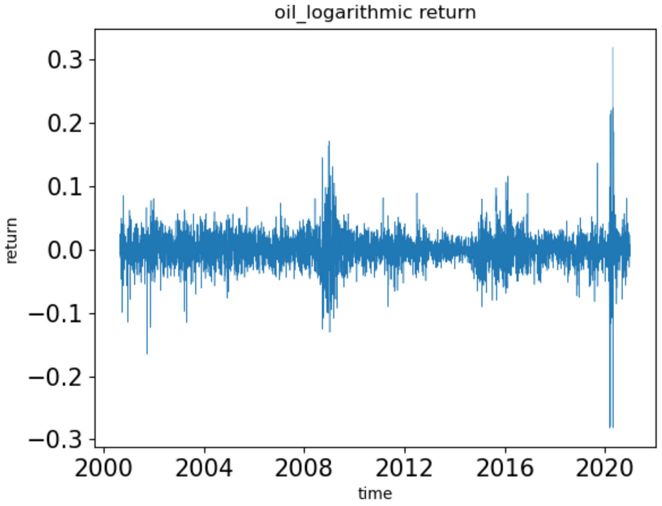
Experimental setting

- logarithmic return : ln((t+1)/t) t : 시점이 t일때 종가(close)
-









- logarithmic return

  - 수익률을 나타내는 지표 중의 하나.

  - 일반 수익률보다 실제 수익률에 더 가까운 지표.

Experimental setting

- Data normalization
  - min-max normalization

- Loss function
  - Used MSE Loss as loss function

- Optimizer
  - Adam

- Metric
  - MSE($e^{predicted\ value}, e^{true\ value}$)
  - MAE($e^{predicted\ value}, e^{true\ value}$)
  - MAPE($e^{predicted\ value}, e^{true\ value}$)

Hyperparameters	**Number**
hidden state, cell state dimension	10
input sequence length	20
output sequence length	1
epoch	30
Batch size	128

Experimental setting

- To determine the detailed performance of a model
  - Add another model evaluation metric.

    - Mean Absolute Percentage Error (MAPE) = $\frac{\sum_{t=1}^{n}\left|\frac{y_t-\hat{y}_t}{y_t}\right|}{n} \cdot 100(\%)$

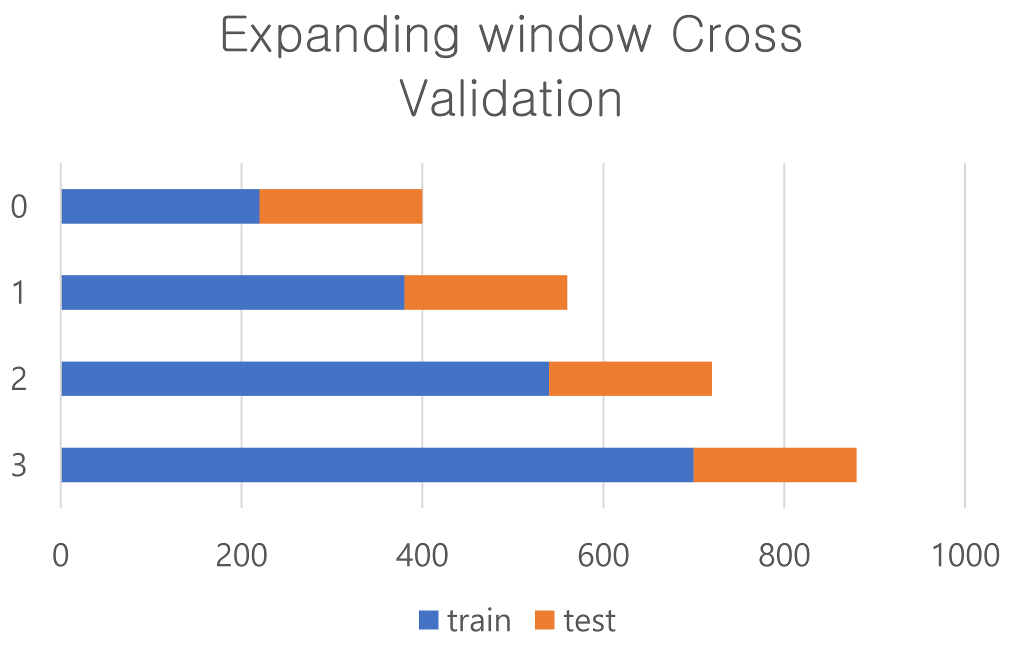
      - MAPE is undefined for data points where the value is zero.
      - MAPE is biased towards prediction that are systemically less than the actual values themselves

