APCC
APEC 기후센터

# 기계학습을 이용한 우리나라 여름철 장기기온예측

2022년 10월 28일

APCC 기후정보서비스 사용자워크숍

예측기술과
이진영 선임연구원


# 오늘의 학습

## I. 기계학습이란

- 인공지능이란
- 기계학습이란
- 기계학습의 종류

## II. 기계학습을 이용한 장기예측

- 장기예측을 위한 준비
- 다양한 기계학습을 이용한 기온예측 – Toy Examples

APCC
APEC 기후센터

# I. 기계학습이란

- 인공지능이란
- 기계학습이란
- 기계학습의 종류

# I. 기계학습이란

## 인공지능이란

"인간의 지능을 필요로 하는 일들을 컴퓨터나 컴퓨터에 의해 제어되는 로봇이 수행하는 능력"

http://movie.naver.com/movie/bi/mi/basic.nhn?code=10200

터미네이터 2 1984~

미래, 인류와 기계의 전쟁은 계속 되는 가운데 스카이넷은 인류 저항군 사령관 존 코너를 없애기 위해 ...

movie.naver.c

http://movie.naver.com/movie/bi/mi/basic.nhn?code=24452

매트릭스 1999

2199년. 인공 두뇌를 가진 컴퓨터(AI: Artificial Intelligence)가 지배하는 세계. 인간들은 태어나자마...

movie.

http://movie.naver.com/movie/bi/mi/basic.nhn?code=38420

아이, 로봇 2004

{법칙 1. 로봇은 인간을 다치게 해선 안되며, 행동하지 않음으로써 인간이 다치도록 방관해서도 안된다....

movie.naver.com

출처: 네이버 영화

https://movie.naver.com/movie/bi/mi/basic.nhn?code=69105

월-E 2008

텅 빈 지구에 홀로 남아 수백 년이라 시간을 이루게...

movie.naver.com

http://movie.naver.com/movie/bi/mi/basic.nhn?code=101950

그녀 2013

'테오도르'(호아킨 피닉스)는 다른 사람들의 편지를 대신 써주는 대필 작가로, 아내(루니 마라)와 별...

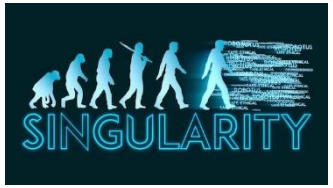
http://movie.naver.com/movie/bi/mi/basic.nhn?code=118361

엑스 마키나 2014

유능한 프로그래머 '칼렙'(돔놀 글리슨)은 치열한 경쟁 끝에 인공지능 분야의 천재 개발자 '네이든'...

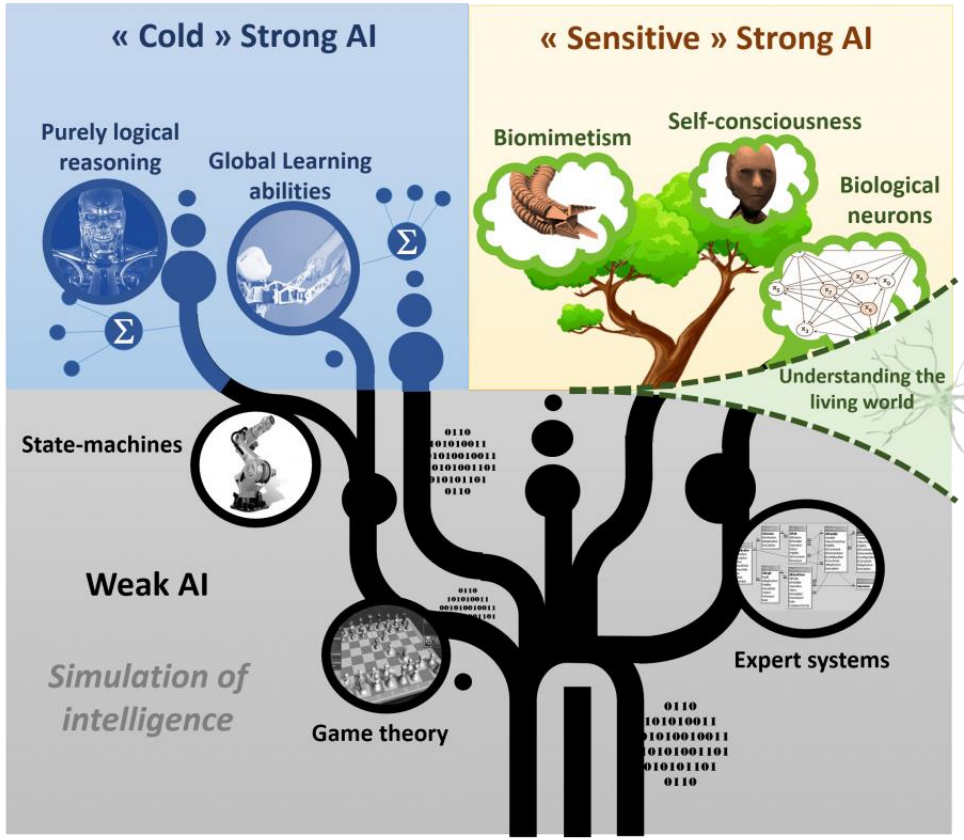
movie.naver.com

출처: 네이버 영화

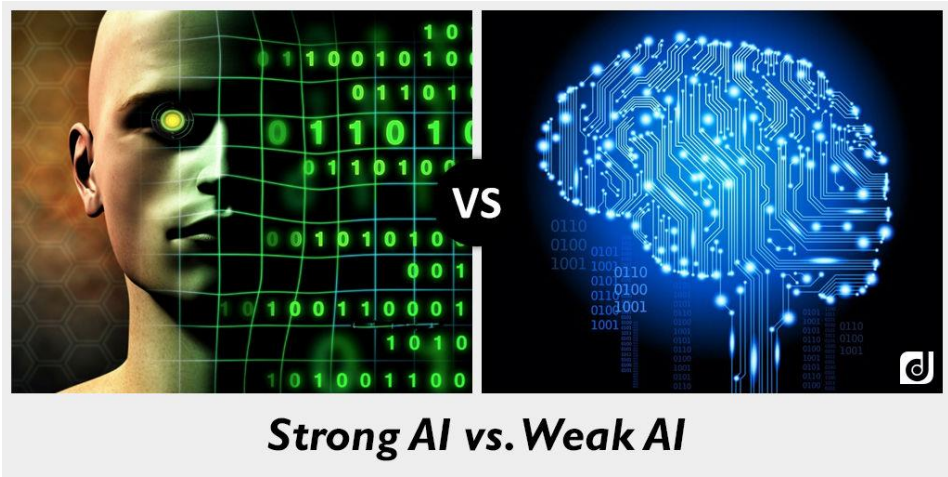


# I. 기계학습이란

## 인공지능이란




이미지 출처: Théophile Gonos



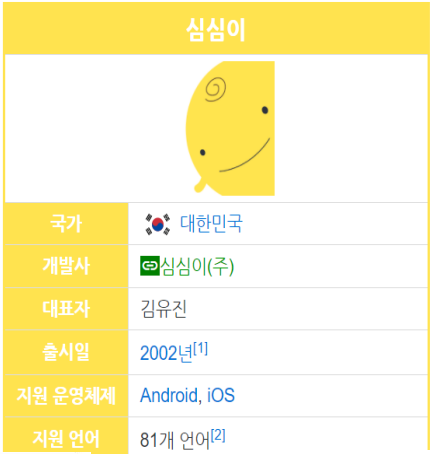

이미지 출처: Difference Between

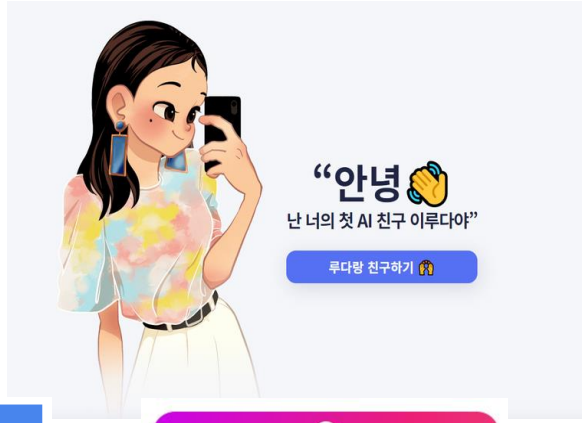
**Strong AI**는 인간을 모방하는 것 뿐 아니라 인간과 유사한 지능을 갖추고 인간처럼 사고하고 행동

