

Comparison Long Short-Term Memory Network and Convolution Neural Network, with additional data

LSTM 네트워크, CNN 네트워크 구현 및
데이터 추가



Geonil Yun,
rjsdlf45@gmail.com
2021. 06. 15.

Missing Value Imputation

Mean-Mode imputation

- fill missing data with mean or mode from data set (or median)

dataset

index	x	y
0	10	7
1	1	7
2	6	missing
3	missing	10
4	3	7

fill 5 (mean)

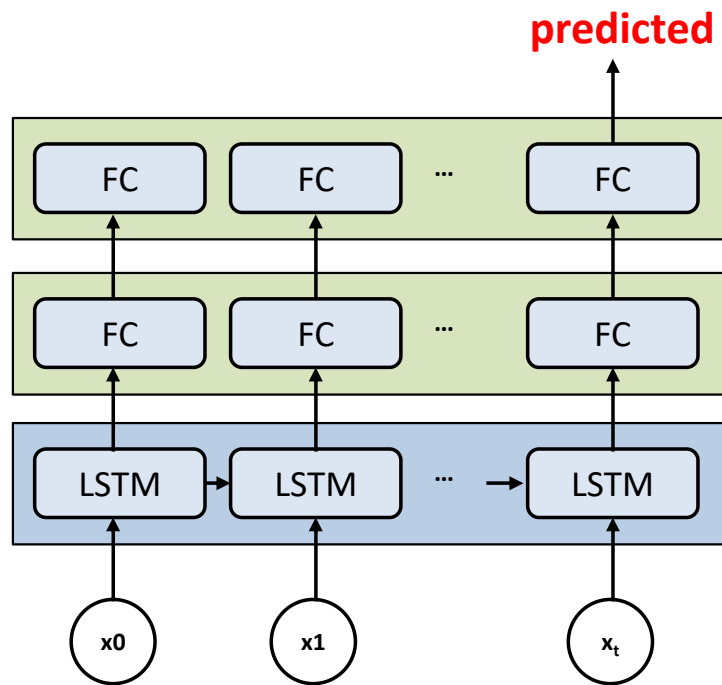
fill 7 (mode)

- using representative value : **fast**
- not consider the **correlation** with other features
- low **accuracy**
- mean reduce **variance**
- mode create **bias**

Experimental setting

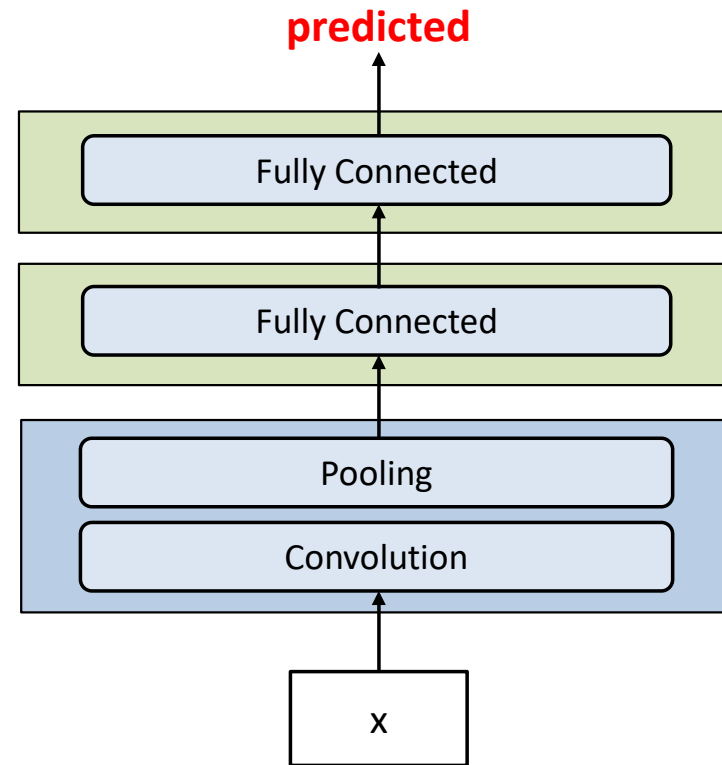
Structure (LSTM)

$$\text{MSE for Loss} = \frac{1}{2} \sum (y - t)^2$$



- input data shape : (batch, timestep, features)

Structure (CNN)

