Encoder Decoder LSTM
구현 및 실험 결과 보고

Tae-Won Lee,
leetee95@gmail.com
2020. 05. 25.

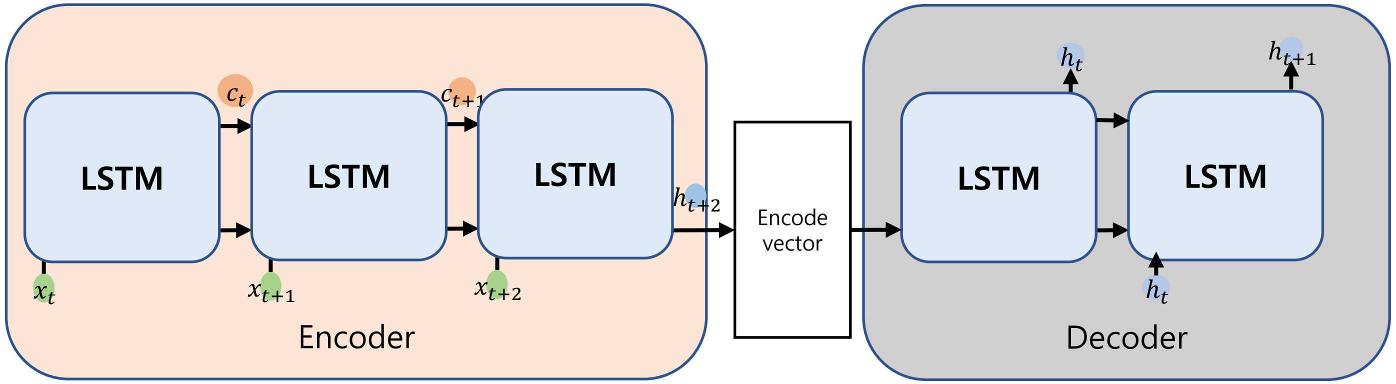
Introduction

- 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

Related works

- 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.

Related works

- Stock data Normalization

학회 명	논문 명	발간 년도	인용 수	normalization 방식	normalization 범위	실제 내용
Physica A: Statistical M echanics and its Applic ations	Financial time series forecasting mod el based on CEEMDAN and LSTM	2018	111	min-max normalizati on	전체 데이터에 적용	Where minx(t) and maxx(t) are the minimum and maximum value of the whole time series respectively.
International Conferenc e on Advances in Com puting, Communication s and Informatics (ICA CCI)	STOCK PRICE PREDICTION USING LS TM,RNN AND CNN-SLIDING WINDO W MODEL	2017	235	min-max normalizati on	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 ra nge, it was subjected to normalization and was mapped to a range of 0 to 1. This normalized data was given to the ne twork for training.
IEEE International Conf erence On Recent Tren ds in Electronics Infor mation & Communicati on Technology (RTEICT)	Short Term Stock Price Prediction Usi ng Deep Learning	2017	41	min-max normalizati on	언급 x	The data of each stock consists of approximately 85000 - 9 0000 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 t he prediction model.
IEEE International Conf erence on Big Data	A LSTM-based method for stock retu rns 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 comp uted from the training set.
International Conferenc e on Business Analytics and Intelligence	Stock Price Prediction Using Machine Learning and Deep Learning Framew orks	2018	19	min-max normalizati on	전체 데이터에 적용	The raw data is normalized using the min-max normalizatio n
Expert Systems With A pplications	ModAugNet: A new forecasting fram ework for stock market index value with an overfitting prevention LSTM module and a prediction LSTM mod ule	2018	80	min-max normalizati on	전체 데이터에 적용	X max and X min are max- imum and minimum value of a specific feature,

Related works

- 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

>> In the result of an experiment, normalization applied to each split sequence shows better performance than normalization applied to entire sequence