**Weak AI**는 인간이 미리 계획한 일을 능숙히 수행하기 위한 기술을 개발하기 위해서 사고

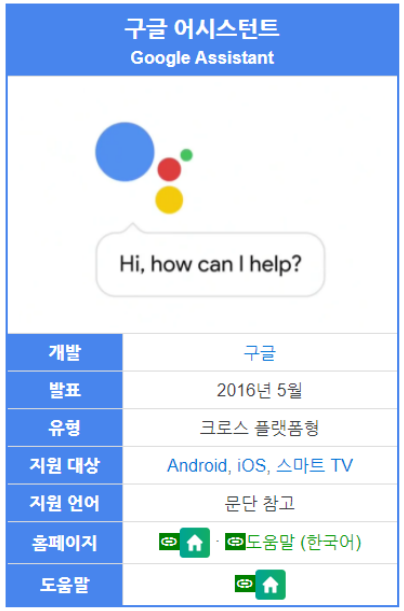
# I. 기계학습이란

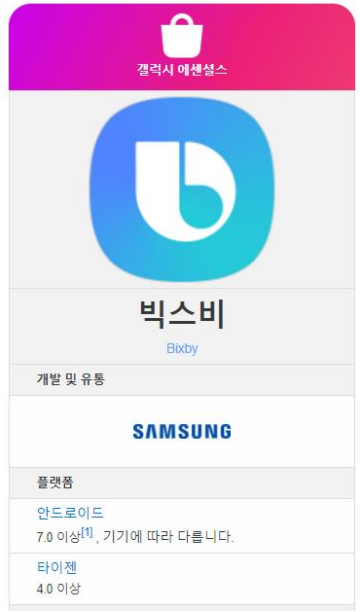
## 인공지능이란





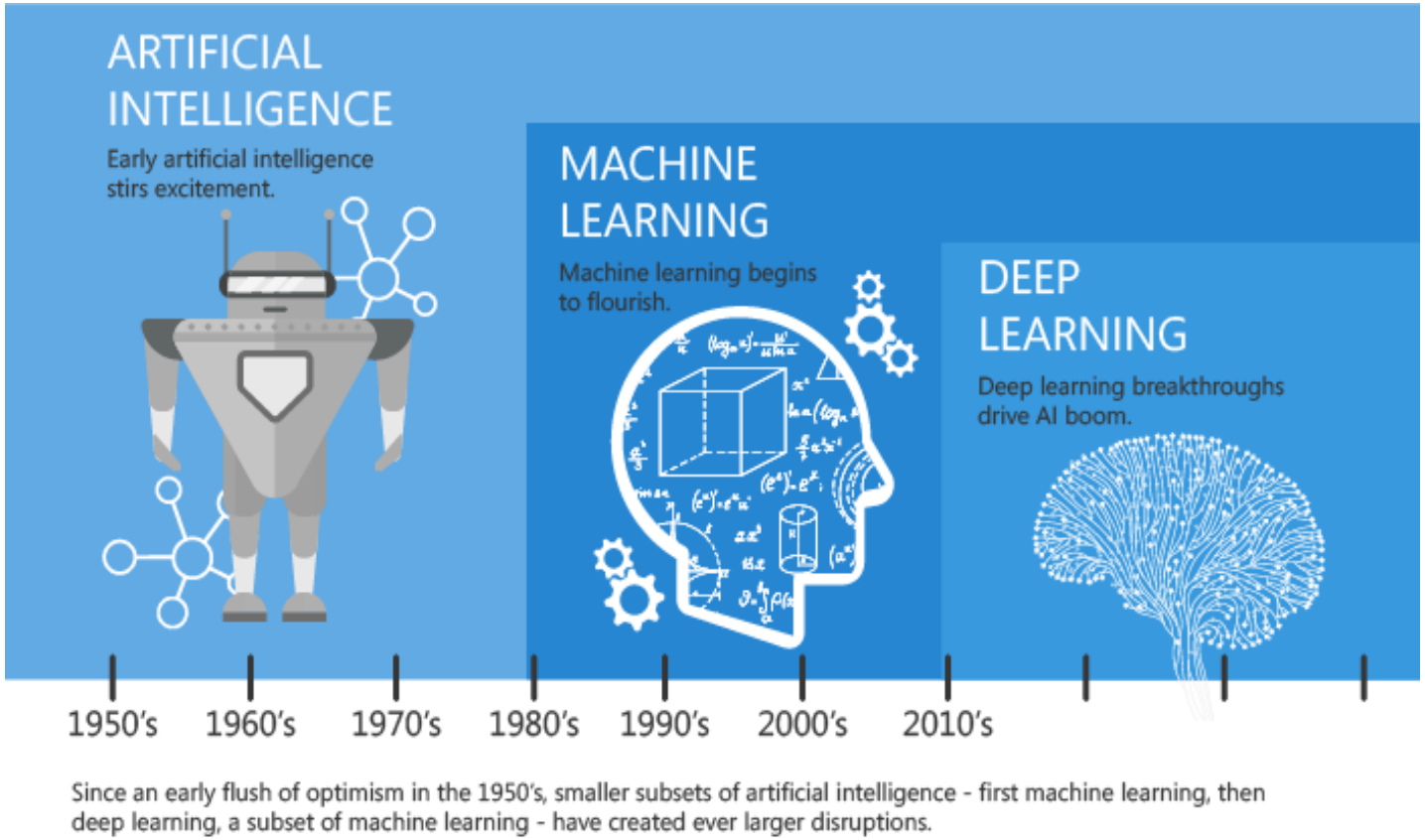




알렉사


앱 스토어와 구글 플레이에서의 로고

# I. 기계학습이란

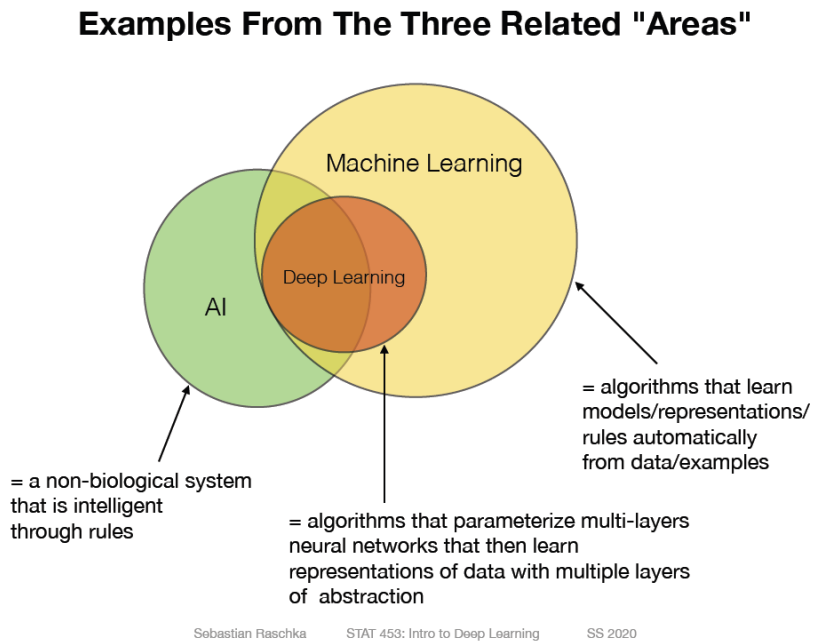
## 기계학습이란




출처: NVIDIA 블로그

처리과정을 학습하는 컴퓨터 모델링: 컴퓨터로 하여금 귀납이나 연역과 같은 특정 추론 전략을 사용하여 현존하는 자료나 이론으로부터 지식을 획득하게 함

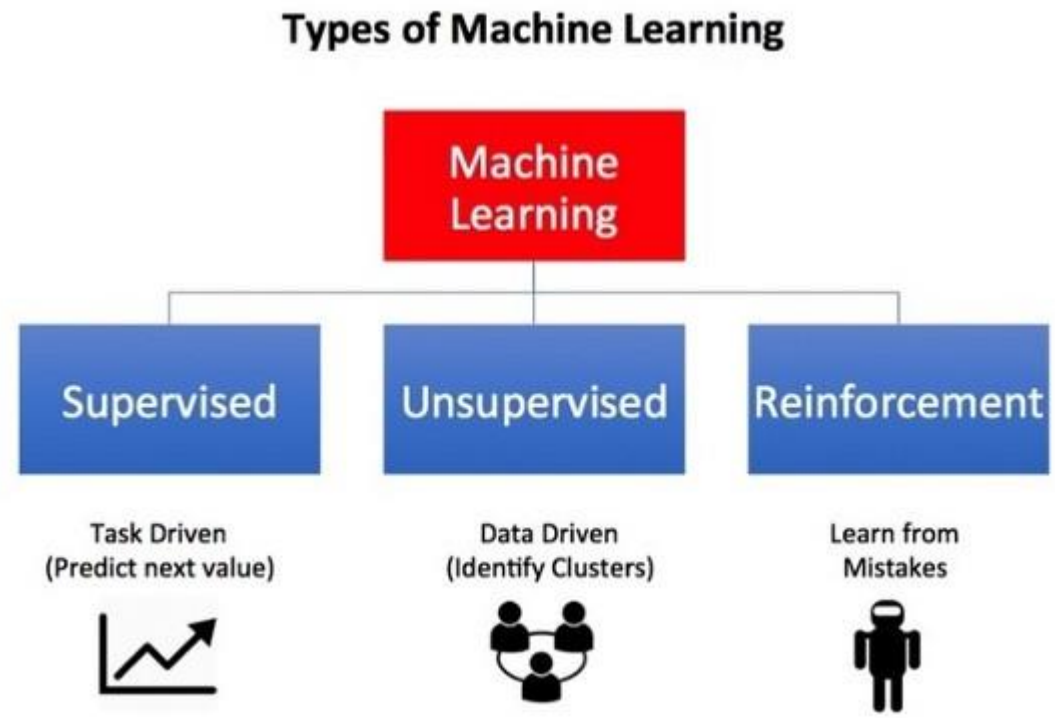
출처: Jensen, 2016



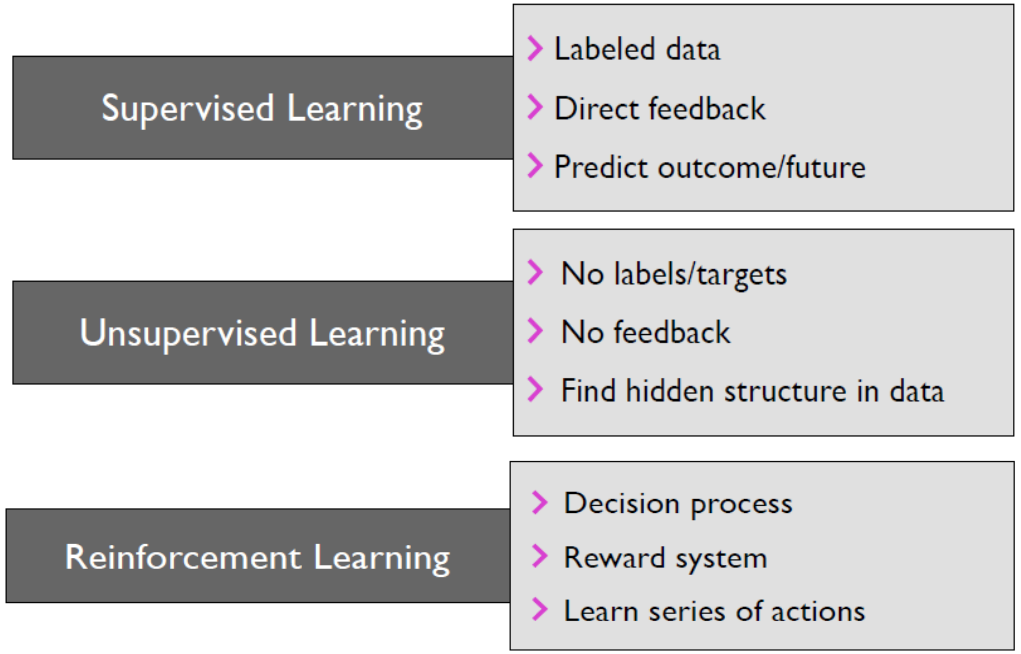
# I. 기계학습이란

## 기계학습의 종류

지도학습 비지도학습 강화학습




출처: towardsdatascience

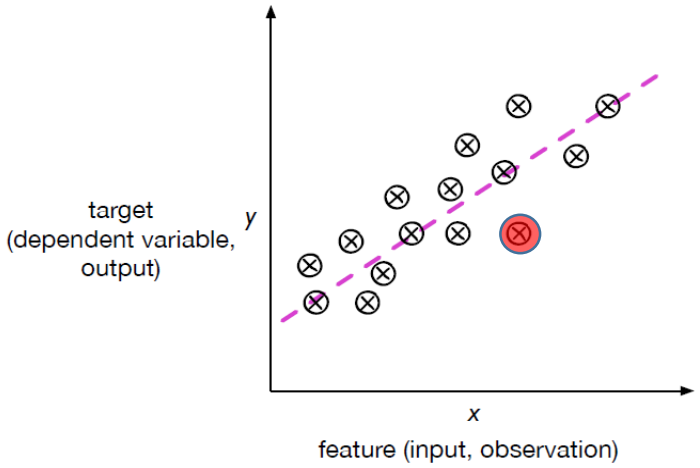



출처: Raschka and Mirjalily, 2019

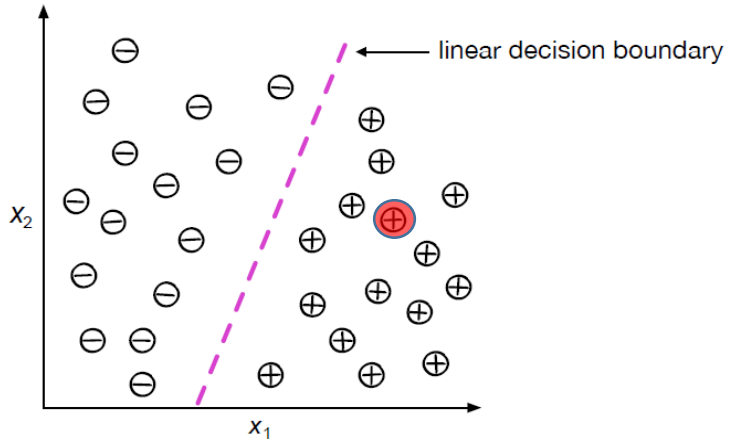
## 기계학습의 종류

### 지도학습

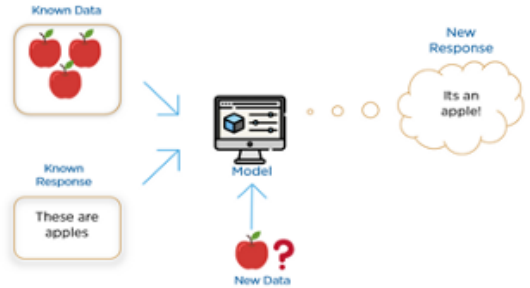
회귀




분류 / 패턴인식




출처: Raschka and Mirjalily, 2019

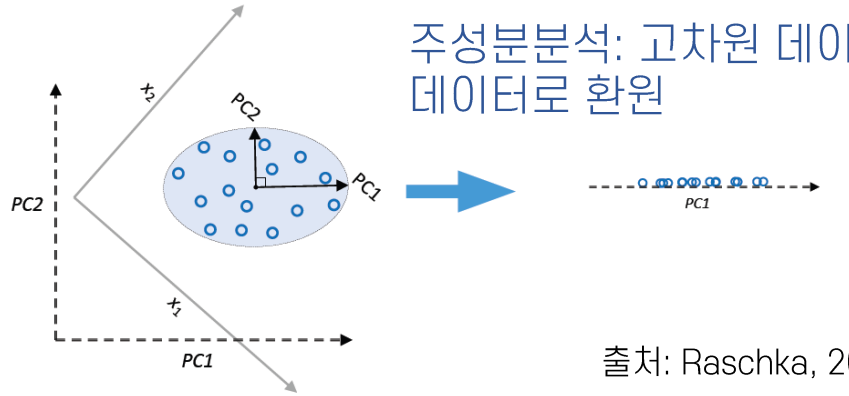



출처: towardsdatascience

## 기계학습의 종류

### 비지도학습

#### 차원 감소



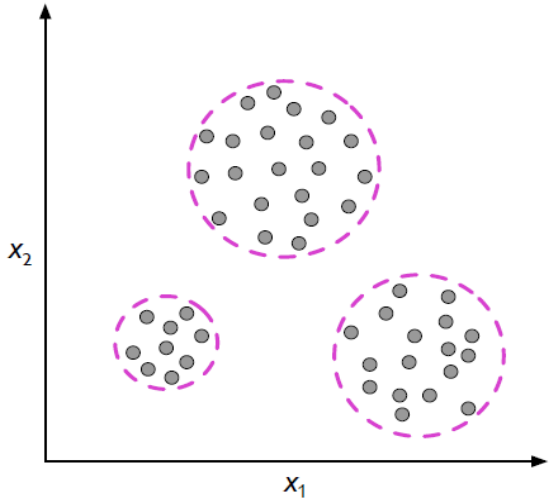

주성분분석: 고차원 데이터를 저차원 데이터로 환원

출처: Raschka, 2020


<그림 6> k개의 eigenface만을 이용한 데이터 복원(reconstruction)

출처: darkpgmr 블로그

#### 군집



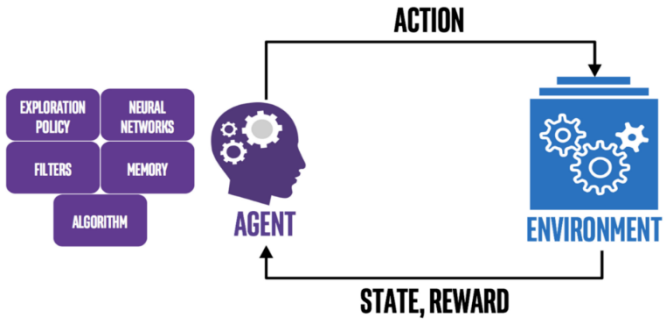

출처: Raschka, 2020


출처: towardsdatascience

# I. 기계학습이란

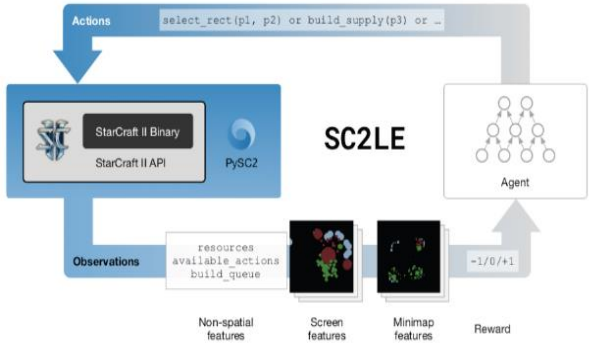
## 기계학습의 종류

### 강화학습




출처: medium.com


Starcraft II




Vinyals, Oriol, Timo Ewalds, Sergey Bartunov, Petko Georgiev, Alexander Sasha Vezhnevets, Michelle Yeo, Alireza Makhzani et al. "Starcraft II: A new challenge for reinforcement learning." *arXiv preprint arXiv:1708.04782* (2017).


