



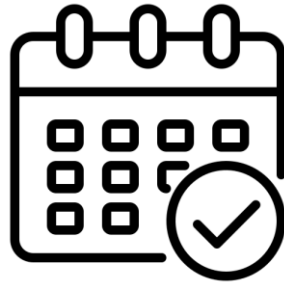
DACON 동서발전 태양광 발전량 예측 AI 경진대회

후반부 캐글/데이콘 2조



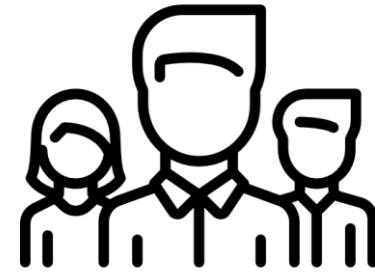
1. 대회 소개

동서발전 태양광 발전량 예측
AI 경진대회



2. 워크로드 소개

대회 일정 / 진행 상황



3. 팀원 소개

12기 이두형, 이나윤
13기 기다연, 박재찬

01. 대회 소개

동서발전 태양광 발전량 예측 AI 경진대회

주제: 시간대별 태양광 발전량 예측

배경:

태양광 발전은 매일 기상 상황과 계절에 따른
일사량의 영향을 받습니다.

이에 대한 예측이 가능하다면 보다 원활한
전력 수급 계획이 가능합니다.

인공지능 기반 태양광 발전량 예측 모델을 만들어주세요.



01. 대회 소개

대회 주요 일정



01. 대회 소개

✓ site_info.csv _ 발전소 정보

Id : 사이트 식별자
Capacity : 발전소 발전용량(MW)
Address : 주소
InstallationAngle : 설치각(°)
IncidentAngle : 입사각(°)
Latitude : 위도
Longitude : 경도

✓ ulsan_obs_data.csv _ 울산지역 발전소 인근 기상 관측 자료

지점 : 지점 코드
지점명 : 관측소 지점
일시 : 관측 시간
기온(°C) : 기온(°C)
풍속(m/s) : 풍속(m/s)
풍향(16방위) : 풍향(°)
습도(%) : 습도(%)
전운량(10분위) : 전운량(낮을 수록 구름이 적음)

✓ ulsan_fcst_data.csv _ 울산지역 발전소 동네 예보

Forecast time : 예보 발표 시점
forecast : 예보 시간
(Forecast time:2018-03-01 11:00:00, forecast:4.0
→ 2018-03-01 11:00:00에 발표한 2018-03-01 15:00:00 예보)

Temperature : 온도(°C)
Humidity : 습도(%)
WindSpeed : 풍속(m/s)
WindDirection : 풍향(°)
Cloud : 하늘상태(1-맑음, 2-구름보통, 3-구름많음, 4-흐림)

✓ energy.csv _ 발전소별 발전량

time : 1시간 단위 계량된 시간 (ex-2018-03-01 1:00:00 => 2018-03-01 00:00:00 ~ 2018-03-01 1:00:00 1시간동안 발전량 계량)
dangjin_floating : 당진수상태양광 발전량(KW)
dangjin_warehouse : 당진자재창고태양광 발전량(KW)
dangjin : 당진태양광 발전량(KW)
ulsan : 울산태양광 발전량(KW)

✓ dangjin_obs_data.csv _ 당진지역 발전소 인근 기상 관측 자료

지점 : 지점 코드
지점명 : 관측소 지점
일시 : 관측 시간
기온(°C) : 기온(°C)
풍속(m/s) : 풍속(m/s)
풍향(16방위) : 풍향(°)
습도(%) : 습도(%)
전운량(10분위) : 전운량(낮을 수록 구름이 적음)

✓ dangjin_fcst_data.csv _ 당진지역 발전소 동네 예보

Forecast time : 예보 발표 시점
forecast : 예보 시간
(Forecast time:2018-03-01 11:00:00, forecast:4.0
→ 2018-03-01 11:00:00에 발표한 2018-03-01 15:00:00 예보)

Temperature : 온도(°C)
Humidity : 습도(%)
WindSpeed : 풍속(m/s)
WindDirection : 풍향(°)
Cloud : 하늘상태(1-맑음, 2-구름보통, 3-구름많음, 4-흐림)