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

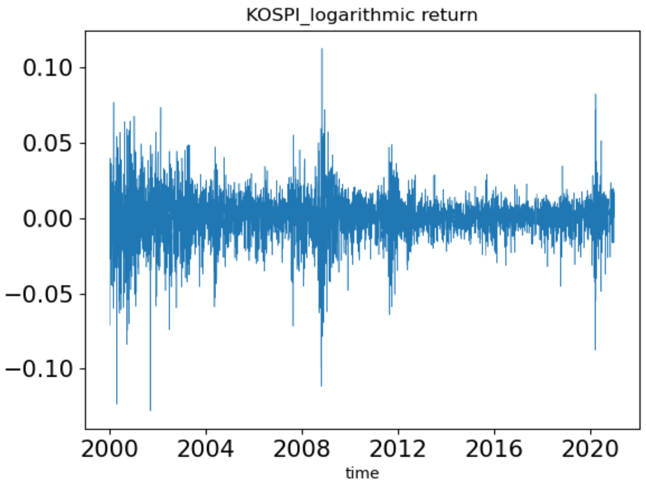
- Statistics

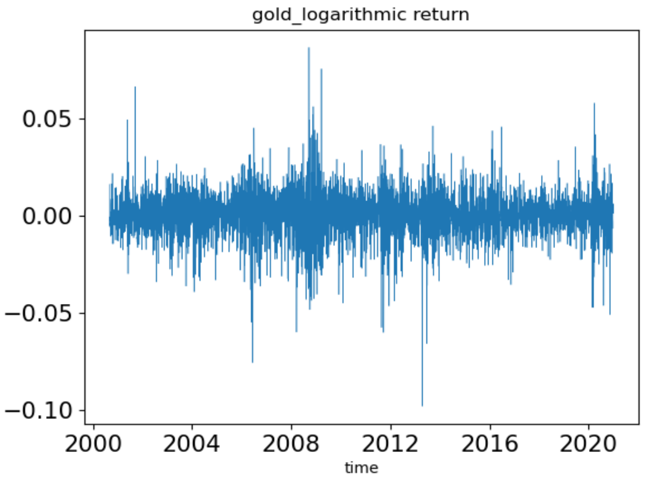
stocks	Mean	Standard deviation	Skewness	Kurtosis(3)	Jerque Bera	ADF
KOSPI	1582.518	587.343	-0.421	1.861	432.61***	-0.438834
KOSDAQ	668.623	103.98	0.46	2.632	78.4663***	-1.173884
NASDAQ	3791.769	2367.556	1.388	4.395	2125.9183***	3.346301
S&P 500	1653.635	674.270	1.033	3.067	941.5057***	1.678640
다우존스	14655.366	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***

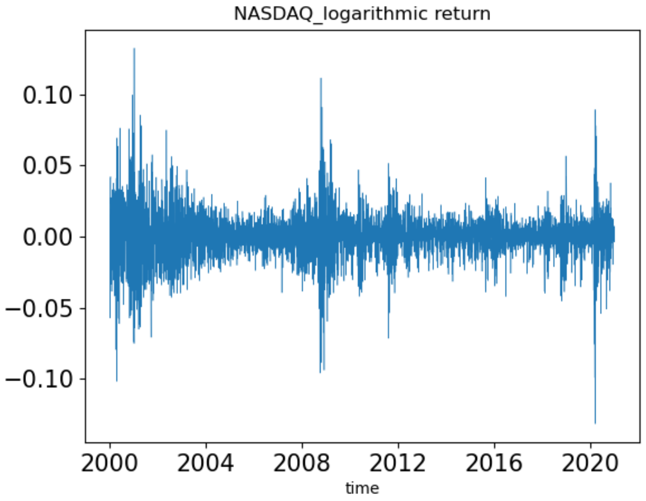
Used Close(종가) variable.

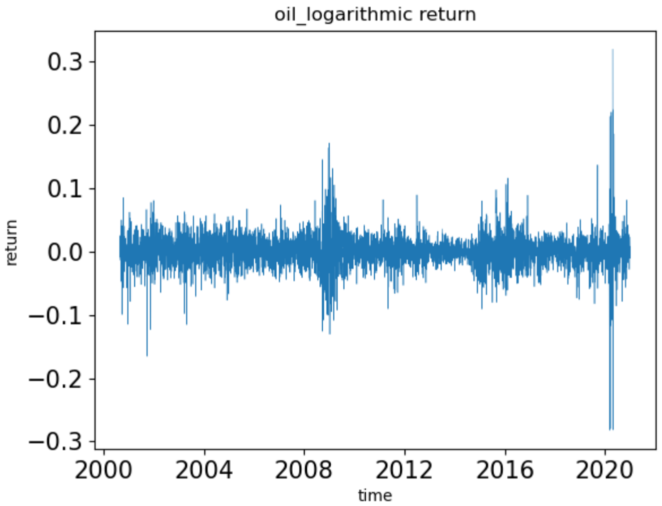
Experimental setting

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

Hyperparameters	**Number**
hidden state, cell state dimension	10
input sequence length	20
output sequence length	1
epoch	120
Batch size	128