어떤 환경 안에서 정의된 에이전트(행위자)가 현재의 상태를 인식하여 선택 가능한 행동들 중 보상을 최대화하는 행동을 선택하는 방법

APCC
APEC 기후센터

# II. 기계학습을 이용한 장기예측

- 장기예측을 위한 준비
- 다양한 기계학습을 이용한 기온예측 - Toy Examples

# II. 기계학습을 이용한 장기예측

## 장기예측을 위한 준비

# II. 기계학습을 이용한 장기예측

## 풀고 싶은 문제를 정의하자: 단정예측


기온 값 자체를 예측
→ 단정예측

**5월(ex. 5/15)에 예측한 6, 7, 8월 평균 기온 예측**

삼분위 카테고리로 분류
→ 각 카테고리별 확률
→ 확률예측

확률분포 예측(베이지안)
→ 각 카테고리별 확률

5월(ex. 5/15)에 예측한 6, 7, 8월 평균 기온 확률 예측

# II. 기계학습을 이용한 장기예측

## 자료를 수집하자


입력자료 X

관측자료 y

- 훈련/검증/테스트 기간의 6, 7, 8월 평균기온

→ 전국 62개 ASOS 지점 자료 평균으로 산정

# II. 기계학습을 이용한 장기예측

## 자료를 수집하자


**입력자료 X**

**관측자료 y**

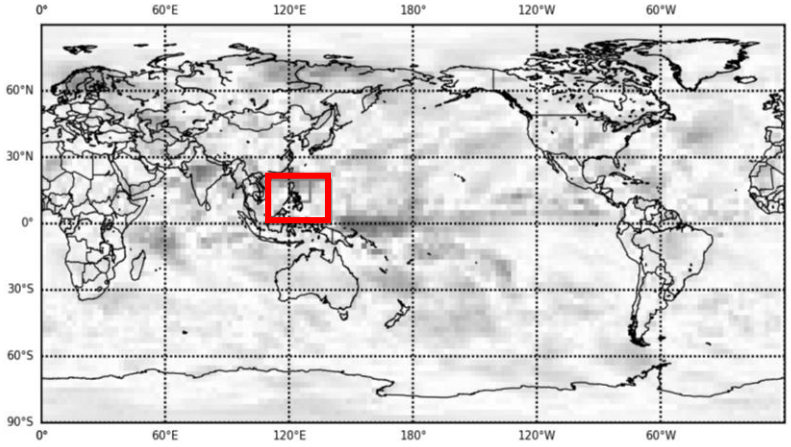
- Outgoing Long wave Radiation (OLR)
- Sea Surface Temperature (SST)
- Precipitation (PRCP)
- Snow Cover Extent (SCE)
- 500 hPa & 850 hPa Geopotential Height (Z500, Z850)
- Sea Ice Area (SIA)

# II. 기계학습을 이용한 장기예측

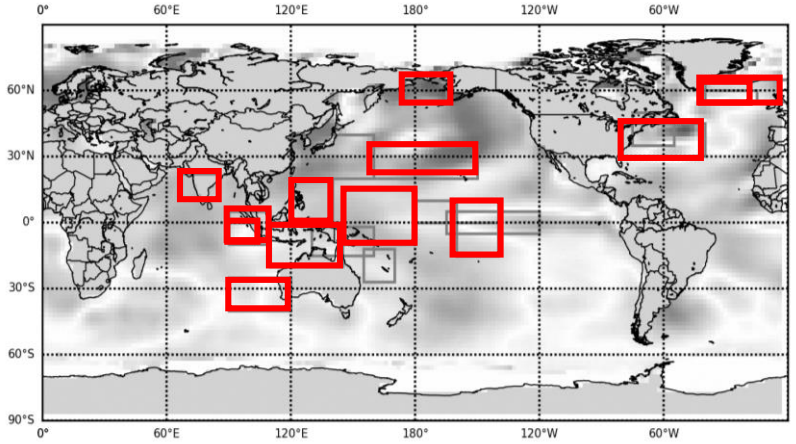
## 자료를 수집하자

- 모델에서 사용할 입력변수들: 장기예보 가이던스 III(기상청, 2018) 활용
  - 기후감시요소의 수가 적으므로 장마기간 강수, 6~8월 기온에 영향을 미치는 요소 모두 사용
  - 쌍극, 삼극 패턴의 경우 개별 및 차이 모두 사용, 전년도 12월 ~ 당해 년도 4월 자료 모두 사용

  >> 약 30여개의 기후감시요소 도출




5월 북서태평양 지역 OLR 편차



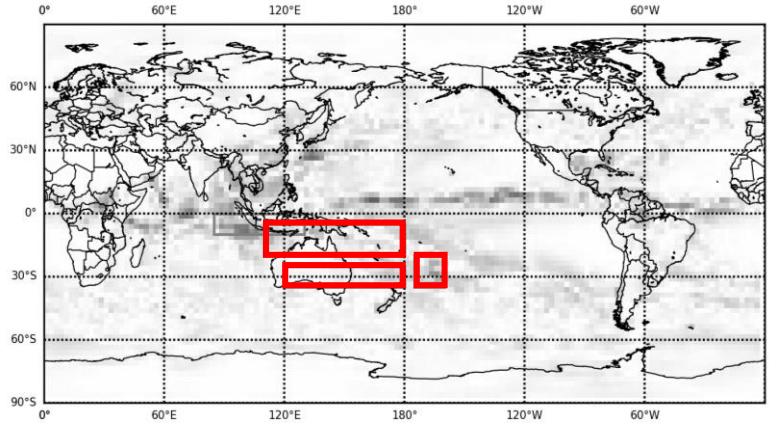

4월 북서태평양 SST 편차 지역차
4월 북태평양 지역 SST 편차(시기차)
4월 열대중태평양 지역 SST 편차
4월 북인도양 지역 SST 편차
3월 베링해 SST 편차
1월 열대서태평양 SST 편차
4월 동인도양~필리핀동쪽 SST 편차 삼극
2월 호주북쪽 지역 SST 편차
4월 열대인도양 지역 SST 편차
4월 북대서양 SST 편차

# II. 기계학습을 이용한 장기예측

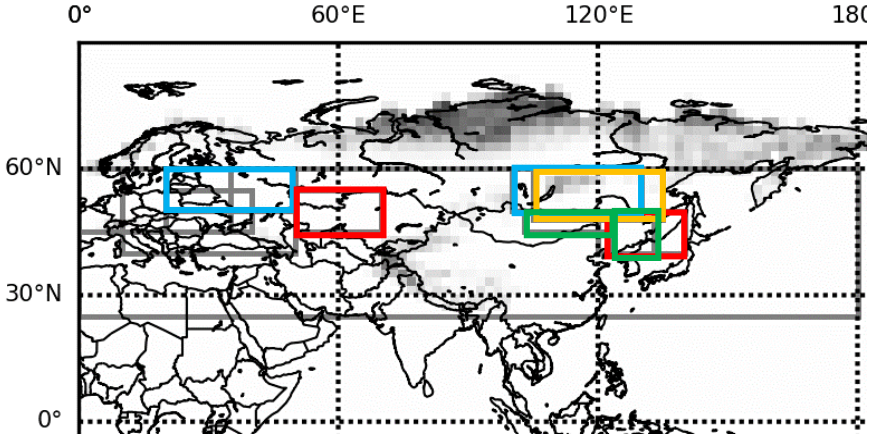
## 자료를 수집하자

- 모델에서 사용할 입력변수들: 장기예보 가이던스 III(기상청, 2018) 활용
  - 기후감시요소의 수가 적으므로 장마기간 강수, 6~8월 기온에 영향을 미치는 요소 모두 사용
  - 쌍극, 삼극 패턴의 경우 개별 및 차이 모두 사용, 전년도 12월 ~ 당해 년도 4월 자료 모두 사용

  >> 약 30여개의 기후감시요소 도출




4월 호주부근 PRCP 편차 지역차
3월 호주북쪽 지역 PRCP 편차




3~5월 유라시아 눈덮임 편차 지역차
3월 중국 북동부 지역 눈덮임 편차
4월 동아시아 / 서아시아 눈덮임 편차 지역차
4월 만주 눈덮임 편차

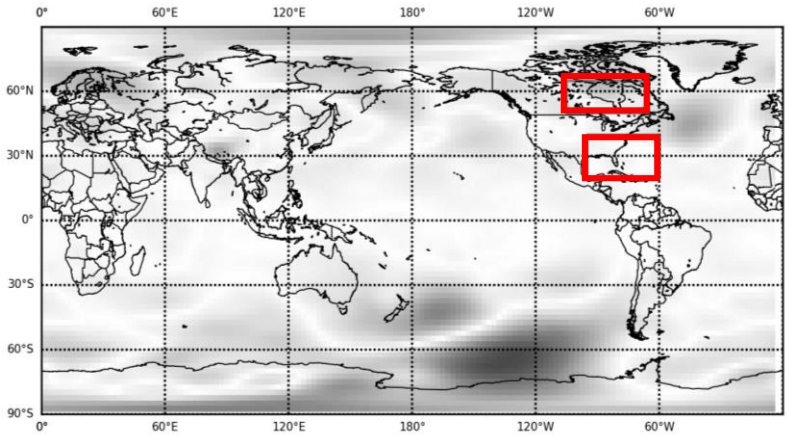
# II. 기계학습을 이용한 장기예측

## 자료를 수집하자

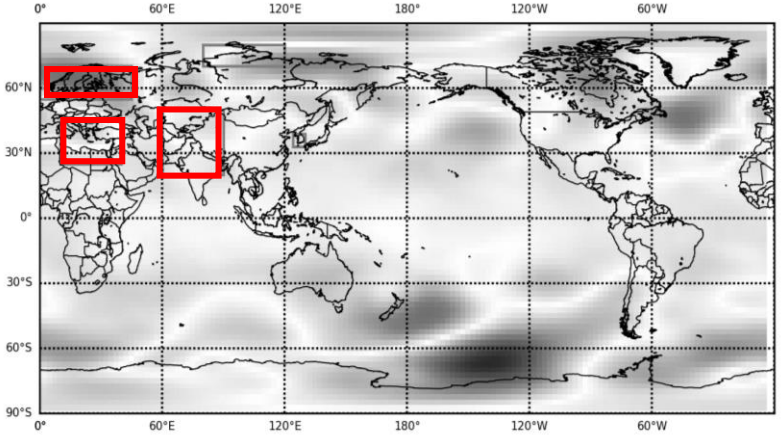
- 모델에서 사용할 입력변수들: 장기예보 가이던스 III(기상청, 2018) 활용
  - 기후감시요소의 수가 적으므로 장마기간 강수, 6~8월 기온에 영향을 미치는 요소 모두 사용
  - 쌍극, 삼극 패턴의 경우 개별 및 차이 모두 사용, 전년도 12월 ~ 당해 년도 4월 자료 모두 사용

  >> 약 30여개의 기후감시요소 도출

3~4월 베링해 해빙 편차




4월 대서양 지역 850 hPa 지위고도 편차 쌍극



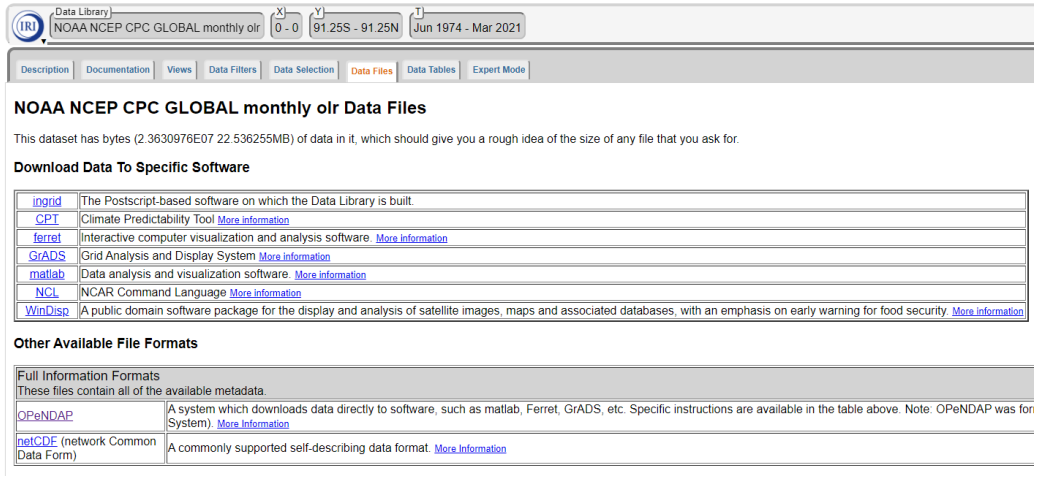

3월 유라시아 지역 500 hPa 지위고도 편차 쌍-삼극

# II. 기계학습을 이용한 장기예측

## 자료를 수집하자

- 모델에서 사용할 입력변수들: 장기예보 가이던스 III(기상청, 2018) 활용
  - 각 기후감시요소에 대하여 시계열 자료 산정

### 자료 수집


Data Library | NOAA NCEP CPC GLOBAL monthly olr | X: 0 - 0 | Y: 91.25S - 91.25N | T: Jun 1974 - Mar 2021

Description | Documentation | Views | Data Filters | Data Selection | Data Files | Data Tables | Expert Mode

**NOAA NCEP CPC GLOBAL monthly olr Data Files**

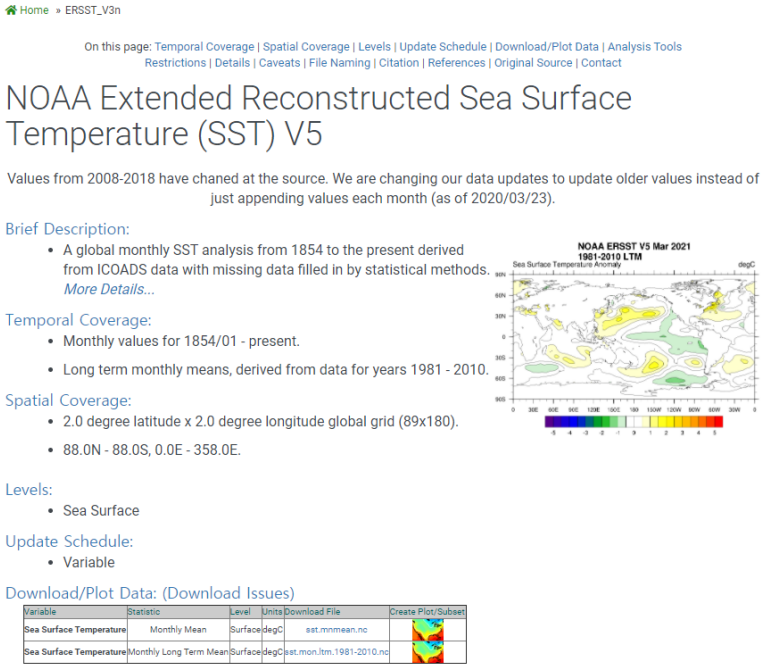
This dataset has bytes (2.3630976E07 22.536255MB) of data in it, which should give you a rough idea of the size of any file that you ask for.