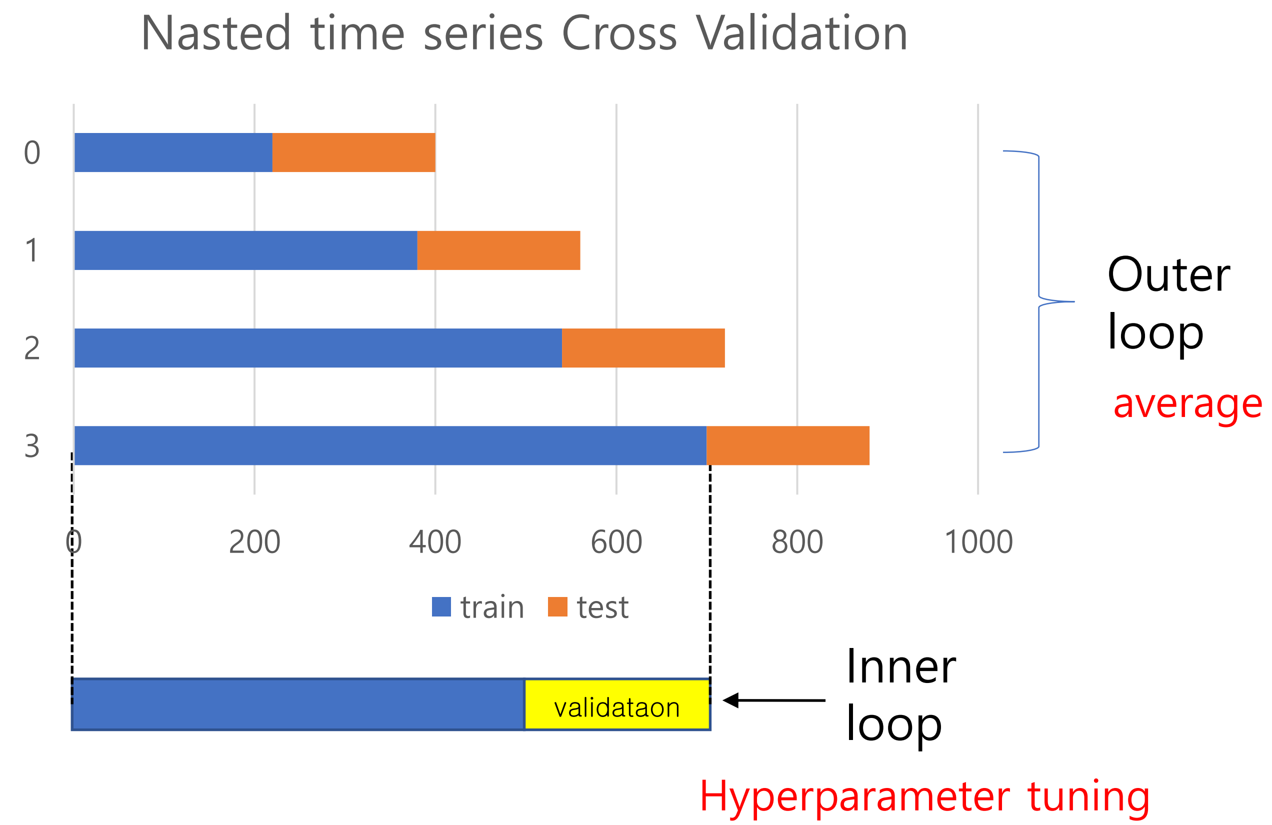
$$y = 20, \hat{y} = 10, n = 1 \ \rightarrow MAPE = 50\%$$
$$y = 10, \hat{y} = 20, n = 1 \ \rightarrow MAPE = 100\%$$