Experimental setting

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

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

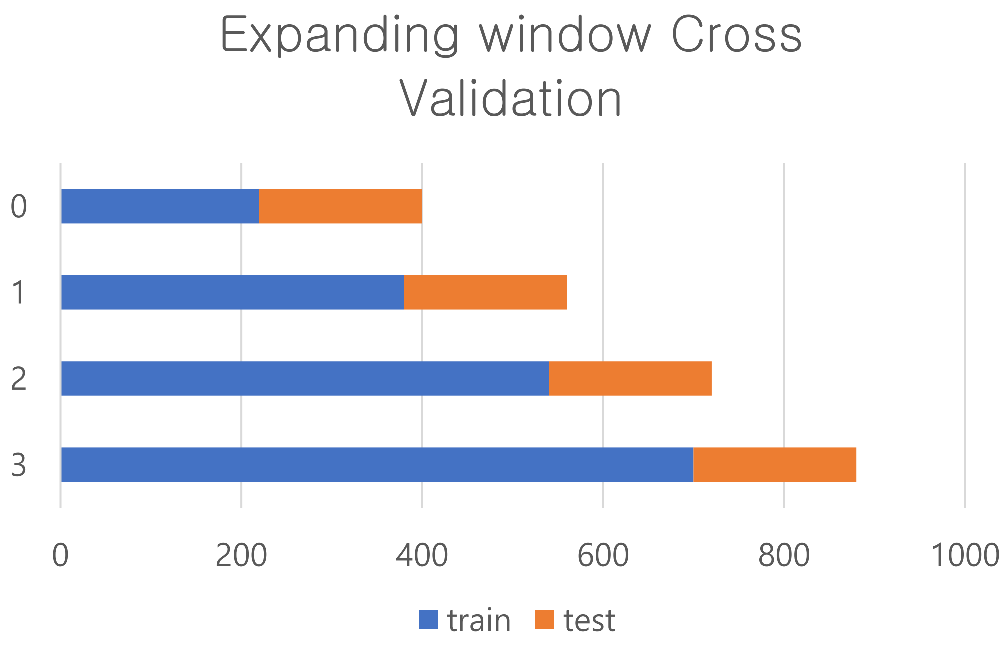
      - MAPE is undefined for data points where the value is zero.
      - MAPE is biased towards prediction that are systemically less than the actual values themselves

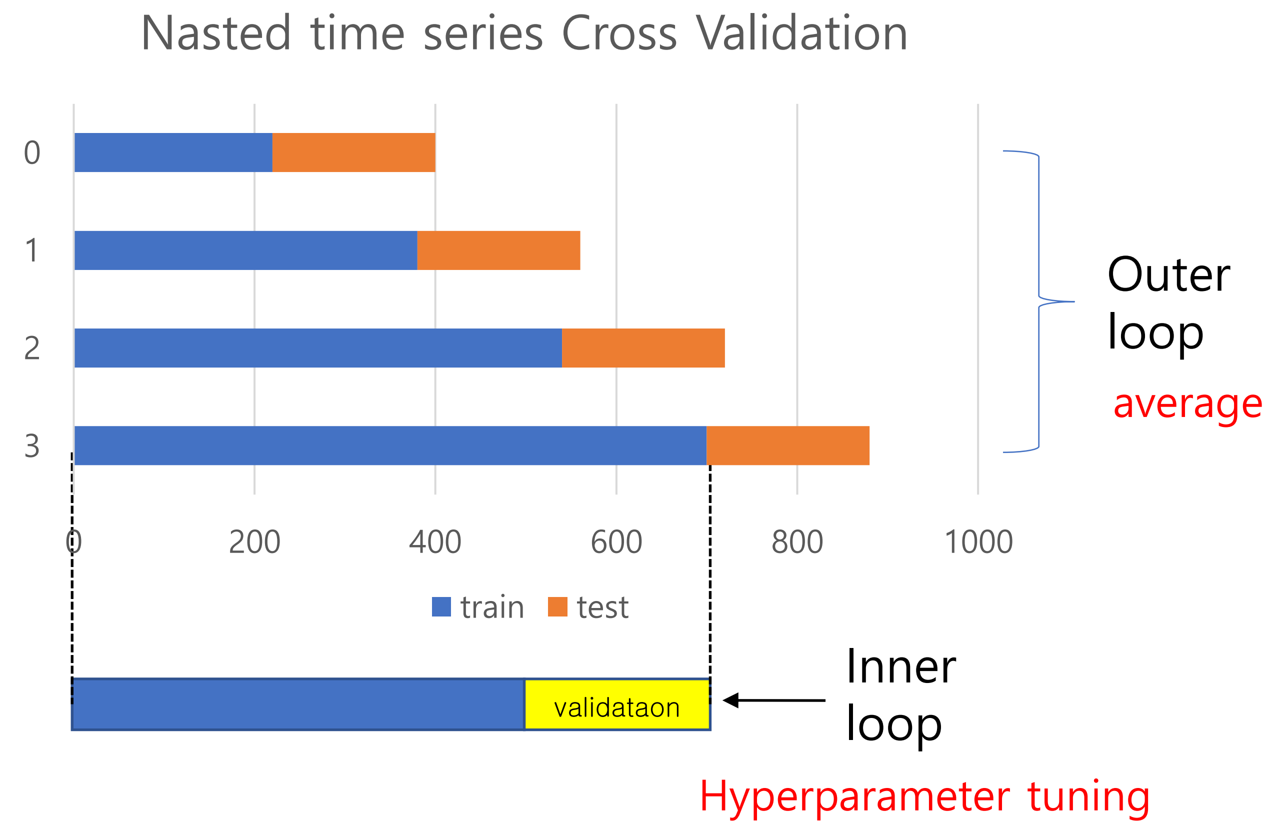
    $y = 20, \hat{y} = 10, n = 1 \ \rightarrow MAPE = 50\%$
    $y = 10, \hat{y} = 20, n = 1 \ \rightarrow MAPE = 100\%$