**Download Data To Specific Software**

| | |
|---|---|
| ingrid | The Postscript-based software on which the Data Library is built. |
| CPT | Climate Predictability Tool More information |
| ferret | Interactive computer visualization and analysis software. More information |
| GrADS | Grid Analysis and Display System More information |
| matlab | Data analysis and visualization software. More information |
| NCL | NCAR Command Language More information |
| WinDisp | A public domain software package for the display and analysis of satellite images, maps and associated databases, with an emphasis on early warning for food security. More information |

**Other Available File Formats**

| Full Information Formats<br>These files contain all of the available metadata. | |
|---|---|
| OPeNDAP | A system which downloads data directly to software, such as matlab, Ferret, GrADS, etc. Specific instructions are available in the table above. Note: OPeNDAP was for System). More Information |
| netCDF (network Common Data Form) | A commonly supported self-describing data format. More Information |


Home » ERSST_V3n

On this page: Temporal Coverage | Spatial Coverage | Levels | Update Schedule | Download/Plot Data | Analysis Tools | Restrictions | Details | Caveats | File Naming | Citation | References | Original Source | Contact

NOAA Extended Reconstructed Sea Surface Temperature (SST) V5

Values from 2008-2018 have chaned at the source. We are changing our data updates to update older values instead of just appending values each month (as of 2020/03/23).

Brief Description:
- A global monthly SST analysis from 1854 to the present derived from ICOADS data with missing data filled in by statistical methods. *More Details...*

NOAA ERSST V5 Mar 2021
1981-2010 LTM
Sea Surface Temperature Anomaly | degC

Temporal Coverage:
- Monthly values for 1854/01 - present.
- Long term monthly means, derived from data for years 1981 - 2010.

Spatial Coverage:
- 2.0 degree latitude x 2.0 degree longitude global grid (89x180).
- 88.0N - 88.0S, 0.0E - 358.0E.

Levels:
- Sea Surface

Update Schedule:
- Variable

Download/Plot Data: (Download Issues)

| Variable | Statistic | Level | Units | Download File | Create Plot/Subset |
|---|---|---|---|---|---|
| Sea Surface Temperature | Monthly Mean | Surface | degC | sst.mnmean.nc | |
| Sea Surface Temperature | Monthly Long Term Mean | Surface | degC | sst.mon.ltm.1981-2010.nc | |

# II. 기계학습을 이용한 장기예측

## 자료를 수집하자

- 모델에서 사용할 입력변수들: 장기예보 가이던스 III(기상청, 2018) 활용
  - 각 기후감시요소에 대하여 시계열 자료 산정

### 자료 수집

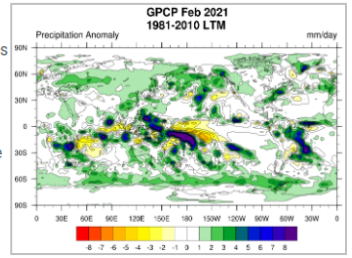
Home » PSL Home » GPCP V2 Precipitation

On this page: Temporal Coverage | Spatial Coverage | Levels | Update Schedule | Download/Plot Data | Analysis Tools
Restrictions | Details | Caveats | File Naming | Citation | References | Original Source | Contact

GPCP Version 2.3 Combined Precipitation Data Set

**Note:** This dataset has been updated to version 2.3 and will be updated regularly. 10/09/2020: Grids from years since 2015 have been replaced. See the NCEI webpage for more information

*Latest available data: Click to Enlarge*




Brief Description:
- Global Precipitation Climatology Project monthly precipitation dataset from 1979-present combines observations and satellite precipitation data into 2.5°x2.5° global grids.

Temporal Coverage:
- Monthly values 1979/01 through Feb 2021 (some months are interim).
- Long term monthly means, derived from years 1981 - 2010.

Spatial Coverage:
- 2.5 degree latitude x 2.5 degree longitude global grid (144x72)
- 88.75N - 88.75S, 1.25E - 358.75E

Levels:
- N/A

Update Schedule:
- Monthly

NOAA NATIONAL CENTERS FOR ENVIRONMENTAL INFORMATION
NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION

Formerly the National Climatic Data Center (NCDC)... more about NCEI »

Home | Climate Information | Data Access | Customer Support | Contact | About | Search

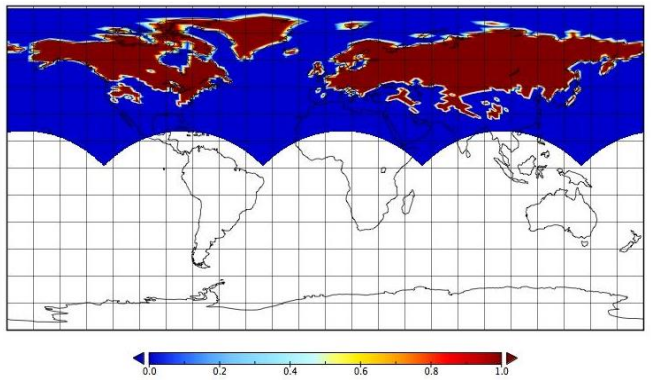
Home > Climate Data Record Program > Terrestrial > Snow Cover Extent (Northern Hemisphere)

Data Access
Download
THREDDS

Documentation
CDR Flyer
Use Agreement
Algorithm Description
Data Flow Diagram
Maturity Matrix
Source Code

Information
Contact
Registration (optional)

Snow Cover Extent (Northern Hemisphere)



# II. 기계학습을 이용한 장기예측

## 자료를 수집하자

- 모델에서 사용할 입력변수들: 장기예보 가이던스 III(기상청, 2018) 활용
  - 각 기후감시요소에 대하여 시계열 자료 산정

### 자료 수집

Home » Gridded Climate Data » NCEP Reanalysis Products: Pressure level variables

On this page: Temporal Coverage | Spatial Coverage | Levels | Update Schedule | Download/Plot Data | Analysis Tools Restrictions | Details | Caveats | File Naming | Citation | References | Original Source | Contact

NCEP/NCAR Reanalysis 1: Pressure

We have transitioned the data files from netCDF3 to netCDF4-classic format on Monday Oct 20th, 2014.

Other Grid Types: **NCEP Reanalysis Main Page** | **Pressure Level Data** | **Surface Data** | **Surface Flux Data** | **Other Flux Data** |**Spectral Coefficents Data** | **Tropopause Data**

Brief Description:
- NCEP/NCAR Reanalysis 1

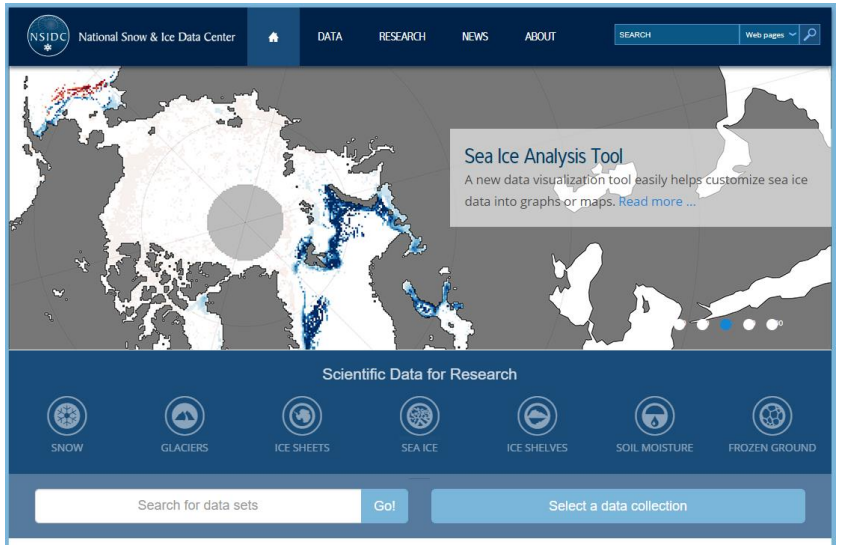
Temporal Coverage:
- 4-times daily, daily and monthly values for 1948/01/01 to present
- Long term monthly means, derived from data for years 1981 - 2010
- Values are instantaneous at the time indicated in the files

Spatial Coverage:
- 2.5 degree x 2.5 degree global grids (144x73)
- 0.0E to 357.5E, 90.0N to 90.0S

Levels:
- 17 Pressure levels (mb): 1000,925,850,700,600,500,400,300,250,200,150,100,70,50,30,20,10
- Some variables have less: omega (to 100mb) and Humidities (to 300mb).

Update Schedule:
- Daily

Download/Plot Data: (Download Issues)

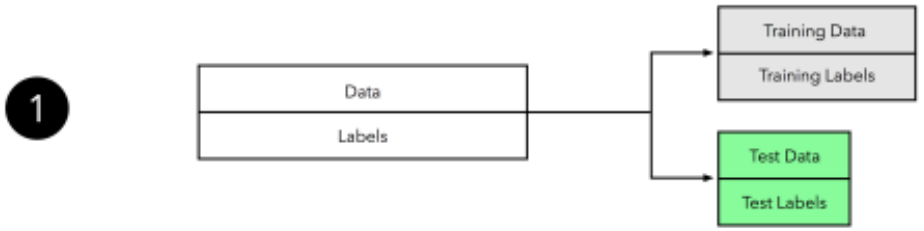


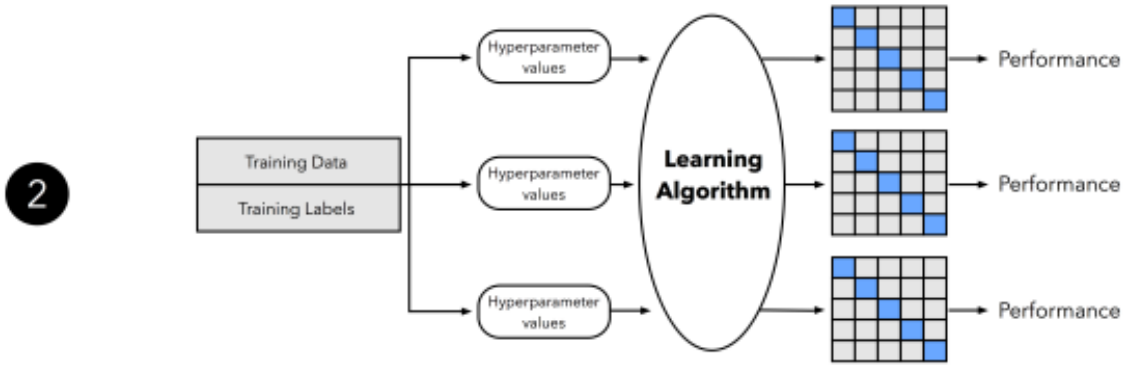
# II. 기계학습을 이용한 장기예측

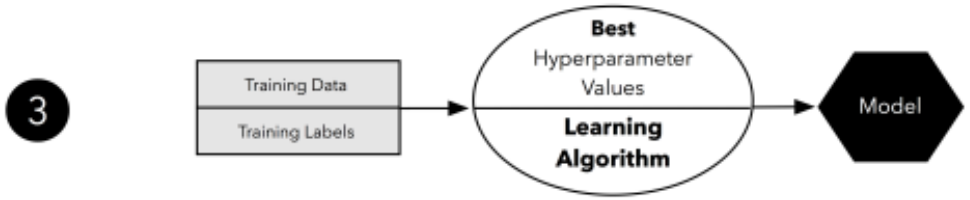
## 자료를 수집하자

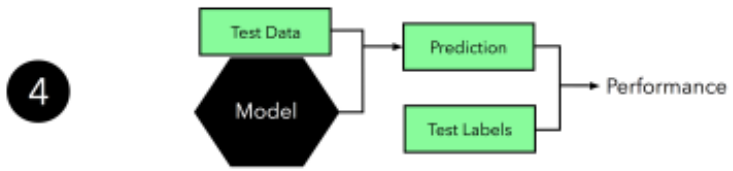
### 모델의 훈련/검증 및 테스트를 위한 자료 분리

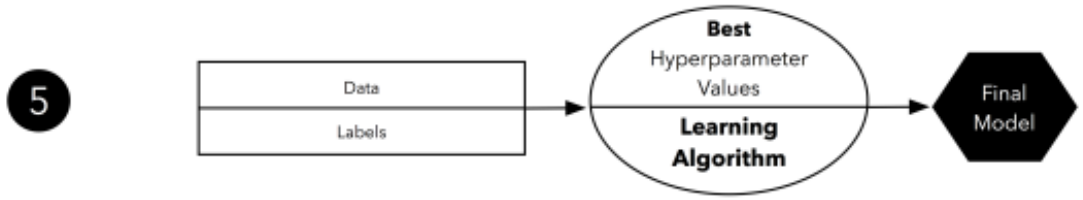
- 본 예시에서는 테스트 수행하지 않고 Leave One-year Out Cross Validation