  - Compare three metrics.
    - MAE
    - MSE
    - MAPE

Experimental setting

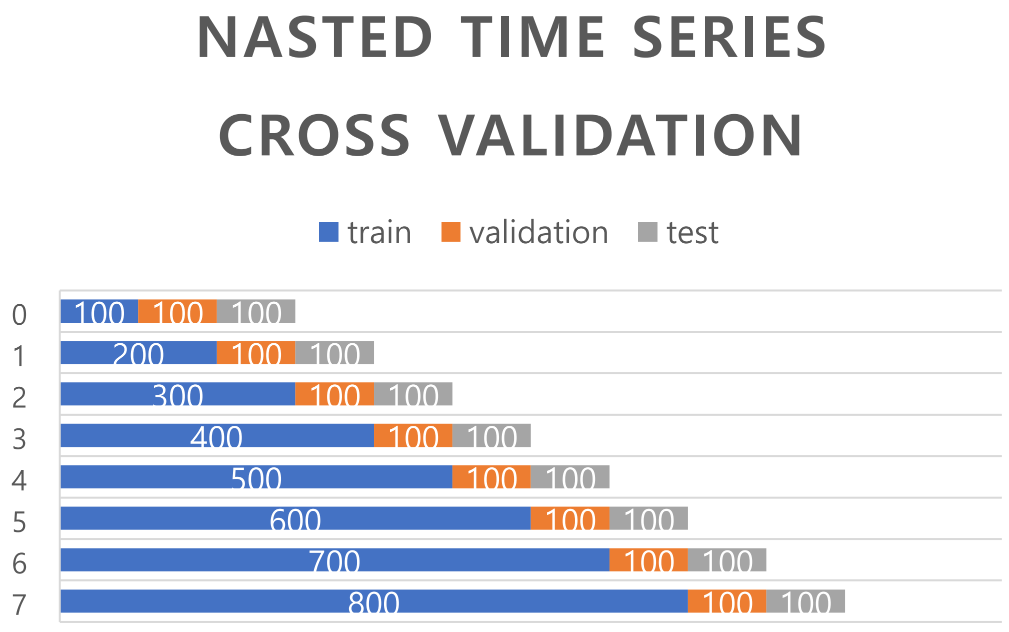
- Stock dataset split in timeseries data





Experimental setting

- Nasted time series Cross Validation



Train : validataion : test = 8 : 1 : 1

- Maximum iteration is 8
- Iteration from 3 to 5 is used universally in timeseries data cross-validtaion.

- More data should be collected.

Experimental setting

- Investment stratege

  Problem determining INPUT length

  - **Value (가치투자)**
  - **Asset allocation (자산배분 전략) – Low correlation, Percentage of assets**
  - **Quality (우량주 투자)**
  - **Momentum (모멘텀 전략)**
  - **Scalping - futures contract**

- Hyperparameter tuning

Experimental setting

- Investment stratege

  - normalization applied to each split sequence

    "normalization was applied for each vector of a sequence by using", A LSTM-based method for stock returns prediction : A case study of China stock market (2015)

  - normalization applied to entire sequence

    "The raw data is normalized using the min-max normalization", Stock Price Prediction Using Machine Learning and Deep Learning Frameworks

>> Both of the strategies can be selected

Comparison results

- Mean Absolute Error (Normalized to each sequence)