  - Compare three metrics.
    - MAE
    - RMSE
    - MAPE

Experimental setting

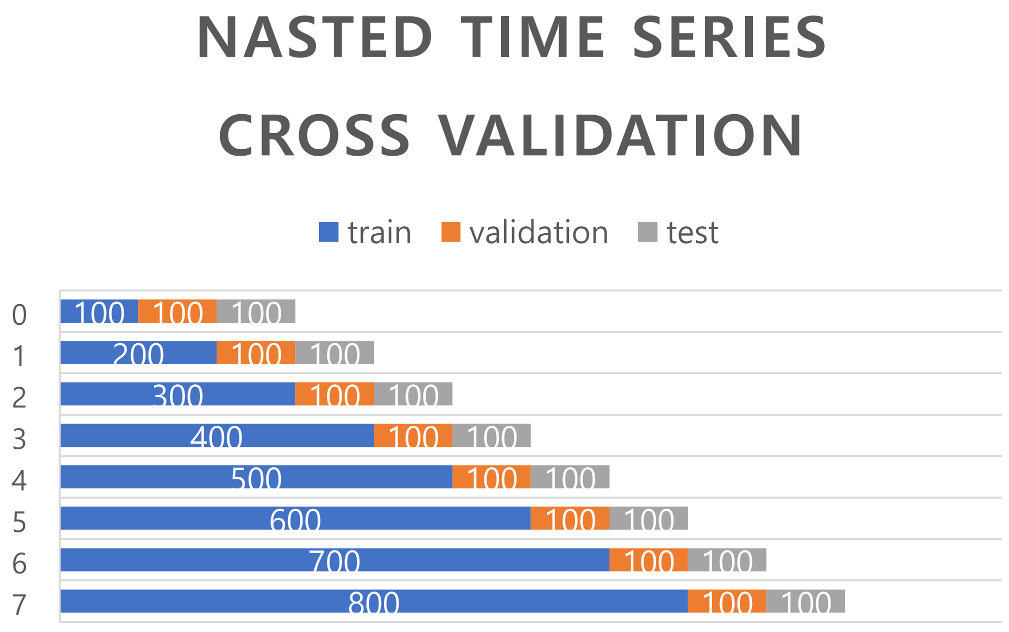
- Stock dataset split in timeseries data





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.

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)			9.7	4.321	9.072	2.430	12.169	5.305	9.249	3.630
인도 선섹스(^BSESN)			522.481	143.644	757.71	311.438	635.134	240.285	409.767	121.605
브라질 보베스파(^BVSP)			1531.036	549.161	1529.299	754.228	2040.192	650.570	2188.41	748.355
다우존스(^DJI)			379.666	155.520	386.981	130.981	383.127	141.125	438.104	263.496
프랑스 CAC(^FCHI)			132.848	64.138	114.632	47.652	86.384	43.260	110.651	51.728
독일 DAX(^GDAXI)			287.096	174.637	236.222	97.034	276.421	131.694	283.876	112.716
S&P 500(^GSPC)			45.127	28.311	40.281	20.761	33.122	14.978	42.765	21.438
홍콩 항생(^HSI)			625.978	235.568	464.189	119.809	592.841	292.263	562.736	159.684
스페인 IBEX(^IBEX)			223.378	159.649	222.907	108.678	245.819	167.813	235.776	104.426
NASDAQ(^IXIC)			109.496	46.546	122.264	41.803	126.881	49.249	157.227	69.242
KOSDAQ(^KQ11)			27.665	9.407	25.22	14.498	23.798	11.050	23.851	10.023
KOSPI(^KS11)			35.608	13.477	39.714	13.813	31.784	11.158	35.751	16.443
일본 니케이(^N225)			537.349	282.333	482.213	198.073	539.397	336.067	692.156	356.340
대만 기권(^TWII)			159.542	66.302	172.253	62.575	157.141	58.127	135.595	61.697
비트코인 암호화폐(BTC-USD)			419.973	274.202	325.423	249.265	542.294	699.560	559.299	763.136
Oil price(CL=F)			2.953	1.809	1.99	0.913	2.678	2.093	2.325	1.501
이더리움 암호화폐(ETH-USD)			40.476	33.185	28.455	21.580	27.205	21.683	42.154	33.436
Gold price(GC=F)			21.052	12.822	20.112	14.340	22.077	14.279	27.564	12.884