출처: Tensorflow 블로그

# II. 기계학습을 이용한 장기예측

## 자료를 수집하자

- 모델에서 사용할 입력변수들: 장기예보 가이던스 III(기상청, 2018) 활용
  - 각 기후감시요소에 대하여 시계열 자료 산정

### 자료 전처리 및 입력변수(Feature) 선정

- NetCDF 등의 자료 읽기
- 기후감시요소 지역에 대해 영역평균

- 각 지역에 대해 전년도 12월 ~ 당해 년도 4월 중 실제값과 상관관계 높은 월 사용(상관관계 산정 시 검증 년도 미포함)
- 본 예시에서는 온난화 트렌드 고려하지 않음

| 변수 | 지역번호 | 기후감시시점(월) | r | p | n |
|---|---|---|---|---|---|
| sst | rg1_sub_2 | 1 | -0.29 | 0.070 | 40 |
| sst | rg2 | 3 | 0.29 | 0.068 | 40 |
| sst | rg3 | 4 | 0.27 | 0.089 | 40 |
| sst | rg5 | 4 | 0.48 | 0.002 | 40 |
| sst | rg27 | 1 | 0.35 | 0.028 | 40 |
| sst | rg26 | 4 | 0.31 | 0.054 | 40 |
| sst | rg30 | 4 | 0.30 | 0.060 | 40 |
| sst | rg14 | 4 | 0.53 | 0.000 | 40 |
| sst | rg20 | 1 | 0.33 | 0.038 | 40 |
| sst | rg19 | 4 | 0.46 | 0.003 | 40 |
| prcp | rg12 | 1 | 0.31 | 0.054 | 40 |
| prcp | rg12_sub_13 | 1 | 0.33 | 0.040 | 40 |
| snow | rg8 | 12 | -0.31 | 0.052 | 40 |
| snow | rg8_sub_9 | 12 | -0.34 | 0.034 | 39 |
| snow | rg10_add_11 | 3 | 0.51 | 0.001 | 40 |
| snow | rg9 | 3 | 0.32 | 0.047 | 40 |
| snow | rg10 | 3 | 0.40 | 0.010 | 40 |
| snow | rg24 | 3 | 0.39 | 0.012 | 40 |
| snow | rg28 | 3 | 0.55 | 0.000 | 40 |
| snow | rg11 | 3 | 0.54 | 0.000 | 40 |
| snow | rg24_sub_25 | 3 | 0.36 | 0.022 | 40 |
| 500hpa | rg16 | 4 | 0.45 | 0.004 | 40 |
| 500hpa | rg15_add_16_sub_17 | 12 | 0.33 | 0.036 | 40 |
| 500hpa | rg15 | 3 | 0.29 | 0.067 | 40 |
| seaice | laptev | 4 | -0.43 | 0.006 | 40 |
| seaice | bering | 1 | -0.32 | 0.046 | 39 |
| seaice | kara | 3 | -0.31 | 0.052 | 40 |
| seaice | barents | 3 | -0.32 | 0.047 | 40 |