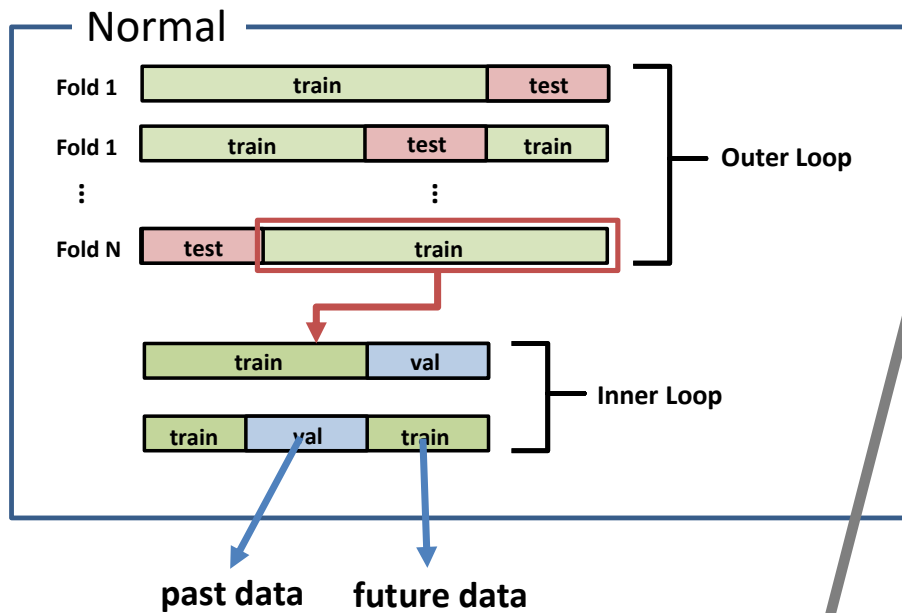


- input data shape : (batch, 1, timestep, features)

Experimental setting

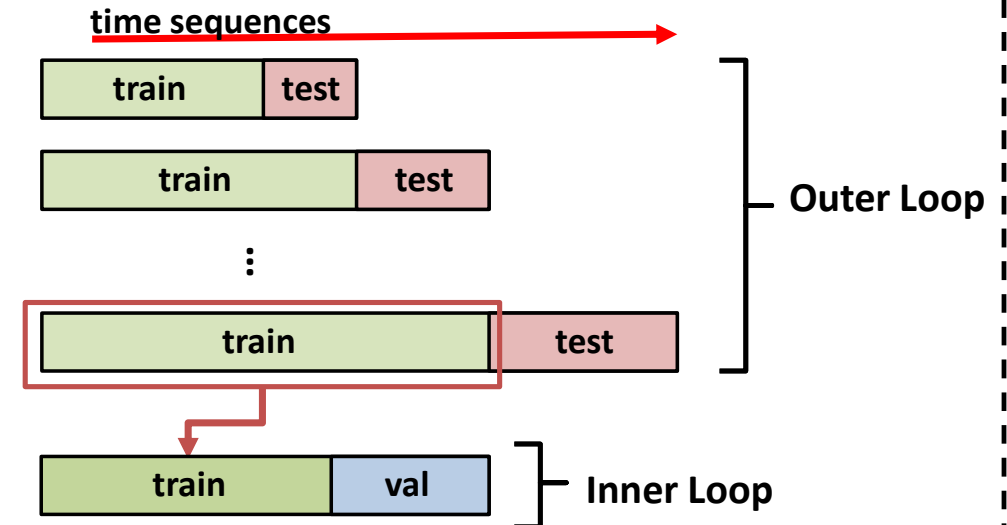
Nested Cross Validation

- Problem in **Time series data**



- train **future** data & predict **past** data

Nested Cross Validation for time series



- train **past** data & predict **future** data
- handling time series data in Nested CV

Experimental setting

- Dataset

- Beijing Air pollution (uci-dataset/00381)
- Hungarian Chickenpox (uci-dataset/00580)
- household power consumption (uci-dataset/00235) → **Missing Value (use mean imputation)**

- **Additional Dataset**

- Air Quality (uci-dataset/00360) → **Missing Value (forward fill imputation)**
- Appliances Energy (uci-dataset/00374)
- Istanbul Stock Exchange (uci-dataset/00247)
- Metro Interstate Traffic Volume (uci-dataset/00492)
- SML 2010 (uci-dataset/00274)