Data	Proposed	Standard Deviation	RNN	Standard Deviation	LSTM	Standard Deviation	GRU	Standard Deviation	Enc-Dec LSTM	Standard Deviation
네덜란드 AEX(^AEX)			0.96207	0.21559	0.87075	0.21181	0.94374	0.14956		
인도 선섹스(^BSESN)			0.89591	0.24998	1.00239	0.20139	0.98904	0.14931		
브라질 보베스파(^BVSP)			0.69429	0.19338	0.71985	0.26356	0.87155	0.22545		
다우존스(^DJI)			0.86084	0.25060	0.94760	0.20296	0.91614	0.26985		
프랑스 CAC(^FCHI)			0.93031	0.15425	0.87530	0.28280	0.80879	0.18792		
독일 DAX(^GDAXI)			0.89896	0.18375	0.88227	0.19661	0.94807	0.16574		
S&P 500(^GSPC)			1.00582	0.16804	0.96865	0.21661	1.01276	0.18808		
홍콩 항생(^HSI)			0.76397	0.24637	0.80247	0.22890	0.74720	0.28785		
스페인 IBEX(^IBEX)			0.87502	0.24435	0.81029	0.10312	0.69536	0.23857		
NASDAQ(^IXIC)			1.05666	0.14763	0.99885	0.22230	1.06603	0.17362		
KOSDAQ(^KQ11)			0.88091	0.21350	0.88883	0.20363	0.93733	0.05612		
KOSPI(^KS11)			0.76880	0.23431	0.82688	0.16073	0.72645	0.18427		
일본 니케이(^N225)			0.84537	0.23163	0.95740	0.15007	0.87397	0.19705		
대만 기권(^TWII)			0.88149	0.30870	0.84268	0.22895	0.87681	0.19366		
비트코인 암호화폐(BTC-USD)			0.88364	0.36082	0.79781	0.27053	0.94435	0.18464		
Gold price(GC=F)			0.70107	0.14618	0.73244	0.20841	0.83734	0.18059		
Oil price(CL=F)			0.72740	0.24628	0.86665	0.27800	0.68963	0.36977		
이더리움 암호화폐(ETH-USD)			0.76278	0.30663	0.64972	0.23933	0.75810	0.27552		
10년 만기 미국 국채										
10년 만기 한국 국채										

Comparison results

- Mean Squared Error (Normalized to each sequence)

Data	Proposed	Standard Deviation	RNN	Standard Deviation	LSTM	Standard Deviation	GRU	Standard Deviation	Enc-Dec LSTM	Standard Deviation
네덜란드 AEX(^AEX)			1.22650	0.44776	1.02455	0.40677	1.19246	0.34049		
인도 선섹스(^BSESN)			1.04893	0.53668	1.31499	0.43717	1.29377	0.34607		
브라질 보베스파(^BVSP)			0.73107	0.29510	0.79004	0.44917	1.11190	0.40448		
다우존스(^DJI)			1.01449	0.49659	1.16652	0.46150	1.12879	0.59395		
프랑스 CAC(^FCHI)			1.23348	0.39812	1.08103	0.55011	0.96631	0.37860		
독일 DAX(^GDAXI)			1.10230	0.36892	1.06544	0.41249	1.16645	0.32853		
S&P 500(^GSPC)			1.29184	0.38246	1.17622	0.49875	1.28634	0.42693		
홍콩 항생(^HSI)			0.91521	0.56533	0.89238	0.53174	0.86247	0.49518		
스페인 IBEX(^IBEX)			1.09938	0.57171	0.92026	0.22723	0.76798	0.45721		
NASDAQ(^IXIC)			1.38485	0.36350	1.27328	0.49485	1.39281	0.38546		
KOSDAQ(^KQ11)			1.05666	0.43742	1.05795	0.36366	1.10322	0.11783		
KOSPI(^KS11)			0.85155	0.45105	0.93961	0.33633	0.75416	0.33836		
일본 니케이(^N225)			1.00534	0.47631	1.21134	0.33909	1.03105	0.41151		
대만 기권(^TWII)			1.07770	0.60618	0.98682	0.39589	1.03540	0.42434		
비트코인 암호화폐(BTC-USD)			1.11735	0.73654	0.90821	0.53519	1.13430	0.36602		
Gold price(GC=F)			0.73026	0.29626	0.79058	0.38358	0.98631	0.39439		
Oil price(CL=F)			0.80273	0.48293	1.13579	0.53472	0.84622	0.74950		
이더리움 암호화폐(ETH-USD)			0.95489	0.64710	0.69582	0.43093	0.86392	0.47846		
10년 만기 미국 국채										
10년 만기 한국 국채										