1980년 6월 평균온도 교차검증 예측을 위한 (1981-2020년 자료 이용) 입력변수 목록

# II. 기계학습을 이용한 장기예측

## 자료를 수집하자

- 모델에서 사용할 입력변수들: 장기예보 가이던스 III(기상청, 2018) 활용
  - 각 기후감시요소에 대하여 시계열 자료 산정

**입력파일 완성!**

기후감시요소 입력변수들

연도

온도 관측값

| year | sst_rg1_sul | sst_rg2_3 | sst_rg3_4 | sst_rg5_4 | sst_rg27_1 | sst_rg26_4 | sst_rg30_4 | sst_rg14_4 | sst_rg20_1 | sst_rg19_4 | prcp_rg12 | prcp_rg12 | snow_rg8 | snow_rg8_s | snow_rg10 | snow_rg9 | snow_rg10 | snow_rg24 | snow_rg28 | snow_rg11 | snow_rg24 | 500hpa_rg | 500hpa_rg1 | 500hpa_rg | seaice_laptev_ | seaice_bering | seaice_kara_3 | seaice_barents_ | obs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1980 | 0.33 | -0.62 | -0.76 | 0.24 | -0.37 | -0.27 | 0.01 | -0.23 | -0.37 | -0.51 | -1.06 | -0.93 | 0.02 | 0.05 | 0.34 | 0.04 | 0.23 | 0.08 | 0.08 | 0.11 | -0.05 | 10.62 | 9.88 | -34.19 | 13156.26 | 90767.88 | 30992.29 | 249572.84 | 21.52 |
| 1981 | 0.53 | -0.79 | -0.23 | -0.11 | -0.43 | -0.19 | -0.12 | -0.43 | -0.43 | -0.27 | -1.40 | -1.49 | -0.14 | | 0.30 | 0.04 | 0.12 | 0.05 | 0.14 | 0.17 | -0.08 | 4.12 | 87.06 | 19.89 | 12450.59 | 87877.26 | 31606.11 | 362016.19 | 20.87 |
| 1982 | 1.51 | -0.73 | -0.70 | -0.51 | -0.31 | -0.47 | -0.36 | -0.07 | -0.38 | -0.24 | 1.48 | 1.60 | 0.17 | 0.18 | -0.29 | -0.03 | -0.04 | -0.05 | -0.21 | -0.25 | -0.18 | 4.77 | 84.42 | -32.24 | 17574.47 | -24101.28 | 38067.11 | 376357.39 | 21.10 |
| 1983 | -0.90 | -0.18 | -1.13 | -1.03 | 0.04 | 0.06 | -0.97 | -0.52 | 0.06 | -0.54 | -3.34 | -3.15 | -0.02 | 0.09 | 0.15 | -0.04 | 0.12 | -0.01 | 0.04 | 0.03 | 0.01 | -19.84 | 38.45 | -7.43 | 5170.61 | 44835.07 | 23456.55 | 48986.60 | 21.00 |
| 1984 | -0.40 | 0.00 | -0.02 | 0.18 | -0.48 | -0.45 | -0.68 | -0.42 | -0.49 | -0.06 | -0.19 | -1.51 | 0.12 | 0.18 | 0.31 | 0.04 | 0.11 | 0.06 | 0.21 | 0.21 | -0.02 | -20.43 | 25.60 | -41.99 | 12220.50 | -116102.43 | 2548.83 | 64610.29 | 21.67 |
| 1985 | -0.29 | -0.58 | -0.29 | -0.22 | -0.50 | -0.11 | -0.36 | -0.55 | -0.41 | -0.16 | -2.66 | -1.57 | -0.14 | -0.14 | 0.41 | 0.04 | 0.22 | 0.07 | 0.17 | 0.19 | -0.07 | -3.19 | -190.92 | -6.63 | 9737.06 | -126793.47 | 12856.76 | 67284.06 | 20.69 |
| 1986 | 0.11 | -0.22 | -0.49 | -0.23 | -0.71 | -0.26 | -0.41 | -0.19 | -0.70 | -0.34 | -2.05 | -2.24 | 0.05 | 0.05 | -0.02 | -0.03 | -0.06 | -0.04 | 0.03 | 0.04 | -0.19 | -2.32 | 96.51 | 13.98 | 5679.50 | -50769.63 | 6043.03 | 158144.73 | 20.92 |
| 1987 | -0.04 | -0.73 | -0.74 | -0.41 | -0.21 | -0.47 | -0.12 | -0.48 | -0.17 | -0.14 | -2.60 | -2.69 | -0.20 | -0.18 | 0.48 | 0.04 | 0.29 | 0.07 | 0.15 | 0.19 | -0.10 | -8.59 | 5.08 | -74.67 | 3337.38 | -8217.90 | 25101.42 | 303503.64 | 21.18 |
| 1988 | 0.59 | -0.57 | 0.24 | 0.38 | -0.08 | 0.02 | -0.35 | -0.09 | -0.04 | 0.18 | -0.61 | 0.04 | 0.13 | 0.13 | 0.39 | 0.04 | 0.30 | 0.07 | 0.09 | 0.09 | 0.04 | 21.40 | 91.38 | -32.36 | 12142.09 | | 27805.66 | 218486.98 | 21.35 |
| 1989 | 0.94 | 0.18 | 0.14 | -0.68 | -0.21 | -0.26 | -0.44 | -0.31 | -0.19 | 0.15 | 2.53 | 0.95 | 0.11 | 0.11 | -0.19 | -0.01 | -0.08 | -0.01 | -0.08 | -0.11 | -0.11 | -36.05 | 41.90 | -2.76 | 7308.94 | 139426.33 | 20607.76 | 121773.41 | 20.02 |
| 1990 | 0.12 | -0.12 | -0.51 | -0.49 | -0.51 | -0.04 | -0.62 | -0.12 | -0.43 | -0.26 | -1.44 | -2.11 | 0.07 | 0.08 | -0.30 | -0.01 | -0.17 | -0.05 | -0.09 | -0.12 | -0.01 | -23.30 | 71.43 | 50.54 | -12443.06 | 58443.83 | 11730.81 | 91690.66 | 21.03 |
| 1991 | -0.66 | -0.04 | 0.17 | -0.04 | 0.19 | -0.54 | -0.33 | 0.09 | 0.14 | -0.39 | -2.08 | -2.03 | 0.03 | 0.03 | 0.05 | 0.01 | 0.09 | 0.01 | 0.02 | -0.05 | -0.06 | -7.57 | -52.81 | 20.39 | -8286.90 | 17281.12 | 32455.76 | 92329.75 | 21.96 |
| 1992 | 0.36 | -0.51 | -0.81 | -1.03 | -0.05 | 0.07 | -0.70 | -0.57 | -0.03 | -0.20 | -2.86 | -2.61 | 0.12 | 0.12 | -0.39 | 0.00 | -0.08 | -0.03 | -0.23 | -0.31 | -0.11 | -12.17 | -81.28 | -17.32 | 9695.72 | 74553.36 | 11150.21 | -68341.76 | 19.96 |
| 1993 | -0.69 | -0.28 | -0.57 | -0.40 | -0.50 | -0.55 | -0.76 | -0.53 | -0.40 | -0.43 | -3.58 | -2.74 | 0.06 | 0.06 | -0.39 | -0.08 | -0.16 | -0.09 | -0.22 | -0.24 | -0.15 | -7.53 | -60.23 | -27.29 | -2744.82 | 16664.91 | 13925.51 | 76145.06 | 20.65 |
| 1994 | -0.50 | -0.31 | -0.51 | 0.04 | -0.27 | -0.44 | -0.84 | -0.09 | -0.28 | 0.05 | -0.59 | 0.36 | -0.03 | -0.03 | 0.19 | 0.01 | 0.13 | 0.03 | 0.06 | 0.06 | -0.09 | -34.19 | 129.66 | 25.39 | 11816.15 | 17426.96 | 29631.89 | 171369.68 | 21.00 |
| 1995 | -0.65 | 0.12 | -0.54 | -0.41 | -0.18 | -0.27 | -0.88 | 0.07 | -0.13 | -0.39 | -2.32 | -3.16 | 0.01 | 0.01 | -0.02 | 0.03 | -0.03 | 0.03 | -0.01 | 0.01 | 0.07 | -11.05 | -23.06 | -20.21 | 17437.51 | 127980.64 | 801.94 | -56449.04 | 20.10 |
| 1996 | 0.70 | -0.46 | -0.22 | -0.17 | -0.31 | -0.12 | -0.17 | -0.20 | -0.38 | -0.14 | 1.91 | 1.59 | 0.06 | 0.06 | -0.30 | 0.02 | -0.15 | -0.01 | -0.14 | -0.15 | -0.09 | -26.49 | -30.29 | -49.18 | 8193.00 | -68125.98 | -5719.89 | 58526.58 | 20.87 |
| 1997 | 0.72 | -0.53 | -0.67 | -0.46 | -0.29 | -0.66 | 0.13 | 0.03 | -0.35 | -0.38 | 0.57 | 0.64 | -0.10 | -0.11 | 0.07 | 0.02 | 0.01 | 0.03 | 0.08 | 0.06 | -0.03 | 1.94 | 24.10 | -50.92 | -17150.81 | -70336.21 | 2614.78 | 119335.06 | 21.73 |
| 1998 | 1.03 | -0.50 | -0.24 | 0.40 | -0.10 | 0.37 | 0.33 | -0.36 | 0.03 | -0.46 | -3.93 | -3.50 | 0.11 | 0.12 | -0.61 | -0.07 | -0.36 | -0.13 | -0.22 | -0.25 | -0.26 | 8.49 | -93.52 | -42.82 | 21580.97 | 117409.94 | 27753.42 | 353139.02 | 20.03 |
| 1999 | 0.00 | 0.36 | 0.79 | -0.06 | -0.25 | 0.28 | 0.10 | -0.50 | -0.20 | 0.16 | -0.91 | -1.56 | 0.12 | 0.13 | 0.29 | 0.03 | 0.11 | 0.05 | 0.15 | 0.18 | 0.18 | -6.95 | 56.91 | -19.14 | 13483.95 | 40229.98 | 33767.72 | 226541.42 | 21.45 |
| 2000 | -0.23 | 0.23 | 0.33 | -0.16 | -0.03 | 0.01 | -0.24 | -0.60 | -0.10 | 0.42 | -0.39 | -0.70 | 0.06 | 0.07 | 0.21 | 0.03 | 0.06 | 0.04 | 0.16 | 0.15 | 0.05 | 7.44 | 185.68 | -0.97 | -1945.48 | 254527.28 | 6544.31 | -37566.61 | 21.43 |
| 2001 | 0.54 | -0.39 | 0.48 | -0.14 | -0.56 | -0.15 | 0.10 | -0.06 | -0.44 | 0.55 | 1.40 | 1.51 | -0.07 | -0.07 | 0.34 | 0.04 | 0.16 | 0.06 | 0.14 | 0.18 | 0.15 | -1.06 | 15.41 | 48.42 | 12145.20 | -74235.78 | 20042.06 | 161490.74 | 21.56 |
| 2002 | 0.62 | 0.24 | -0.29 | 0.14 | 0.28 | 0.23 | 0.35 | 0.28 | 0.16 | -0.01 | 0.30 | 0.04 | 0.17 | 0.22 | -0.56 | -0.05 | -0.28 | -0.09 | -0.25 | -0.29 | 0.11 | -15.09 | -58.55 | 15.75 | 13378.99 | 137515.25 | 27840.63 | 130491.66 | 20.94 |
| 2003 | 1.10 | -0.25 | -0.14 | 0.19 | 0.53 | 0.06 | 0.68 | 0.24 | 0.46 | -0.03 | -0.01 | 0.99 | 0.09 | 0.09 | -0.06 | -0.03 | 0.16 | -0.04 | -0.17 | -0.23 | -0.19 | -6.46 | -106.46 | -28.39 | 5904.55 | -102683.95 | 26373.79 | 250264.69 | 20.72 |
| 2004 | 0.65 | -0.55 | 0.10 | 0.10 | 0.25 | -0.09 | 0.59 | 0.18 | 0.19 | 0.23 | -2.47 | -2.56 | -0.03 | -0.04 | 0.18 | 0.04 | 0.20 | 0.06 | -0.02 | -0.02 | 0.15 | 3.47 | 26.84 | 45.04 | 24455.15 | -120947.84 | 28384.77 | 109684.72 | 21.60 |
| 2005 | -0.07 | -0.28 | -0.28 | -0.16 | 0.17 | 0.09 | 0.46 | 0.13 | 0.18 | -0.30 | 2.11 | 2.00 | 0.06 | 0.06 | 0.13 | 0.00 | 0.05 | 0.00 | 0.07 | 0.08 | -0.10 | 5.12 | 42.15 | 4.26 | -1282.67 | -134542.88 | 28147.27 | 38469.88 | 22.19 |
| 2006 | 0.49 | -0.33 | 0.39 | -0.08 | -0.31 | -0.30 | 0.38 | 0.00 | -0.27 | 0.00 | 2.26 | 0.91 | 0.15 | 0.15 | 0.06 | 0.03 | -0.03 | 0.04 | 0.07 | 0.09 | 0.24 | -1.62 | -17.75 | -9.13 | 27155.52 | 184211.68 | 539.03 | -49603.02 | 21.14 |
| 2007 | 0.66 | -0.39 | 0.06 | 0.17 | 0.02 | -0.33 | 0.38 | 0.23 | 0.06 | -0.08 | -3.48 | -3.38 | -0.45 | -0.46 | 0.31 | 0.04 | 0.11 | 0.06 | 0.15 | 0.20 | -0.01 | 33.65 | 36.43 | -2.16 | -58048.93 | 62775.40 | -17239.01 | -182295.64 | 21.51 |
| 2008 | 0.51 | 0.29 | 0.86 | -0.21 | 0.03 | -0.22 | 0.32 | -0.41 | 0.02 | -0.06 | 3.43 | 2.90 | 0.00 | 0.01 | -0.54 | -0.07 | -0.29 | -0.12 | -0.23 | -0.25 | 0.05 | -8.59 | -125.42 | 3.26 | 15051.47 | 25480.15 | 14253.06 | -131101.67 | 20.29 |
| 2009 | 0.41 | -0.02 | 0.32 | 0.04 | 0.04 | 0.21 | 0.41 | 0.00 | -0.05 | 0.14 | 1.40 | 1.94 | -0.09 | -0.09 | 0.26 | 0.02 | 0.14 | 0.05 | 0.12 | 0.12 | 0.10 | -14.27 | -80.59 | -27.55 | -11064.24 | 140131.41 | 11308.57 | -47090.96 | 21.65 |
| 2010 | 0.83 | -0.60 | -0.27 | 0.78 | -0.01 | 0.47 | 0.42 | 0.20 | 0.05 | 0.02 | -0.61 | -0.09 | -0.02 | -0.02 | 0.57 | 0.04 | 0.26 | 0.08 | 0.27 | 0.31 | 0.07 | 20.75 | -14.84 | 48.49 | 10296.96 | 40599.23 | 9059.35 | 93943.96 | 22.16 |
| 2011 | 1.49 | -0.44 | 0.46 | -0.32 | -0.36 | -0.47 | 0.31 | 0.12 | -0.42 | -0.25 | 1.02 | -0.13 | 0.15 | 0.15 | 0.31 | 0.03 | 0.16 | 0.05 | 0.13 | 0.14 | -0.01 | 7.18 | 18.16 | 17.28 | -20602.20 | -12749.90 | -45224.64 | -68234.32 | 21.75 |
| 2012 | -0.48 | 0.50 | 0.34 | -0.41 | -0.07 | -0.07 | -0.06 | -0.15 | -0.07 | 0.31 | 1.01 | 0.36 | -0.14 | -0.15 | 0.21 | 0.01 | 0.08 | 0.02 | 0.12 | 0.13 | -0.05 | 4.79 | 84.82 | 5.31 | 705.69 | 200162.07 | -95340.79 | -221954.65 | 21.94 |
| 2013 | 0.03 | -0.05 | -0.04 | 0.06 | 0.17 | 0.24 | 0.40 | -0.06 | 0.16 | 0.66 | 1.11 | 1.18 | 0.14 | 0.14 | 0.59 | 0.04 | 0.26 | 0.08 | 0.28 | 0.33 | 0.15 | -2.34 | -39.77 | 7.11 | -13673.05 | 73535.21 | 31999.84 | 31710.27 | 22.45 |
| 2014 | -0.51 | 0.43 | -0.20 | -0.03 | 0.17 | 0.23 | 0.28 | 0.21 | 0.09 | 0.13 | 2.84 | 3.20 | 0.05 | 0.05 | -0.19 | -0.07 | -0.15 | -0.08 | -0.06 | -0.04 | -0.11 | -8.24 | 16.99 | -2.33 | -32702.16 | -77948.29 | -38759.28 | -215194.05 | 21.64 |
| 2015 | -1.70 | 0.53 | -0.34 | 0.36 | 0.02 | 0.32 | -0.66 | 0.38 | 0.00 | -0.22 | 0.73 | 0.55 | 0.06 | 0.05 | -0.01 | 0.00 | -0.08 | 0.00 | 0.03 | 0.07 | -0.09 | 9.44 | 79.18 | 16.00 | 6497.71 | -193875.34 | -18219.05 | -178333.10 | 21.50 |
| 2016 | -1.54 | 0.91 | -0.35 | 0.74 | 1.04 | 0.79 | -0.33 | 0.16 | 1.02 | 0.27 | -1.38 | -1.86 | -0.28 | -0.28 | -0.15 | 0.01 | -0.08 | -0.01 | -0.08 | -0.07 | 0.05 | 7.12 | 52.18 | -0.60 | 9682.55 | -85664.05 | -53245.50 | -318016.58 | 22.16 |
| 2017 | -1.26 | 0.54 | 0.49 | 0.09 | -0.11 | -0.06 | -0.12 | 0.37 | -0.06 | -0.09 | 0.48 | 0.16 | 0.11 | 0.11 | 0.11 | 0.00 | 0.09 | 0.01 | 0.02 | 0.02 | -0.07 | 14.72 | 4.52 | 3.35 | -6333.25 | -199262.00 | -7916.68 | -183763.36 | 21.60 |
| 2018 | -1.08 | 0.65 | 0.00 | 0.25 | -0.33 | 0.61 | -0.24 | 0.25 | -0.23 | 0.16 | 1.69 | 1.39 | -0.01 | -0.02 | -0.09 | -0.05 | -0.10 | -0.06 | 0.00 | 0.01 | -0.09 | 11.84 | 95.44 | 18.04 | -464.35 | -247555.38 | 13871.31 | -62679.55 | 22.05 |
| 2019 | -0.14 | 0.56 | 0.27 | 0.24 | 0.19 | -0.16 | -0.12 | 0.11 | 0.24 | -0.04 | 0.43 | 2.02 | 0.13 | 0.12 | -0.49 | -0.01 | -0.24 | -0.03 | -0.26 | -0.25 | 0.05 | 14.47 | -59.47 | 10.82 | -5701.21 | -55179.64 | 3685.39 | -41747.17 | 21.18 |
| 2020 | -0.59 | 0.58 | 0.40 | 0.47 | 0.63 | 0.46 | -0.19 | 0.47 | 0.60 | 0.35 | 1.94 | 2.58 | -0.36 | -0.36 | -0.05 | -0.01 | -0.04 | -0.02 | 0.02 | -0.01 | 0.09 | 13.22 | 25.17 | 9.09 | -27481.00 | -87377.40 | -21239.77 | -51071.01 | 22.68 |

1980년 6월 평균온도 예측을 위한 입력파일

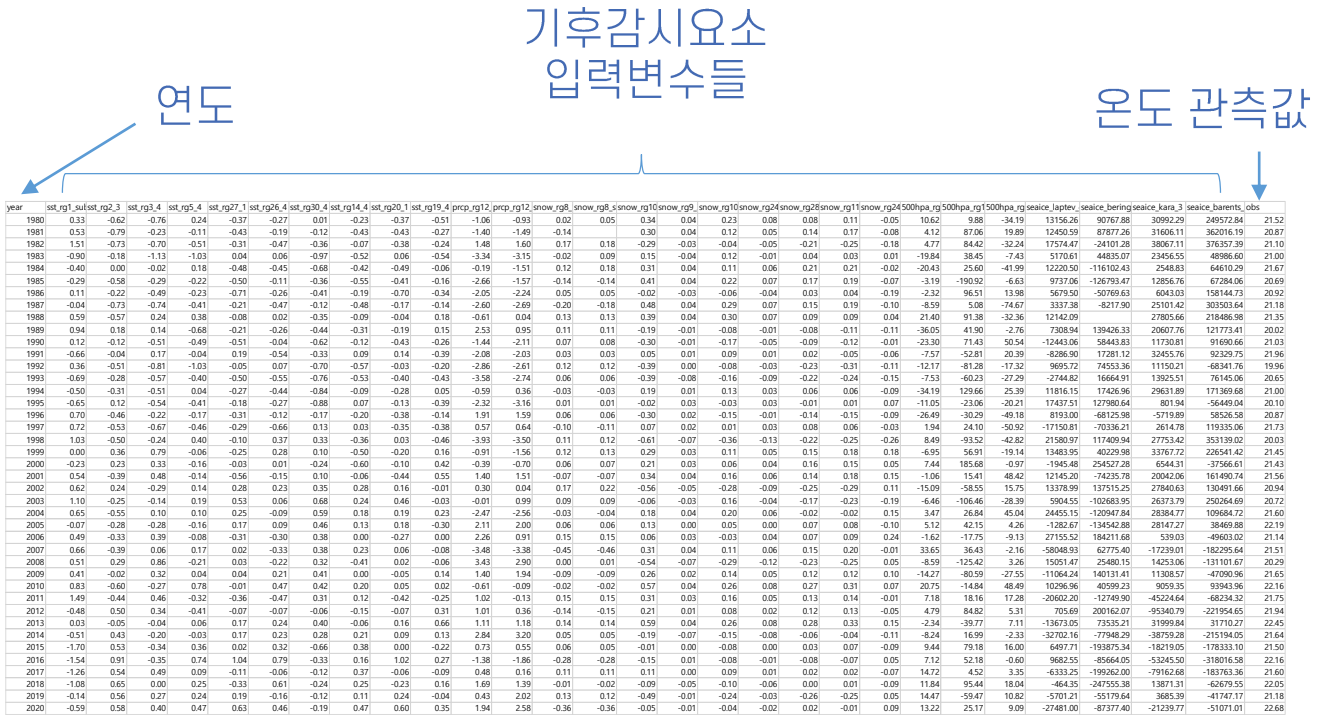
# II. 기계학습을 이용한 장기예측

## 자료를 수집하자

- 모델에서 사용할 입력변수들: 장기예보 가이던스 III(기상청, 2018) 활용
  - 각 기후감시요소에 대하여 시계열 자료 산정

### 훈련 및 테스트

- 1년 씩 제외하여 교차검증
- 테스트 자료 따로 정해야 하나 **본 예시에서는 교차검증 평균 결과만 살펴봄**
- 장기예보 가이던스에서 기후감시요소 지역 도출한 기간 제외해야 하나 본 예시에는 포함