R Vinayakumar; E. A Gopalakrishnan; Vijay Krishna Menon; K. P. Soman, Stock price prediction using LSTM, RNN and CNN-sliding window model (2017)

Comparison results

- Root 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)			11.304	5.039	10.566	3.010	14.227	6.222	10.572	3.943
인도 선섹스(^BSESN)			625.698	182.463	869.819	402.124	727.703	268.761	506.799	170.489
브라질 보베스파(^BVSP)			1858.968	676.142	1829.228	913.358	2366.848	728.495	2558.67	883.479
다우존스(^DJI)			451.991	207.007	447.909	167.888	449.206	177.294	517.604	345.870
프랑스 CAC(^FCHI)			154.815	76.128	132.486	56.154	102.502	47.582	131.174	64.944
독일 DAX(^GDAXI)			339.719	193.275	274.045	103.857	319.372	161.814	328.517	124.275
S&P 500(^GSPC)			52.531	35.028	47.002	25.976	39.627	19.338	49.762	25.417
홍콩 항생(^HSI)			751.627	282.089	540.112	120.722	687.794	313.541	671.273	197.581
스페인 IBEX(^IBEX)			262.575	175.966	256.591	116.325	278.552	174.035	275.532	119.596
NASDAQ(^IXIC)			127.552	57.086	144.6	50.219	147.346	57.546	177.893	77.217
KOSDAQ(^KQ11)			31.153	9.875	29.395	16.572	27.159	11.997	27.257	10.410
KOSPI(^KS11)			41.436	14.432	44.939	14.383	37.307	13.317	42.37	18.105
일본 니케이(^N225)			610.658	288.148	550.857	213.837	614.475	368.112	778.45	376.288
대만 기권(^TWII)			183.53	70.931	197.271	68.478	179.187	59.845	157.723	69.263
비트코인 암호화폐(BTC-USD)			536.844	363.039	427.415	327.390	666.699	817.629	683.818	889.893
Oil price(CL=F)			3.894	2.953	2.794	2.216	3.594	3.203	3.18	2.238
이더리움 암호화폐(ETH-USD)			50.98	44.595	36.175	26.770	34.514	27.134	50.385	39.829
Gold price(GC=F)			26.2	16.438	25.728	17.718	28.299	18.431	33.724	16.270
10년 만기 미국 국채										
10년 만기 한국 국채										

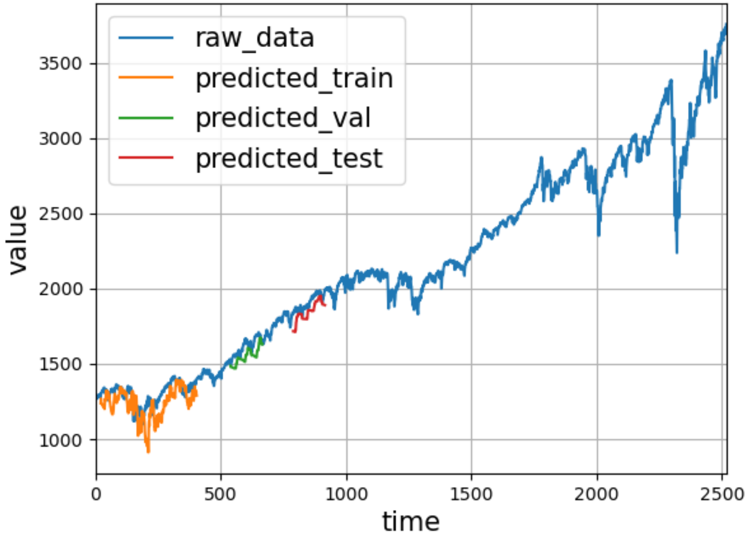
Comparison results

- Mean Absolute Percentage Error (Normalized to each sequence)