Experimental setting

Cross validation

- hyperparameter settings

Dataset	Beijing air pollution		Hungarian Chickenpox		household power consumption	
Model	CNN	LSTM	CNN	LSTM	CNN	LSTM
filter size	3, 5		3, 5		3, 5	
filter num	30		30		30	
hidden size	100		100		100	
input length	24		8		30	
output length	1		1		1	
epoch	10	10	30	30	30	30
batch size	64		8, 16		32, 64	
learning rate	0.1, 0.01	0.1, 0.01, 0.001	0.1, 0.01	0.01, 0.001	0.1, 0.01	0.01, 0.001

Experimental setting

Cross validation

- Same hyperparameter in Others(Air quality, ISE, Metro, SML)

Dataset	Appliances Energy		Ohters	
Model	CNN	LSTM	CNN	LSTM
filter size	3, 5		3, 5	
filter num	30		30	
hidden size	100			
input length	30		24	
output length	1			
epoch	30			
batch size	32, 64			
learning rate	0.1, 0.01	0.01, 0.001	0.1, 0.01	0.01, 0.001

Comparison results

- RMSE (Root Mean Squared Error)

Data		LSTM	CNN
Beijing Air Pollution		21.47384	22.61476
Hungary Chickenpox		70.46656	69.41699
Household Power Consumption		0.49048	18.71137
Air Quality		0.74999	257.05172
Appliances Energy		74.74476	97.68157
Istanbul Stock Exchange	ISE	2.34701	352.61487
	SP	2.25855	71.80231
Metro Interstate Traffic Volume		392.73909	598.22332
SML 2010		0.10628	0.40663

Comparison results

- MAE (Mean Absolute Error)

Data		LSTM	CNN
Beijing Air Pollution		13.37477	13.74278
Hungary Chickenpox		39.21686	45.16213
Household Power Consumption		0.27045	18.12832
Air Quality		0.52713	254.11711
Appliances Energy		39.91200	60.79923
Istanbul Stock Exchange	ISE	2.34508	352.56304
	SP	2.25635	71.78394
Metro Interstate Traffic Volume		262.32308	454.25089
SML 2010		0.08351	0.29864

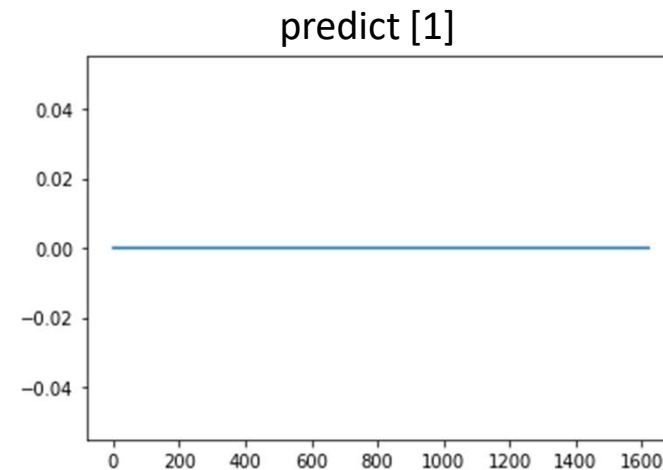
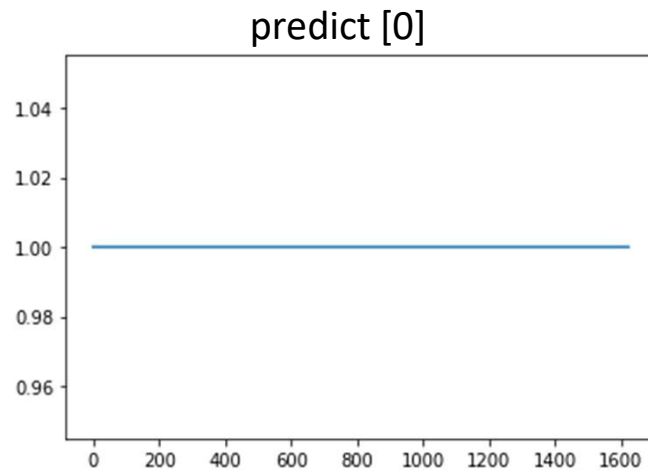
ISSUE