| year | sst_rg1_sul | sst_rg2_3 | sst_rg3_4 | sst_rg5_4 | sst_rg27_1 | sst_rg26_4 | sst_rg30_4 | sst_rg14_4 | sst_rg20_1 | sst_rg19_4 | prcp_rg12 | prcp_rg12 | snow_rg8 | snow_rg8_s | snow_rg10 | snow_rg9 | snow_rg10 | snow_rg24 | snow_rg28 | snow_rg11 | snow_rg24 | 500hpa_rg | 500hpa_rg1 | 500hpa_rg | seaice_laptev | seaice_bering | seaice_kara_3 | seaice_barents | obs |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1980 | 0.33 | -0.62 | -0.76 | 0.24 | -0.37 | -0.27 | 0.01 | -0.23 | -0.37 | -0.51 | -1.06 | -0.93 | 0.02 | 0.05 | 0.34 | 0.04 | 0.23 | 0.08 | 0.08 | 0.11 | -0.05 | 10.62 | 9.88 | -34.19 | 13156.26 | 90767.88 | 30992.29 | 249572.84 | 21.52 |
| 1981 | 0.53 | -0.79 | -0.23 | -0.11 | -0.43 | -0.19 | -0.12 | -0.43 | -0.43 | -0.27 | -1.40 | -1.49 | -0.14 | | 0.30 | 0.04 | 0.12 | 0.05 | 0.14 | 0.17 | -0.08 | 4.12 | 87.06 | 19.89 | 12450.59 | 87877.26 | 31606.11 | 362016.19 | 20.87 |
| 1982 | 1.51 | -0.73 | -0.70 | -0.51 | -0.31 | -0.47 | -0.36 | -0.07 | -0.38 | -0.24 | 1.48 | 1.60 | 0.17 | 0.18 | -0.29 | -0.03 | -0.04 | -0.05 | -0.21 | -0.25 | -0.18 | 4.77 | 84.42 | -32.24 | 17574.47 | -24101.28 | 38067.11 | 376357.39 | 21.10 |
| 1983 | -0.90 | -0.18 | -1.13 | -1.03 | 0.04 | 0.06 | -0.97 | -0.52 | 0.06 | -0.54 | -3.34 | -3.15 | -0.02 | 0.09 | 0.15 | -0.04 | 0.12 | -0.01 | 0.04 | 0.03 | 0.01 | -19.84 | 38.45 | -7.43 | 5170.61 | 44835.07 | 23456.55 | 48986.60 | 21.00 |
| 1984 | -0.40 | 0.00 | -0.02 | 0.18 | -0.48 | -0.45 | -0.68 | -0.42 | -0.49 | -0.06 | -0.19 | -1.51 | 0.12 | 0.18 | 0.31 | 0.04 | 0.11 | 0.06 | 0.21 | 0.21 | -0.02 | -20.43 | 25.60 | -41.99 | 12220.50 | -116102.43 | 2548.83 | 64610.29 | 21.67 |
| 1985 | -0.29 | -0.58 | -0.29 | -0.22 | -0.50 | -0.11 | -0.36 | -0.55 | -0.41 | -0.16 | -2.66 | -1.57 | -0.14 | -0.14 | 0.41 | 0.04 | 0.22 | 0.07 | 0.17 | 0.19 | -0.07 | -3.19 | -190.92 | -6.63 | 9737.06 | -126793.47 | 12856.76 | 67284.06 | 20.69 |
| 1986 | 0.11 | -0.22 | -0.49 | -0.23 | -0.71 | -0.26 | -0.41 | -0.19 | -0.70 | -0.34 | -2.05 | -2.24 | 0.05 | 0.05 | -0.02 | -0.03 | -0.06 | -0.04 | 0.03 | 0.04 | -0.19 | -2.32 | 96.51 | 13.98 | 5679.50 | -50769.63 | 6043.03 | 158144.73 | 20.92 |
| 1987 | -0.04 | -0.73 | -0.74 | -0.41 | -0.21 | -0.47 | -0.12 | -0.48 | -0.17 | -0.14 | -2.60 | -2.69 | -0.20 | -0.18 | 0.48 | 0.04 | 0.29 | 0.07 | 0.15 | 0.19 | -0.10 | -8.59 | 5.08 | -74.67 | 3337.38 | -8217.90 | 25101.42 | 303503.64 | 21.18 |
| 1988 | 0.59 | -0.57 | 0.24 | 0.38 | -0.08 | 0.02 | -0.35 | -0.09 | -0.04 | 0.18 | -0.61 | 0.04 | 0.13 | 0.13 | 0.39 | 0.04 | 0.30 | 0.07 | 0.09 | 0.09 | 0.04 | 21.40 | 91.38 | -32.36 | 12142.09 | | 27805.66 | 218486.98 | 21.35 |
| 1989 | 0.94 | 0.18 | 0.14 | -0.68 | -0.21 | -0.26 | -0.44 | -0.31 | -0.19 | 0.15 | 2.53 | 0.95 | 0.11 | 0.11 | -0.19 | -0.01 | -0.08 | -0.01 | -0.08 | -0.11 | -0.11 | -36.05 | 41.90 | -2.76 | 7308.94 | 139426.33 | 20607.76 | 121773.41 | 20.02 |
| 1990 | 0.12 | -0.12 | -0.51 | -0.49 | -0.51 | -0.04 | -0.62 | -0.12 | -0.43 | -0.26 | -1.44 | -2.11 | 0.07 | 0.08 | -0.30 | -0.01 | -0.17 | -0.05 | -0.09 | -0.12 | -0.01 | -23.30 | 71.43 | 50.54 | -12443.06 | 58443.83 | 11730.81 | 91690.66 | 21.03 |
| 1991 | -0.66 | -0.04 | 0.17 | -0.04 | 0.19 | -0.54 | -0.33 | 0.09 | 0.14 | -0.39 | -2.08 | -2.03 | 0.03 | 0.03 | 0.05 | 0.01 | 0.09 | 0.01 | 0.02 | -0.05 | -0.06 | -7.57 | -52.81 | 20.39 | -8286.90 | 17281.12 | 32455.76 | 92329.75 | 21.96 |
| 1992 | 0.36 | -0.51 | -0.81 | -1.03 | -0.05 | 0.07 | -0.70 | -0.57 | -0.03 | -0.20 | -2.86 | -2.61 | 0.12 | 0.12 | -0.39 | 0.00 | -0.08 | -0.03 | -0.23 | -0.31 | -0.11 | -12.17 | -81.28 | -17.32 | 9695.72 | 74553.36 | 11150.21 | -68341.76 | 19.96 |
| 1993 | -0.69 | -0.28 | -0.57 | -0.40 | -0.50 | -0.55 | -0.76 | -0.53 | -0.40 | -0.43 | -3.58 | -2.74 | 0.06 | 0.06 | -0.39 | -0.08 | -0.16 | -0.09 | -0.22 | -0.24 | -0.15 | -7.53 | -60.23 | -27.29 | -2744.82 | 16664.91 | 13925.51 | 76145.06 | 20.65 |
| 1994 | -0.50 | -0.31 | -0.51 | 0.04 | -0.27 | -0.44 | -0.84 | -0.09 | -0.28 | 0.05 | -0.59 | 0.36 | -0.03 | -0.03 | 0.19 | 0.01 | 0.13 | 0.03 | 0.06 | 0.06 | -0.09 | -34.19 | 129.66 | 25.39 | 11816.15 | 17426.96 | 29631.89 | 171369.68 | 21.00 |
| 1995 | -0.65 | 0.12 | -0.54 | -0.41 | -0.18 | -0.27 | -0.88 | 0.07 | -0.13 | -0.39 | -2.32 | -3.16 | 0.01 | 0.01 | -0.02 | 0.03 | -0.03 | 0.03 | -0.01 | 0.01 | 0.07 | -11.05 | -23.06 | -20.21 | 17437.51 | 127980.64 | 801.94 | -56449.04 | 20.10 |
| 1996 | 0.70 | -0.46 | -0.22 | -0.17 | -0.31 | -0.12 | -0.17 | -0.20 | -0.38 | -0.14 | 1.91 | 1.59 | 0.06 | 0.06 | -0.30 | 0.02 | -0.15 | -0.01 | -0.14 | -0.15 | -0.09 | -26.49 | -30.29 | -49.18 | 8193.00 | -68125.98 | -5719.89 | 58526.58 | 20.87 |
| 1997 | 0.72 | -0.53 | -0.67 | -0.46 | -0.29 | -0.66 | 0.13 | 0.03 | -0.35 | -0.38 | 0.57 | 0.64 | -0.10 | -0.11 | 0.07 | 0.02 | 0.01 | 0.03 | 0.08 | 0.06 | -0.03 | 1.94 | 24.10 | -50.92 | -17150.81 | -70336.21 | 2614.78 | 119335.06 | 21.73 |
| 1998 | 1.03 | -0.50 | -0.24 | 0.40 | -0.10 | 0.37 | 0.33 | -0.36 | 0.03 | -0.46 | -3.93 | -3.50 | 0.11 | 0.12 | -0.61 | -0.07 | -0.36 | -0.13 | -0.22 | -0.25 | -0.26 | 8.49 | -93.52 | -42.82 | 21580.97 | 117409.94 | 27753.42 | 353139.02 | 20.03 |
| 1999 | 0.00 | 0.36 | 0.79 | -0.06 | -0.25 | 0.28 | 0.10 | -0.50 | -0.20 | 0.16 | -0.91 | -1.56 | 0.12 | 0.13 | 0.29 | 0.03 | 0.11 | 0.05 | 0.15 | 0.18 | 0.18 | -6.95 | 56.91 | -19.14 | 13483.95 | 40229.98 | 33767.72 | 226541.42 | 21.45 |
| 2000 | -0.23 | 0.23 | 0.33 | -0.16 | -0.03 | 0.01 | -0.24 | -0.60 | -0.10 | 0.42 | -0.39 | -0.70 | 0.06 | 0.07 | 0.21 | 0.03 | 0.06 | 0.04 | 0.16 | 0.15 | 0.05 | 7.44 | 185.68 | -0.97 | -1945.48 | 254527.28 | 6544.31 | -37566.61 | 21.43 |
| 2001 | 0.54 | -0.39 | 0.48 | -0.14 | -0.56 | -0.15 | 0.10 | -0.06 | -0.44 | 0.55 | 1.40 | 1.51 | -0.07 | -0.07 | 0.34 | 0.04 | 0.16 | 0.06 | 0.14 | 0.18 | 0.15 | -1.06 | 15.41 | 48.42 | 12145.20 | -74235.78 | 20042.06 | 161490.74 | 21.56 |
| 2002 | 0.62 | 0.24 | -0.29 | 0.14 | 0.28 | 0.23 | 0.35 | 0.28 | 0.16 | -0.01 | 0.30 | 0.04 | 0.17 | 0.22 | -0.56 | -0.05 | -0.28 | -0.09 | -0.25 | -0.29 | 0.11 | -15.09 | -58.55 | 15.75 | 13378.99 | 137515.25 | 27840.63 | 130491.66 | 20.94 |
| 2003 | 1.10 | -0.25 | -0.14 | 0.19 | 0.53 | 0.06 | 0.68 | 0.24 | 0.46 | -0.03 | -0.01 | 0.99 | 0.09 | 0.09 | -0.06 | -0.03 | 0.16 | -0.04 | -0.17 | -0.23 | -0.19 | -6.46 | -106.46 | -28.39 | 5904.55 | -102683.95 | 26373.79 | 250264.69 | 20.72 |
| 2004 | 0.65 | -0.55 | 0.10 | 0.10 | 0.25 | -0.09 | 0.59 | 0.18 | 0.19 | 0.23 | -2.47 | -2.56 | -0.03 | -0.04 | 0.18 | 0.04 | 0.20 | 0.06 | -0.02 | -0.02 | 0.15 | 3.47 | 26.84 | 45.04 | 24455.15 | -120947.84 | 28384.77 | 109684.72 | 21.60 |
| 2005 | -0.07 | -0.28 | -0.28 | -0.16 | 0.17 | 0.09 | 0.46 | 0.13 | 0.18 | -0.30 | 2.11 | 2.00 | 0.06 | 0.06 | 0.13 | 0.00 | 0.05 | 0.00 | 0.07 | 0.08 | -0.10 | 5.12 | 42.15 | 4.26 | -1282.67 | -134542.88 | 28147.27 | 38469.88 | 22.19 |
| 2006 | 0.49 | -0.33 | 0.39 | -0.08 | -0.31 | -0.30 | 0.38 | 0.00 | -0.27 | 0.00 | 2.26 | 0.91 | 0.15 | 0.15 | 0.06 | 0.03 | -0.03 | 0.04 | 0.07 | 0.09 | 0.24 | -1.62 | -17.75 | -9.13 | 27155.52 | 184211.68 | 539.03 | -49603.02 | 21.14 |
| 2007 | 0.66 | -0.39 | 0.06 | 0.17 | 0.02 | -0.33 | 0.38 | 0.23 | 0.06 | -0.08 | -3.48 | -3.38 | -0.45 | -0.46 | 0.31 | 0.04 | 0.11 | 0.06 | 0.15 | 0.20 | -0.01 | 33.65 | 36.43 | -2.16 | -58048.93 | 62775.40 | -17239.01 | -182295.64 | 21.51 |
| 2008 | 0.51 | 0.29 | 0.86 | -0.21 | 0.03 | -0.22 | 0.32 | -0.41 | 0.02 | -0.06 | 3.43 | 2.90 | 0.00 | 0.01 | -0.54 | -0.07 | -0.29 | -0.12 | -0.23 | -0.25 | 0.05 | -8.59 | -125.42 | 3.26 | 15051.47 | 25480.15 | 14253.06 | -131101.67 | 20.29 |
| 2009 | 0.41 | -0.02 | 0.32 | 0.04 | 0.04 | 0.21 | 0.41 | 0.00 | -0.05 | 0.14 | 1.40 | 1.94 | -0.09 | -0.09 | 0.26 | 0.02 | 0.14 | 0.05 | 0.12 | 0.12 | 0.10 | -14.27 | -80.59 | -27.55 | -11064.24 | 140131.41 | 11308.57 | -47090.96 | 21.65 |
| 2010 | 0.83 | -0.60 | -0.27 | 0.78 | -0.01 | 0.47 | 0.42 | 0.20 | 0.05 | 0.02 | -0.61 | -0.09 | -0.02 | -0.02 | 0.57 | 0.04 | 0.26 | 0.08 | 0.27 | 0.31 | 0.07 | 20.75 | -14.84 | 48.49 | 10296.96 | 40599.23 | 9059.35 | 93943.96 | 22.16 |
| 2011 | 1.49 | -0.44 | 0.46 | -0.32 | -0.36 | -0.47 | 0.31 | 0.12 | -0.42 | -0.25 | 1.02 | -0.13 | 0.15 | 0.15 | 0.31 | 0.03 | 0.16 | 0.05 | 0.13 | 0.14 | -0.01 | 7.18 | 18.16 | 17.28 | -20602.20 | -12749.90 | -45224.64 | -68234.32 | 21.75 |
| 2012 | -0.48 | 0.50 | 0.34 | -0.41 | -0.07 | -0.07 | -0.06 | -0.15 | -0.07 | 0.31 | 1.01 | 0.36 | -0.14 | -0.15 | 0.21 | 0.01 | 0.08 | 0.02 | 0.12 | 0.13 | -0.05 | 4.79 | 84.82 | 5.31 | 705.69 | 200162.07 | -95340.79 | -221954.65 | 21.94 |
| 2013 | 0.03 | -0.05 | -0.04 | 0.06 | 0.17 | 0.24 | 0.40 | -0.06 | 0.16 | 0.66 | 1.11 | 1.18 | 0.14 | 0.14 | 0.59 | 0.04 | 0.26 | 0.08 | 0.28 | 0.33 | 0.15 | -2.34 | -39.77 | 7.11 | -13673.05 | 73535.21 | 31999.84 | 31710.27 | 22.45 |
| 2014 | -0.51 | 0.43 | -0.20 | -0.03 | 0.17 | 0.23 | 0.28 | 0.21 | 0.09 | 0.13 | 2.84 | 3.20 | 0.05 | 0.05 | -0.19 | -0.07 | -0.15 | -0.08 | -0.06 | -0.04 | -0.11 | -8.24 | 16.99 | -2.33 | -32702.16 | -77948.29 | -38759.28 | -215194.05 | 21.64 |
| 2015 | -1.70 | 0.53 | -0.34 | 0.36 | 0.02 | 0.32 | -0.66 | 0.38 | 0.00 | -0.22 | 0.73 | 0.55 | 0.06 | 0.05 | -0.01 | 0.00 | -0.08 | 0.00 | 0.03 | 0.07 | -0.09 | 9.44 | 79.18 | 16.00 | 6497.71 | -193875.34 | -18219.05 | -178333.10 | 21.50 |
| 2016 | -1.54 | 0.91 | -0.35 | 0.74 | 1.04 | 0.79 | -0.33 | 0.16 | 1.02 | 0.27 | -1.38 | -1.86 | -0.28 | -0.28 | -0.15 | 0.01 | -0.08 | -0.01 | -0.08 | -0.07 | 0.05 | 7.12 | 52.18 | -0.60 | 9682.55 | -85664.05 | -53245.50 | -318016.58 | 22.16 |
| 2017 | -1.26 | 0.54 | 0.49 | 0.09 | -0.11 | -0.06 | -0.12 | 0.37 | -0.06 | -0.09 | 0.48 | 0.16 | 0.11 | 0.11 | 0.11 | 0.00 | 0.09 | 0.01 | 0.02 | 0.02 | -0.07 | 14.72 | 4.52 | 3.35 | -6333.25 | -199262.00 | -7916.68 | -183763.36 | 21.60 |
| 2018 | -1.08 | 0.65 | 0.00 | 0.25 | -0.33 | 0.61 | -0.24 | 0.25 | -0.23 | 0.16 | 1.69 | 1.39 | -0.01 | -0.02 | -0.09 | -0.05 | -0.10 | -0.06 | 0.00 | 0.01 | -0.09 | 11.84 | 95.44 | 18.04 | -464.35 | -247555.38 | 13871.31 | -62679.55 | 22.05 |
| 2019 | -0.14 | 0.56 | 0.27 | 0.24 | 0.19 | -0.16 | -0.12 | 0.11 | 0.24 | -0.04 | 0.43 | 2.02 | 0.13 | 0.12 | -0.49 | -0.01 | -0.24 | -0.03 | -0.26 | -0.25 | 0.05 | 14.47 | -59.47 | 10.82 | -5701.21 | -55179.64 | 3685.39 | -41747.17 | 21.18 |
| 2020 | -0.59 | 0.58 | 0.40 | 0.47 | 0.63 | 0.46 | -0.19 | 0.47 | 0.60 | 0.35 | 1.94 | 2.58 | -0.36 | -0.36 | -0.05 | -0.01 | -0.04 | -0.02 | 0.02 | -0.01 | 0.09 | 13.22 | 25.17 | 9.09 | -27481.00 | -87377.40 | -21239.77 | -51071.01 | 22.68 |