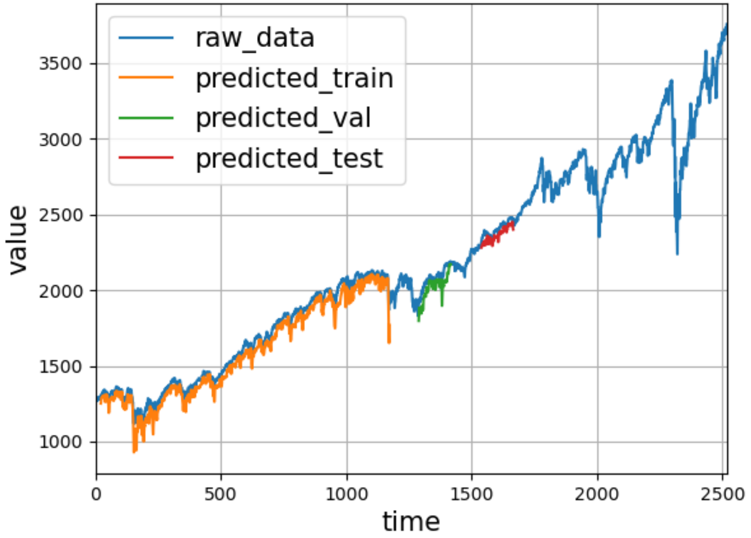
Data	Proposed	Standard Deviation	RNN	Standard Deviation	LSTM	Standard Deviation	GRU	Standard Deviation	Enc-Dec LSTM	Standard Deviation
네덜란드 AEX(^AEX)			0.02149	0.01151	0.01938	0.00575	0.02651	0.01254	0.02022	0.00925
인도 선섹스(^BSESN)			0.01874	0.00637	0.027	0.01201	0.02263	0.00982	0.01509	0.00650
브라질 보베스파(^BVSP)			0.02367	0.00858	0.02444	0.01348	0.032	0.01248	0.03465	0.01565
다우존스(^DJI)			0.01884	0.00642	0.01926	0.00540	0.01879	0.00609	0.02153	0.01073
프랑스 CAC(^FCHI)			0.0292	0.01604	0.02491	0.01165	0.01842	0.00936	0.02364	0.01140
독일 DAX(^GDAXI)			0.02807	0.01809	0.02219	0.00922	0.02619	0.01290	0.02743	0.01374
S&P 500(^GSPC)			0.01934	0.01087	0.01704	0.00705	0.01425	0.00514	0.01949	0.01068
홍콩 항생(^HSI)			0.02536	0.01034	0.01907	0.00639	0.02369	0.01168	0.02286	0.00739
스페인 IBEX(^IBEX)			0.02469	0.01909	0.02437	0.01341	0.02631	0.01815	0.02631	0.01407
NASDAQ(^IXIC)			0.01863	0.00593	0.02108	0.00644	0.02229	0.00838	0.02689	0.01003
KOSDAQ(^KQ11)			0.03983	0.01357	0.03586	0.02034	0.03414	0.01655	0.03399	0.01325
KOSPI(^KS11)			0.0169	0.00552	0.01907	0.00631	0.01553	0.00614	0.01715	0.00803
일본 니케이(^N225)			0.03159	0.02126	0.02795	0.01461	0.0323	0.02505	0.04045	0.02692
대만 기권(^TWII)			0.01696	0.00828	0.01824	0.00751	0.01643	0.00649	0.01453	0.00793
비트코인 암호화폐(BTC-USD)			0.07405	0.03593	0.05769	0.02862	0.08609	0.07321	0.07933	0.06390
Oil price(CL=F)			0.0705	0.08610	0.04872	0.05775	0.06203	0.07696	0.04769	0.03948
이더리움 암호화폐(ETH-USD)			0.14305	0.09475	0.10787	0.09124	0.11357	0.07691	0.14004	0.08726
Gold price(GC=F)			0.01556	0.00919	0.01484	0.01028	0.01643	0.01052	0.02061	0.00972
10년 만기 미국 국채										
10년 만기 한국 국채										

Comparison results

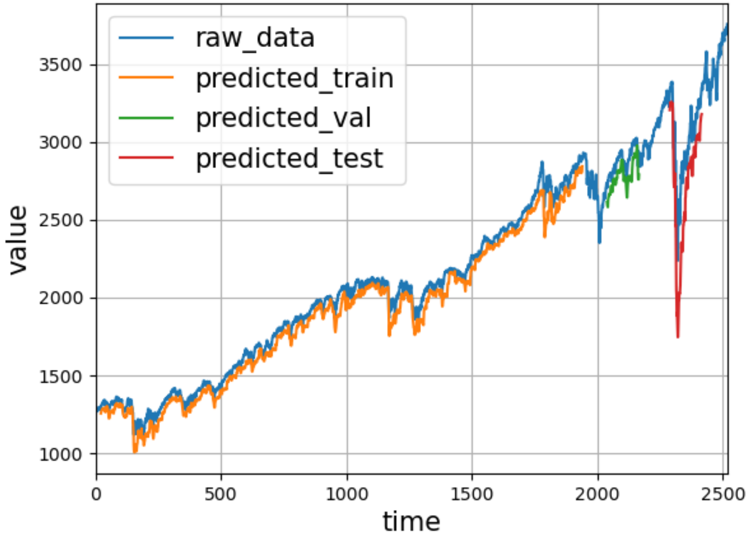
- Predicted Graph(LSTM – S&P 500 Index) –each seq