Classification

- using CEE(Cross Entropy Error) in classification task

$$CEE = - \sum_i t_i \log(y_i)$$

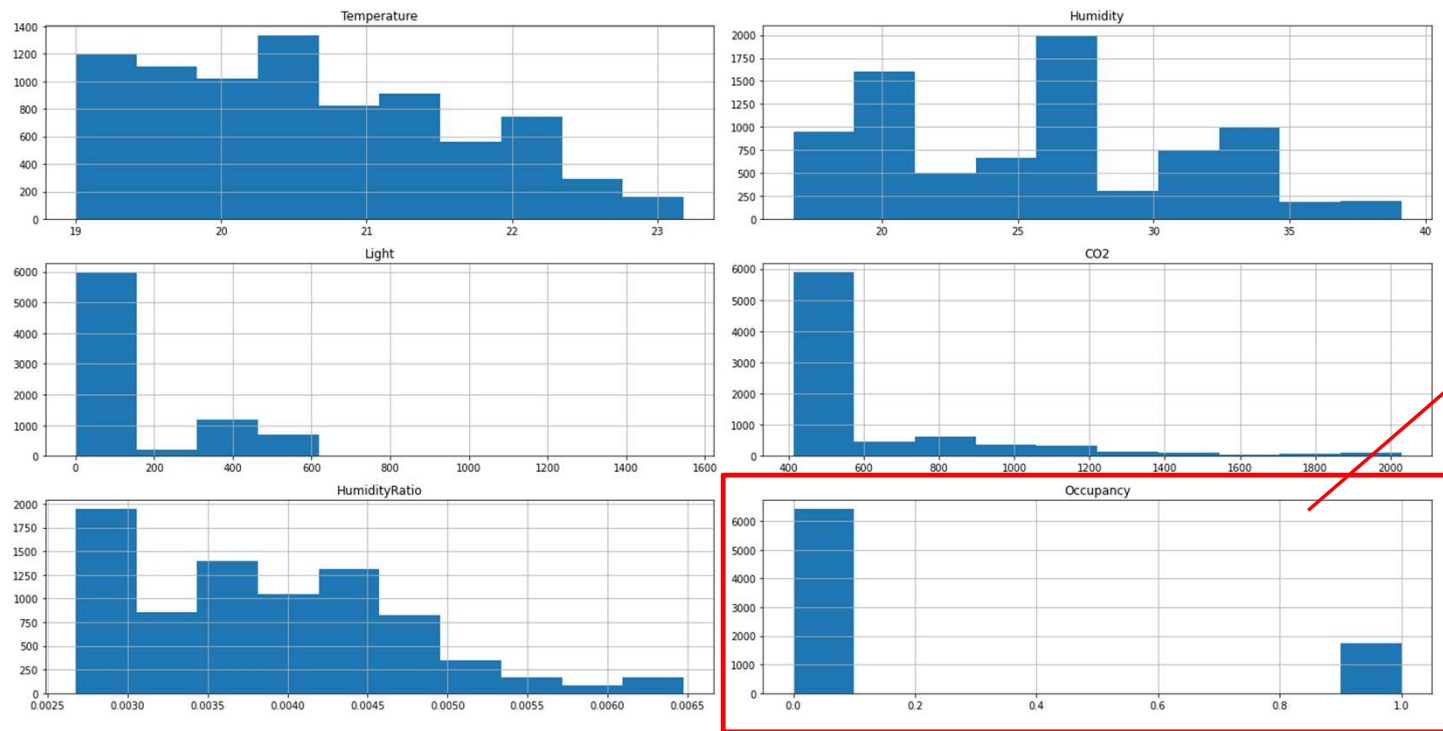
- error = 0 , in second fold of nested cross validation



ISSUE

Classification

- using CEE(Cross Entropy Error) in classification task



Imbalance data

Thank you

Appendix : Missing Value

Deletion

- Likewise deletion

dataset

index	x	y
0	10	7
1	3	5
2	6	missing
3	missing	10
4	3	2

remove

- Remove index 2, 3 from the dataset

- Dropping variable

dataset

index	x	y	z
0	10	7	missing
1	3	5	missing
2	6	missing	missing
3	missing	10	missing
4	3	2	7

