[Code Review] Network Traffic Prediction

2023-2 URP | SKKU AAI Oh Soo Jeong & Park Su Yeon

WHAT IS TIME SERIES FORECASTING?

- The task of fitting a model to historical, time-stamped data in order to predict future values.
- Time dependent: the values for every period are not only affected by outside factors, but also by the values of past periods.
- The most popular benchmark dataset: ETTh1 dataset
- Typical evaluation metrics: MSE, RMSE, MAE, MAPE, SMAPE, ⋯

TIME SERIES DATA의 속성

- Seasonal: 주기적으로 반복되는 패턴
- Cyclical: 장기적인 주기 변동
- Trend: 장기적으로 상승하거나 하락하는 패턴
- Random(residual): 예측 불가능한 임의의 변동성, 불규칙 요소

↓

Time series data decomposition: 시계열에 영향을 주는 요인들을 분리하여 분석

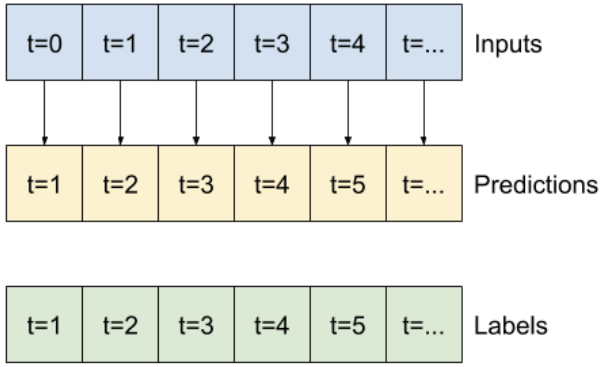
TIME SERIES FORECASTING MODEL

- Traditional approaches include moving average, exponential smoothing, and ARIMA.
- RNN, Transformers, XGBoost, ...
- 1) Smoothing: real world data의 노이즈 최소화를 위한 알고리즘. 과거부터 현재까지 수집한 관측한 데이터를 활용하여 현재까지의 히스토리 전체를 재추정. 데이터의 추세 두드러지게 확인하기 위해 사용.
- 2) ARIMA: 데이터의 비정상성(Non-stationarity)를 정상성을 가진 형태로 변환시키고, 그 후 AR 모델과 MA 모델을 결합하여 예측 수행하는 통계적 방법론.
- 3) LSTM: RNN의 한 종류로 데이터 사이의 장기적인 의존성을 파악하는데 유용하게 사용.

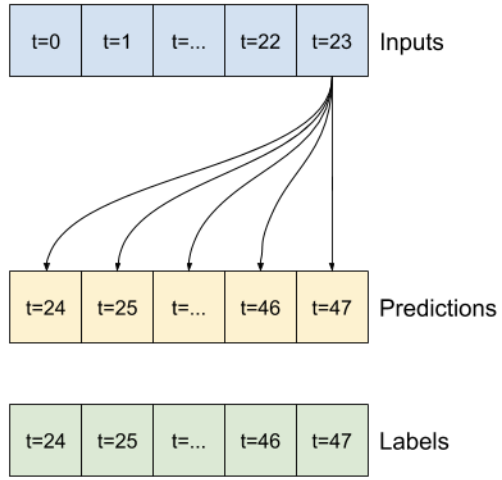
* Stationary(정상성): 시계열이 시간에 상관 없이 일정한 성질을 띄우는 것. 주기성이나 추세가 보이지 않으며, 임의의 시점 t에 대한 기대 값과 분산이 어느 시점에서 관측해도 변하지 않음을 의미함.

TYPES OF TIME SERIES FORECASTING MODEL

- Single-step model
  입력 데이터로 현재 시점 까지의 정보를 사용하고, 다음의 single step을 예측



- Multi-step model
  입력 데이터로 현재 시점 까지의 정보를 사용하고 미래의 여러 step을 예측



TIME SERIES DATA SPLITTING

- Training/ Validation/ Test
- The data is not being randomly shuffled before splitting.
  - Ensures that chopping the data into windows of consecutive samples is still possible
  - Ensures that the validation/test results are more realistic, being evaluated on the data collected after the model was trained

TIME SERIES DATA PREPROCESSING

Data scaling

- important to scale features before training neural network
- prevent overflow or underflow

- Standard scaler: 기본 scale로 평균을 제외하고 표준편차를 나누어 반환
- Min-max scaler: 0~1 또는 -1~1 사이의 값으로 반환, 분포의 모형을 그대로 지켜줌. 각 feature가 정규분포가 아니거나 표준편차가 매우 작을 때 효과적
  -> outlier에 취약하므로, 표준편차가 커지면 취약해짐
- Robust scaler: min-max 대신 IQR (25%, 75%)값을 사용하여 반환, outlier에 영향을 최소화하였기에 outlier가 있는 데이터에 효과적임
- Normalizer: 각 변수(feature)를 전체 n개 모든 feature들의 크기들로 나누어서 변환, 각 변수들의 값은 원점으로부터 반지름 1만큼 떨어진 범위 내로 변환