1th-Iteration



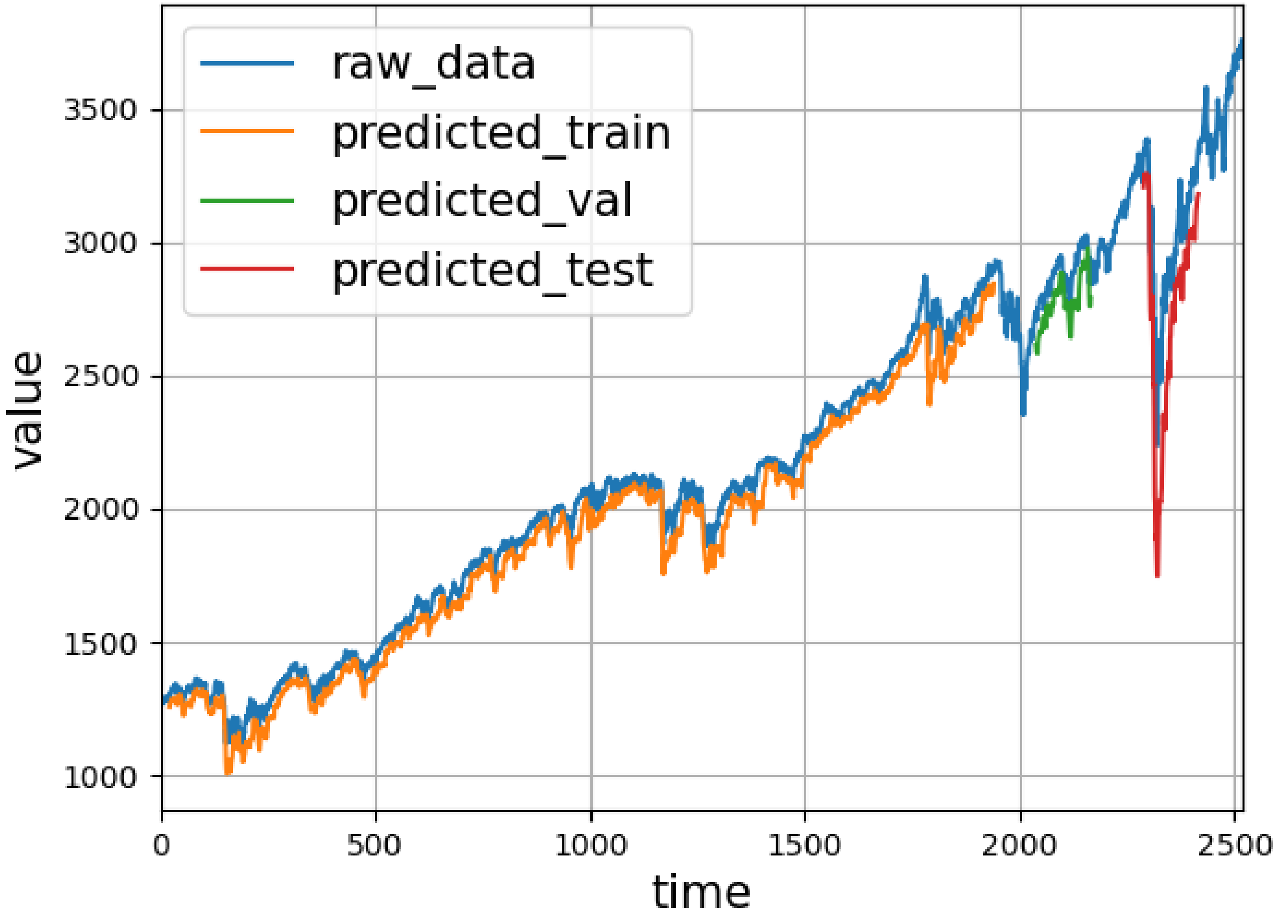
4th-Iteration



7th-Iteration

Comparison results

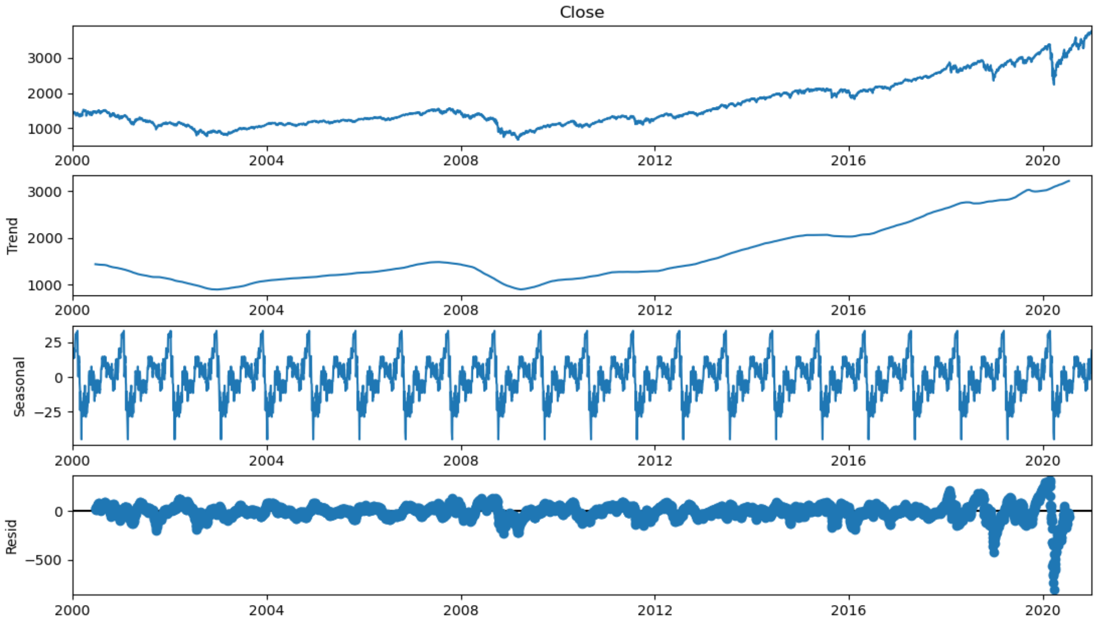
- Predicted Graph –each seq



>> predicted value is lagged

Issue

- Time series decomposition (S&P 500 Index)



1) 기존의 데이터를 이동평균을 이용하여 Trend를 계산
(이동평균의 개수는 freqency를 그래프를 보고 지정)

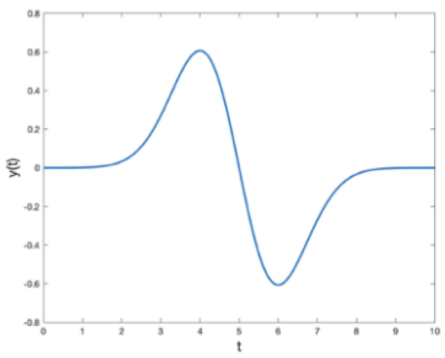
2) 전체 데이터에서 계산된 Trend를 뺀 후 지정된 frequency 만큼의 term을 뒤 평균을 낸 값을 seasonal로 사용.

3) (Raw data) – (seasonal) – (trend) = Residual

>> 추출된 Residual을 arima등 전통적인 시계열 예측을 위해 정상시계열(Stationary)로 만드는데 사용됨
>> frequency를 지정해줘야함.

Issue

- 주식 가격의 노이즈 제거 및 분해를 위해 사용되는 방식.
  - STL(Seasonal and Trend decomposition using Loess)
    - loess regression을 이용
      1) 적합한 근사 추세선을 찾는다.
      2) 추세이탈 시계열에서 seasonal 시계열을 찾는다.
      3) De-seasonalised data에서 다항회귀를 적용한다.
  - Wavelet Transform
    - global frequency information를 포착하는 Fourier Transform의 대안으로 사용됨
    - 우측의 그림과 같은 파형을 생성해 위치를 옮겨가며
      convolution 연산을 시켜주면 계수를 얻을 수 있고,
      얻어진 계수로 wavelet scale 을 증가시고 해당 과정을 반복함.



Issue

- 주식 가격의 노이즈 제거 및 분해를 위해 사용되는 방식.

  - Empirical Mode Decomposition(EMD) – (Hilbert-Huang transform)

    - EMD는 raw series를 n개의 IMF(Intrinsic mode functions, 고유진동함수)들과 residual로 분해한다.

    - Parameter를 정의할 필요가 없다.

    - EMD는 비선형(non-linear), 비정상(non-stationary) 시계열에 적용이 가능한 시공간 분석 방법(time-space analysis method)이다.

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.