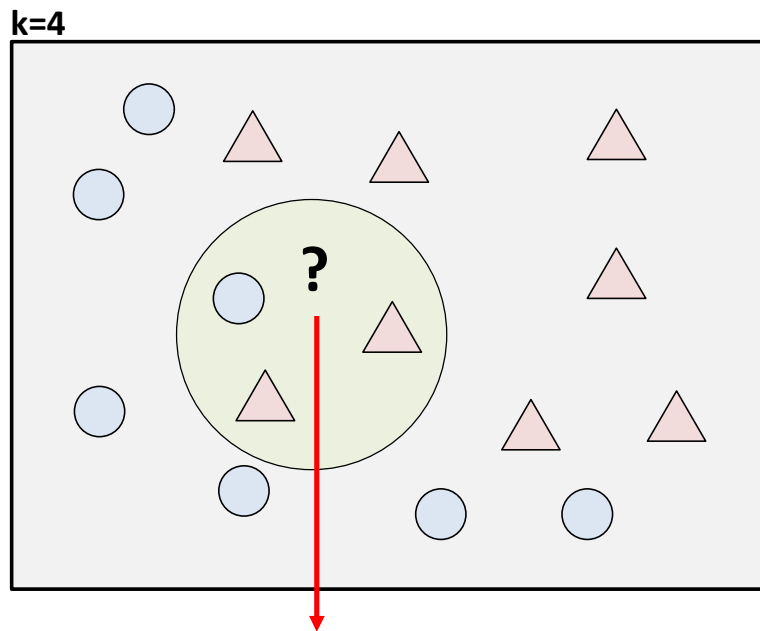
remove

- Remove 'z' column from the dataset

Appendix : Value Imputation

KNN imputation

- fill missing value by using KNN(K-Nearest Neighbor)

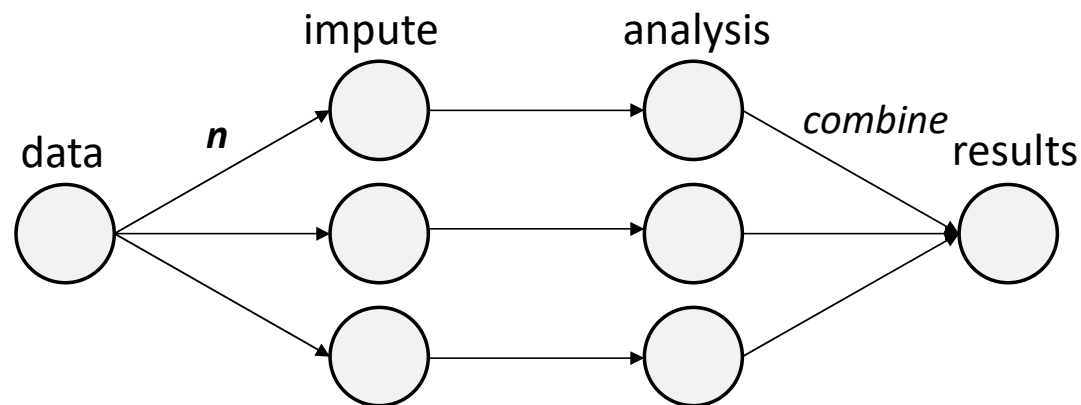


- More **accurate** than M-M impute
- Sensitive to **feature scale**
- Too many **calculations**

Appendix : Missing Value Imputation

Multiple Imputation

- fill missing value by using Multiple Imputation
 - ① **Imputation** : impute n datasets.
 - ② **Analysis** : analysis result of n datasets.
 - ③ **Combine** : combine result of n datasets.



Appendix : Code

Model	view with code
LSTM	https://colab.research.google.com/drive/1oNW4HzBzmrjyWcWZbjFMtksP1vTV9IO2?usp=sharing
CNN	https://colab.research.google.com/drive/1Ccd7l65bJwiD2HW8XmOKTHO1PEUMuQI3?usp=sharing
LSTM-Beijing air pollution	https://colab.research.google.com/drive/1De0Tn_iD5NQBKAu_eKr3PWZwvcnNTu97?usp=sharing
CNN-Beijing air pollution	https://colab.research.google.com/drive/1HmjdyXbwUASZxzS40v4opc9myAuuhcrF?usp=sharing
LSTM-Hungarian chickenpox	https://colab.research.google.com/drive/1eeLLtCJGzQ2u_V09pXF0dHg2Oz4YHfxn?usp=sharing
CNN-Hungarian chickenpox	https://colab.research.google.com/drive/1E4PywHa-xACYhHTiMiHRwEkCHT9eBz4A?usp=sharing
LSTM-Power consumption	https://colab.research.google.com/drive/1HVLLJMPGJeixJ4H9O3lvzfUhUWp8DAXN?usp=sharing
CNN-Power consumption	https://colab.research.google.com/drive/1d7wu9u7zo_URAXYd2E01uROGcsdoSc87?usp=sharing