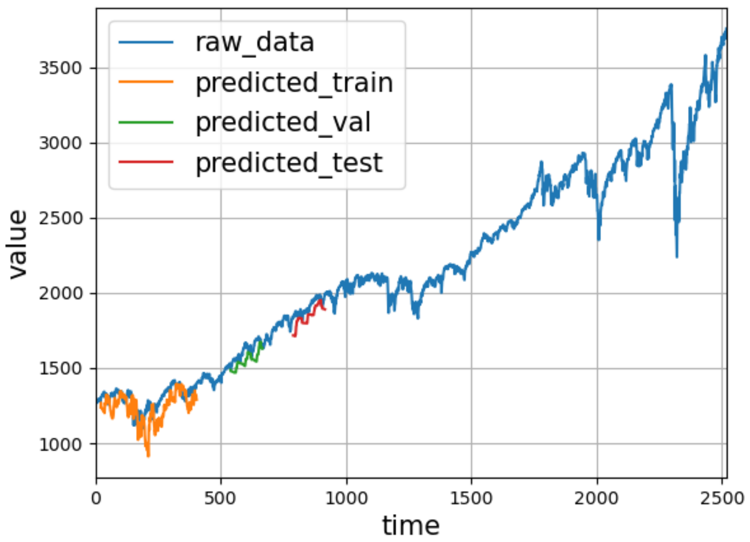
Comparison results

- Mean Absolute Percentage Error (Normalized to each sequence)

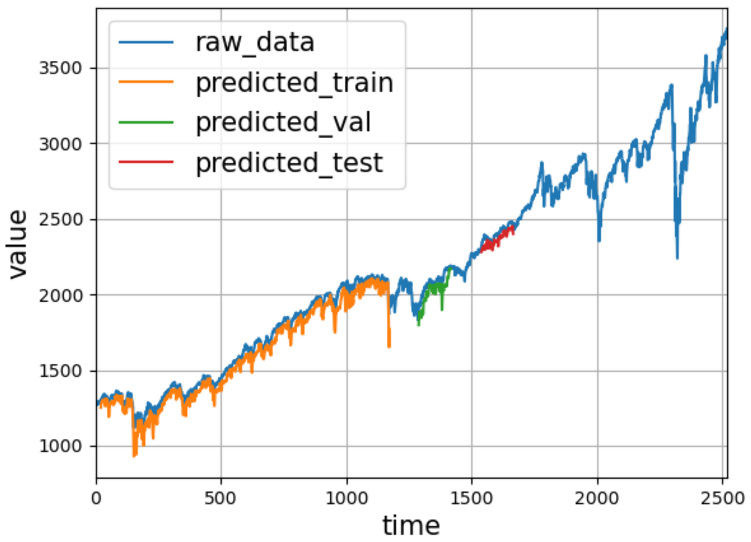
Data	Proposed	Standard Deviation	RNN	Standard Deviation	LSTM	Standard Deviation	GRU	Standard Deviation	Enc-Dec LSTM	Standard Deviation
네덜란드 AEX(^AEX)			0.44518	0.09171	0.40539	0.10271	0.43188	0.06942		
인도 선섹스(^BSESN)			0.43434	0.07771	0.46884	0.08388	0.45814	0.05276		
브라질 보베스파(^BVSP)			0.34040	0.09554	0.35737	0.12688	0.40995	0.10208		
다우존스(^DJI)			0.38940	0.09696	0.43311	0.07100	0.42279	0.10788		
프랑스 CAC(^FCHI)			0.42830	0.06250	0.41519	0.12807	0.37230	0.07585		
독일 DAX(^GDAXI)			0.43219	0.08924	0.42404	0.09270	0.45985	0.08355		
S&P 500(^GSPC)			0.44663	0.07040	0.43601	0.09132	0.45420	0.09175		
홍콩 항생(^HSI)			0.37032	0.10301	0.40924	0.10745	0.35350	0.11606		
스페인 IBEX(^IBEX)			0.43694	0.10753	0.40345	0.05278	0.33714	0.10416		
NASDAQ(^IXIC)			0.47022	0.06222	0.44743	0.10498	0.48110	0.07792		
KOSDAQ(^KQ11)			0.41345	0.08314	0.41407	0.09285	0.44519	0.02210		
KOSPI(^KS11)			0.35576	0.08301	0.38959	0.06275	0.34838	0.08557		
일본 니케이(^N225)			0.39678	0.09363	0.45135	0.07503	0.41156	0.08446		
대만 기권(^TWII)			0.41164	0.12117	0.38861	0.10381	0.41071	0.08376		
비트코인 암호화폐(BTC-USD)			0.41763	0.13608	0.38153	0.10557	0.46262	0.06817		
Gold price(GC=F)			0.34794	0.06486	0.36199	0.09089	0.41714	0.09247		
Oil price(CL=F)			0.37069	0.10529	0.42546	0.09894	0.33773	0.15417		
이더리움 암호화폐(ETH-USD)			0.39220	0.12508	0.32564	0.09519	0.38650	0.10056		
10년 만기 미국 국채										
10년 만기 한국 국채										