BANDWIDTH USAGE OF UNIVERSITY DATASET

- This dataset present the traffic volume of a university campus to the Internet during 6 months (2019 Jen-Jul).
- Each line of the dataset represents an hour of recorded data (bandwidth usage in bps)
  - Bps: bit per second
- Dataset shape: (4319, 6)
- Timestamp transform

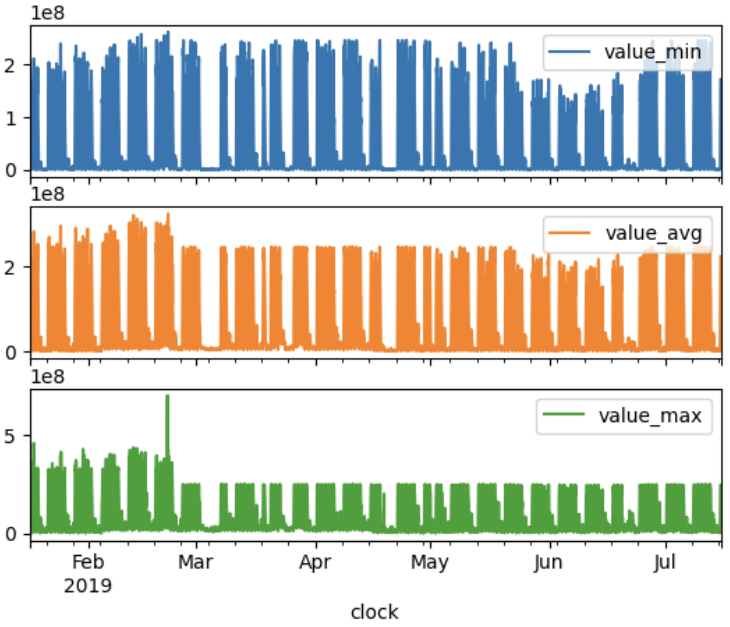
```
df['clock'] = pd.to_datetime(df['clock'], unit='s')
df
```

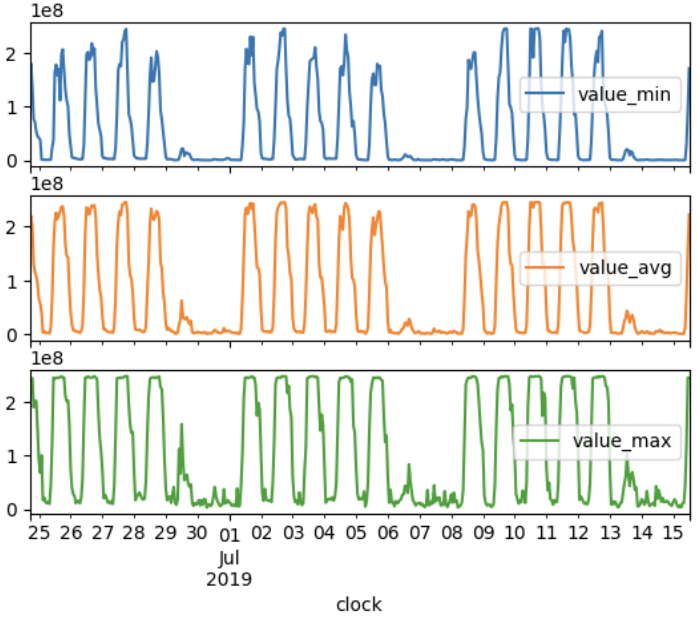
✓ 0.0s Python

	itemid	clock	num	value_min	value_avg	value_max
0	159138	2019-07-15 12:00:00	60	170651112	222446310	245355432
1	159138	2019-07-15 11:00:00	60	92258968	156333926	245517360
2	159138	2019-07-15 10:00:00	60	29854256	73581269	121096792
3	159138	2019-07-15 09:00:00	60	333720	12801917	43525128
4	159138	2019-07-15 08:00:00	60	183976	1146488	11797152

BANDWIDTH USAGE OF UNIVERSITY DATASET

- Data Visualization





REFERENCE

- https://www.tensorflow.org/tutorials/structured_data/time_series
- https://wikidocs.net/120249
- https://pseudo-lab.github.io/Tutorial-Book/chapters/time-series/Ch1-Time-Series.html
- https://paperswithcode.com/task/time-series-forecasting/codeless#datasets
- https://ieee-dataport.org/documents/bandwidth-usage-university-campus

THANK YOU

Q&A