01. 대회 소개

✓ sample_submission.csv _ 예측한 발전량 제출 양식

public LB : 2021년 2월 예측

private LB : 2021년 6월 9일 ~ 2021년 7월 8일 30일간 예측, 평가기간 제출 가능, 예측 전날 선택된 제출물 평가

time : 지난 한시간동안 발전량 예측

dangjin_floating : 당진수상태양광 예측 발전량(KW)

dangjin_warehouse : 당진자재창고태양광 예측 발전량(KW)

ulsan : 울산태양광 예측 발전량(KW)

	time	dangjin_floating	dangjin_warehouse	dangjin	ulsan
0	2021-02-01 01:00:00	0	0	0	0
1	2021-02-01 02:00:00	0	0	0	0
2	2021-02-01 03:00:00	0	0	0	0
3	2021-02-01 04:00:00	0	0	0	0
4	2021-02-01 05:00:00	0	0	0	0
5	2021-02-01 06:00:00	0	0	0	0
6	2021-02-01 07:00:00	0	0	0	0
7	2021-02-01 08:00:00	0	0	0	0
8	2021-02-01 09:00:00	0	0	0	0
9	2021-02-01 10:00:00	0	0	0	0

```
import pandas as pd
import numpy as np

def sola_nmae(answer_df, submission_df):
    submission = submission_df[submission_df['time'].isin(answer_df['time'])]
    submission.index = range(submission.shape[0])

    # 시간대별 총 발전량
    sum_submission = submission.iloc[:,1:].sum(axis=1)
    sum_answer = answer_df.iloc[:,1:].sum(axis=1)

    # 발전소 발전용량
    capacity = {
        'dangjin_floating':1000, # 당진수상태양광 발전용량
        'dangjin_warehouse':700, # 당진자재창고태양광 발전용량
        'dangjin':1000, # 당진태양광 발전용량
        'ulsan':500 # 울산태양광 발전용량
    }

    # 총 발전용량
    total_capacity = np.sum(list(capacity.values()))

    # 총 발전용량 절대오차
    absolute_error = (sum_answer - sum_submission).abs()

    # 발전용량으로 정규화
    absolute_error /= total_capacity

    # 총 발전용량의 10% 이상 발전한 데이터 인덱스 추출
    target_idx = sum_answer[sum_answer>=total_capacity*0.1].index

    # NMAE(%)
    nmae = 100 * absolute_error[target_idx].mean()

    return nmae
```

02. 워크로드 소개

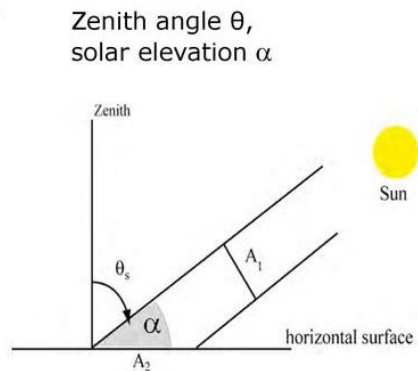
방향 설정 (5월 3일~ 5월 14일)

: EDA, 태양광 발전 관련 지표(GHI, DHI, DNI),
외부 데이터 추출, 데이터 전처리

Direct, Diffuse and Global Irradiance

When measuring solar irradiance, the following components are of particular interest:

- ❑ **Direct normal irradiance (DNI)**
(also: beam irradiance)
- ❑ **Diffuse horizontal irradiance (DHI)**
(also: diffuse sky radiation)
- ❑ **Global horizontal irradiance (GHI)**
(also: total solar irradiance)



$$GHI = DHI + DNI * \sin(\alpha)$$

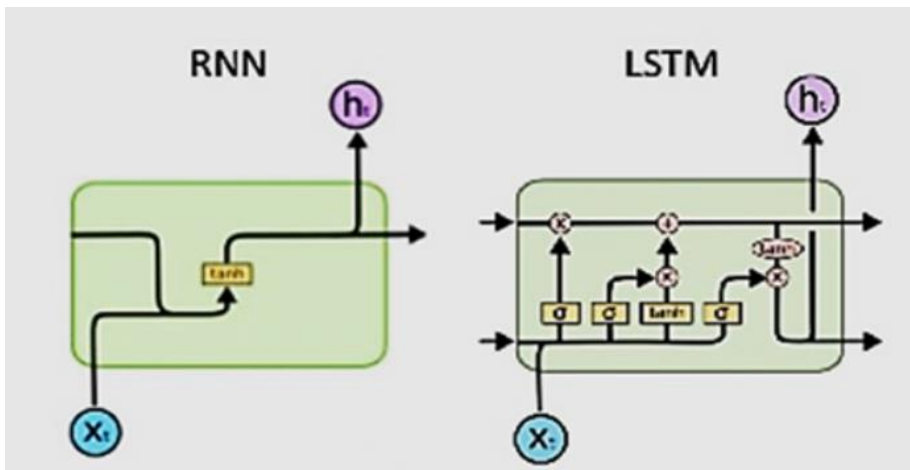
2021년 5월

SUN	MON	TUE	WED	THU	FRI	SAT
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5
6	7	8				