Comparison results

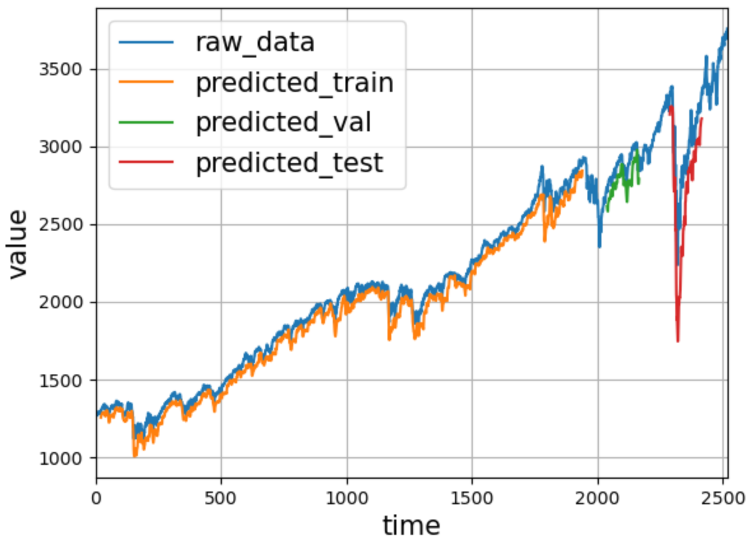
- Predicted Graph(LSTM – S&P 500 Index)



1th-Iteration



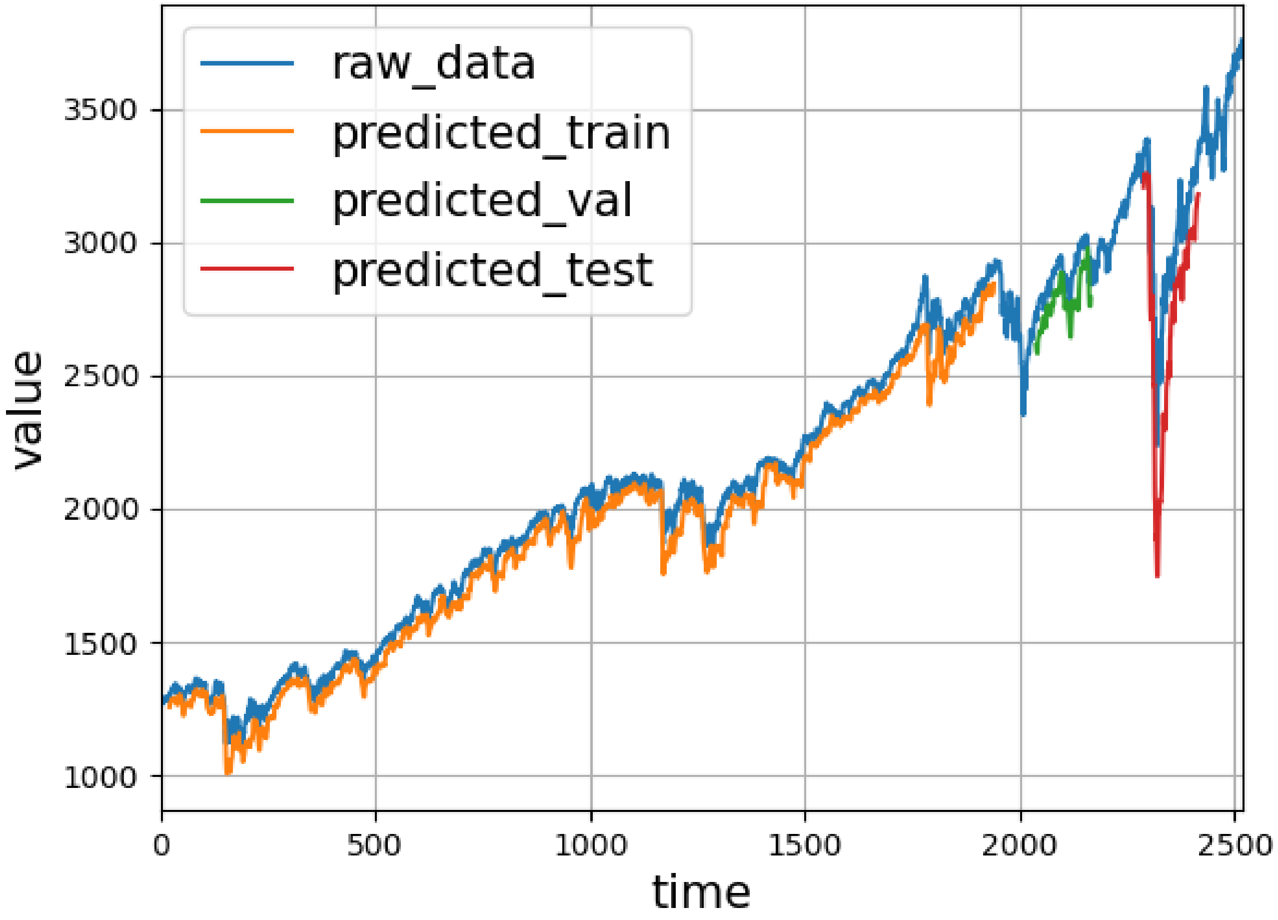
4th-Iteration



7th-Iteration

Comparison results

- Predicted Graph



>> predicted value is lagged

Reference

1 Ha Young Kim, & Chang Hyun Won (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Systems With Applications, Vol.103. 25-37. https://doi.org/10.1016/j.eswa.2018.03.002

2 Qiang Jiang, Chenglin Tang, Chen Chen, Xin Wang and Qing Huang(2019), Stock Price Forecast Based on LSTM Neural Network. Proceedings of the Twelfth International Conference on Management Science and Engineering Management. Vol. 32. 393-408. https://doi.org/10.1007/978-3-319-93351-1_32

3 Sima Siami-Namini, Neda Tavakoli, Akbar Siami Namin(2018), A Comparison of ARIMA and LSTM in Forecasting Time Series. IEEE International Conference on Machine Learning and Applications, Vol.17. 10.1109/ICMLA.2018.00227

4 Hui li, Yunpeng cui, Shuo wang, (Member, IEEE), Juan liu, Jinyuan qin, And Yilin Yang (2020), Multivariate Financial Time-Series Prediction With Certified Robustness. IEEE Access, Vol. 8. 109133 - 109143. 10.1109/ACCESS.2020.3001287

5 Hadi Rezaei, Hamidreza Faaljou, Gholamreza Mansourfar (2020), Stock Price Prediction using Deep Learning and Frequency Decomposition. Expert Systems with Applications, Vol. 169. 114332. https://doi.org/10.1016/j.eswa.2020.114332

Reference

1 Ha Young Kim, & Chang Hyun Won (2018). Forecasting the volatility of stock price index: A hybrid model integrating LSTM with multiple GARCH-type models. Expert Systems With Applications, Vol.103. 25-37. https://doi.org/10.1016/j.eswa.2018.03.002

Appendix

Appendix

- Jarque-bera test

 - 왜도와 첨도가 정규분포와 일치하는지 판단하는 적합도 검정

- ADF test

 - 시계열 데이터가 정상성(stationary)을 만족하는지에 대한 검정

- LM-ARCH 검정

 - ARCH, GARCH와 같은 모형을 적용할 수 있는지에 대한 검정.