02. 워크로드 소개

모델 만들기 (5월 19일~ 6월 4일)

: 태양광 발전량 알고리즘 관련 논문 읽기,
RNN, LSTM 등의 시계열 예측 모델



5월 30일: 팀 병합

6월 8일: Public 평가 종료일

2021년 5월

SUN	MON	TUE	WED	THU	FRI	SAT
						1
2	3	4	5	6	7	8
9	10	11	12	13	14	15
16	17	18	19	20	21	22
23	24	25	26	27	28	29
30	31	1	2	3	4	5
6	7	8				

02. 워크로드 소개

태양광 발전량 예측 모델

(1) ARIMA

(1) ARIMA

ARIMA(Auto-Regressive Integrated Moving Average)는 auto-regressive(AR) 모델과 moving average(MA)

모형을 동시에 포함하는 시간의 흐름에 따른 연속값들의 확률구조를 고려한 모델로, 귀차니성 p , 차분 차수 d , 이동평균 차수 q 를 가진다. 시계열 Z_t 가 $ARIMA(p, d, q)$ 모형으로 표현된다.

$$\phi_p(B)(1-B)^d Z_t = \theta_q(B)(1-B)^d Z_t, \\ \text{where, } \phi_p(B) = (1 - \phi_1 B - \dots - \phi_p B^p), \\ \theta_q(B) = (1 - \theta_1 B - \dots - \theta_q B^q), \\ \theta_0: \text{the deterministic trend}$$

p : 자기회귀(AR) 차수로서, t 시점의 값에 $(t-1), \dots, (t-p)$ 시점 값이 영향을 준다.
 d : 차분을 의미하며, t 시점의 값에 $(t-d)$ 시점의 값을 빼서 비정상 시계열을 정상화한다.
 q : 이동평균(MA) 차수로서, t 시점의 값에 $(t-1), \dots, (t-q)$ 시점의 연속적인 오차의 평균을 더한다.

SARIMA 모델을 이용한 예보 모형 구축 Solar Power Generation Forecast ARIMA

<http://eds.b.ebscohost.com/ssl.cpa.korea.ac.kr/eds/detail/detail?vid=38bc95-6b91f0db2af640pdc-v-sssmg01&bddata=jmxbmc9a28mc202T11Z4H4MtbG12Z5ZzY29wZT1zaS8&db=edsici>

(2) ARIMAX

2.2. ARIMAX 모형

ARIMAX 모형은 auto-regressive integrated moving average model (ARIMA) 기반 모형이다. 평균이 μ 이고 ARIMA의 차수가 p, d, q 일 때 외생변수 x_{it} 가 있는 식은 다음과 같다.

$$(1 - B^d) (Y_t - \mu) = \frac{\theta(B)}{\phi(B)} \epsilon_t + \sum_{i=1}^k \beta_i x_{it}, \\ \text{where } \phi(B) = 1 - \phi_1 B - \phi_2 B^2 - \dots - \phi_p B^p, \\ \theta(B) = 1 - \theta_1 B - \theta_2 B^2 - \dots - \theta_q B^q$$

여기, $\phi(B)$ 는 auto-regressive (AR)에 관한 다항식이며, $\theta(B)$ 는 moving-average이다. B 는 후진연산자(backward shift operator)이다. Y_t 는 t 시점 ($t = 1, 2, \dots$)

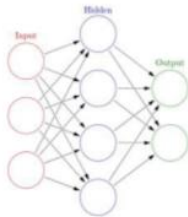


Figure 2.1. Artificial neural network.

시계열 예측 모델 방법론

11 Classical Time Series Forecasting Methods in Python (Cheat Sheet)
Machine learning methods can be used for classification and forecasting on time series problems. Before exploring machine learning methods for time series, it is important to understand the basic concepts of time series forecasting. <https://machinelearningmastery.com/time-series-forecasting-methods-in-python/>



1) Autoregression (AR)

: 자기 자신의 과거 데이터/관측값 사용 - 이후 자신의 관측값에 영향을 주는 모형

- Whitening으로 임의의 노이즈 삽입 (불확실성 주입)
- 장점: 최신 과거에 기반한 변동 패턴 파악하기에 적합
- 단점: 트렌드가 위아래로 변동하는 경우에는 부적합

2) Moving Average (MA)

: 트렌드가 변화하는 상황에서 적합한 모형

3) Autoregressive Moving Average (ARMA) / Seasonal (SARMA)

: 과거의 상태와 오차값을 사용해 현재의 상태를 예측하는 모형

- AR+MA를 섞어놓은 모형 (분석의 정확도 상승)
- 단점: 불규칙적 시계열 데이터를 제대로 예측하지 못함

4) Autoregressive Integrated Moving Average (ARIMA) / Seasonal (SARIMA)

: ARMA + 관측치 사이에 차분(difference) 개념 사용해 불규칙적 시계열 데이터를 규칙적으로 만들기 위한 수 있도록 변환함

- 차분: 현재 상태 변수 - 바로 전 상태 변수 - Whitening의 효과

5) Deep Neural Network (DNN)

6) Recurrent Neural Network (RNN)

: 순환신경망(반복적, 순차적인 시계열 데이터 학습에 특화된 인공신경망)

- 짧은 기간의 정보 바탕으로 학습

7) K-Nearest Neighbors (KNN): K-근접 신경망

: 데이터 기준으로 K개의 가까운 점을 찾아 평균값

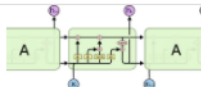
8) Long Short-Term Memory Network (LSTM)

: RNN의 변형 버전

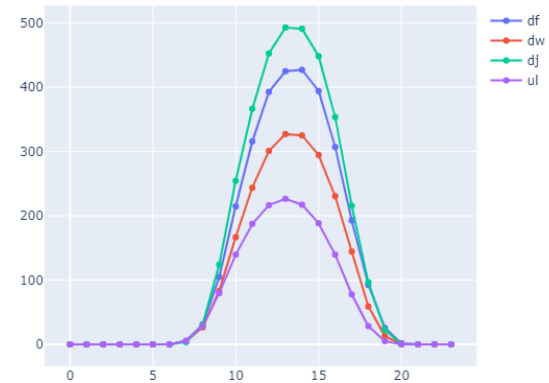
- RNN의 단점 보완 (긴 시간에도 의존(long-term dependencies))

Long Short-Term Memory (LSTM) 이해하기

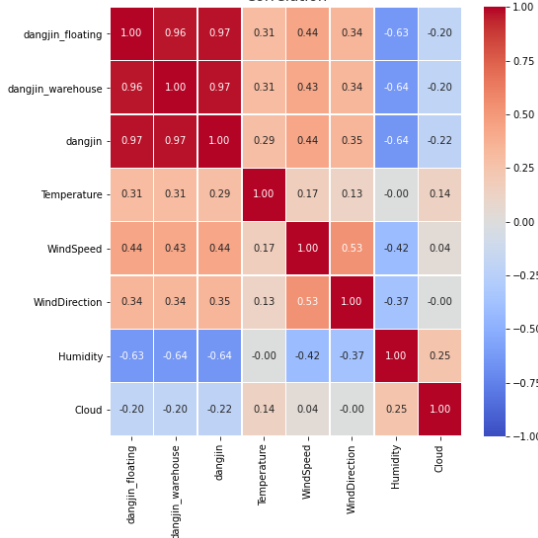
이 글은 Christopher Olah가 2015년 8월에 쓴 구글 수리 발표 번역한 것이다.
Recurrent neural network의 개념을 쉽게 설명했고, 그 중 특기적인 모델인 LSTM
T: <https://dgkim5360.tistory.com/entry/understanding-long-short-term-memory>



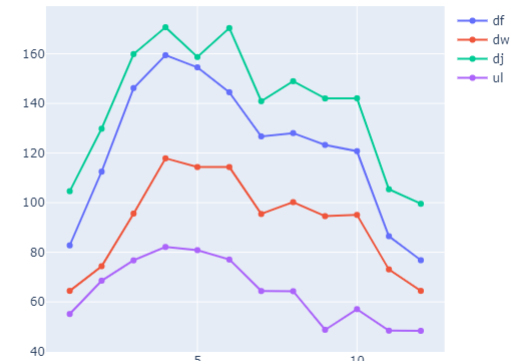
HOUR_energy



Correlation



MONTH_energy



03. 팀원 소개



12기

이두형 (수학과 18)

이나운 (통계학과 19)

13기

기다연 (통계학과 19)

박재찬 (의예과 20)

An aerial photograph of a city at dusk or dawn. The city is situated along a large body of water, with mountains visible in the background. The foreground shows a large parking lot with several cars and a building with a flat roof. The text "THANK YOU" is overlaid in the center of the image.

THANK YOU