1980년 6월 평균온도 예측을 위한 입력파일

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

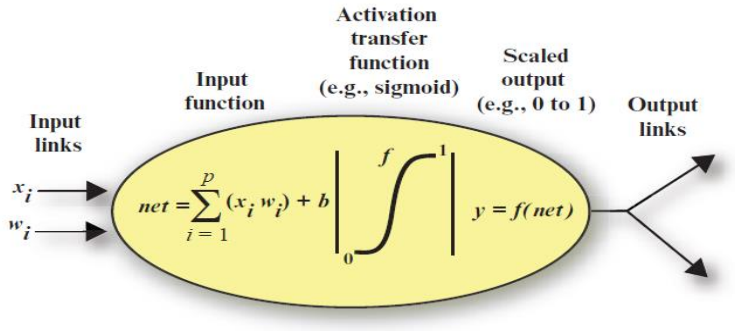
### 지도학습, 회귀모델

- 인공신경망
- 트리를 이용한 기계학습
    - 앙상블 / 부스팅 기법
- 서포트 벡터 머신

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 인공신경망

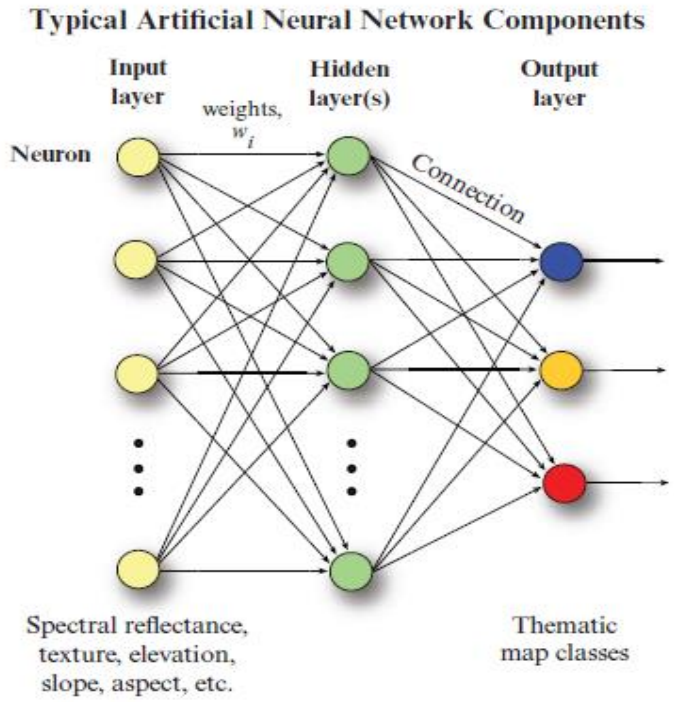



- 뉴런의 출력값은 y로 입력값(xi)과 각각의 가중치(wi)를 곱한 것의 합에 편향(b)을 더하여 활성 함수(activation function)를 적용한 값

- 신경망은 **입력층, 은닉층, 출력층**으로 구성되며 각 층은 노드를 통해 서로 연결됨

- 상호 연결을 통해 정보는 다중 방향으로 흐르게 되고 신경망이 훈련됨

- **상호 연결의 세기(혹은 가중치)**는 신경망에 의해 학습되고 저장되며 검증 단계에서 사용됨




출처: Jensen, 2016

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 인공신경망

**지도학습**

- **Multilayer Feedforward Backpropagation Network**
- Radial Basis Function (RBF) Network
- Convolutional Neural Network (CNN)
- Recurrent Neural Network (RNN) 등

비지도학습

- Self Organizing Map (SOM)
- Autoencoder
- Generative Adversarial Networks (GAN) 등

강화학습

- Finding the balance between exploration (new territory) and exploitation (known areas/knowledge)

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 인공신경망

성능과 정확도에 영향을 미치는 요소들

- **은닉층과 뉴런의 개수**
- 활성 함수
- 학습 알고리즘
- 학습률 등과 같은 매개변수 등

```
#ANN
from sklearn.neural_network import MLPRegressor

mymodel = MLPRegressor(hidden_layer_sizes = (100), \
                       activation = 'relu', \
                       solver = 'sgd', \
                       alpha = 0.1, \
                       learning_rate = 'constant', \
                       learning_rate_init = 0.001, \
                       tol = 1e-2, \
                       max_iter = 1000, \
                       momentum = 0.9, \
                       random_state = 42, \
                       )
```

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 인공신경망

성능과 정확도에 영향을 미치는 요소들

- 은닉층과 뉴런의 개수
- **활성 함수**
- 학습 알고리즘
- 학습률 등과 같은 매개변수 등

```
#ANN
from sklearn.neural_network import MLPRegressor

mymodel = MLPRegressor(hidden_layer_sizes = (100), \
                       activation = 'relu', \
                       solver = 'sgd', \
                       alpha = 0.1, \
                       learning_rate = 'constant', \
                       learning_rate_init = 0.001, \
                       tol = 1e-2, \
                       max_iter = 1000, \
                       momentum = 0.9, \
                       random_state = 42, \
                       )
```

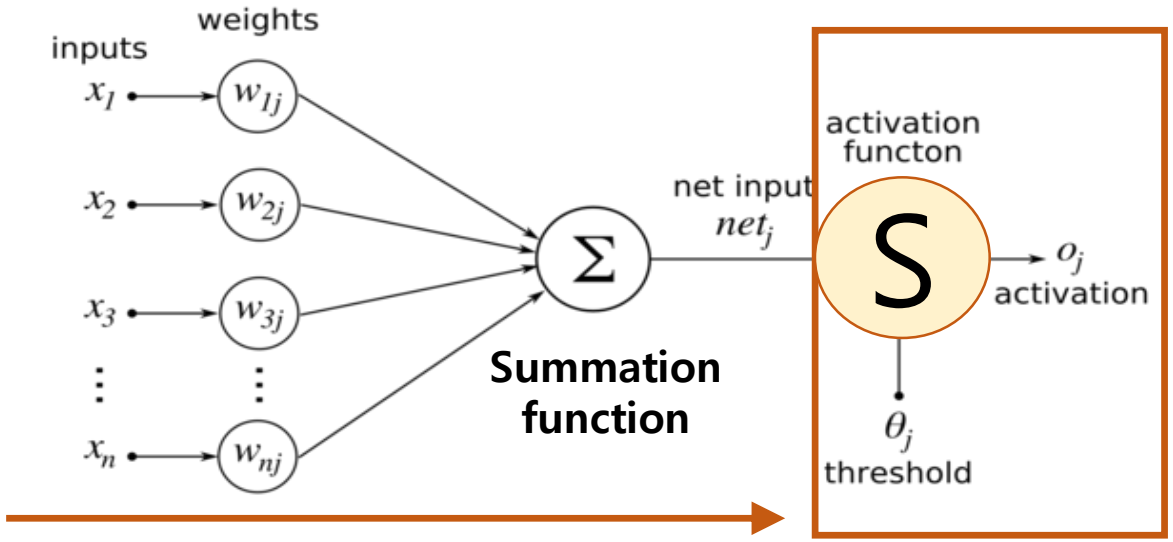
# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 활성 함수

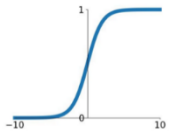
자료의 비선형성을 다루기 위해
선형 자료를 비선형 자료로 전환

임계점 이상이면 활성

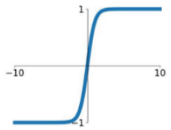



출처: UNIST IRIS Lab

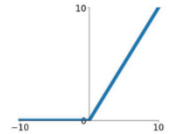
**Sigmoid**
$\sigma(x) = \frac{1}{1+e^{-x}}$



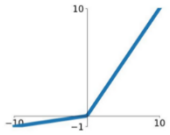
**tanh**
$\tanh(x)$



**ReLU**
$\max(0, x)$



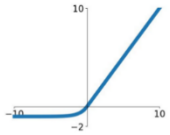
**Leaky ReLU**
$\max(0.1x, x)$



**Maxout**
$\max(w_1^T x + b_1, w_2^T x + b_2)$

**ELU**

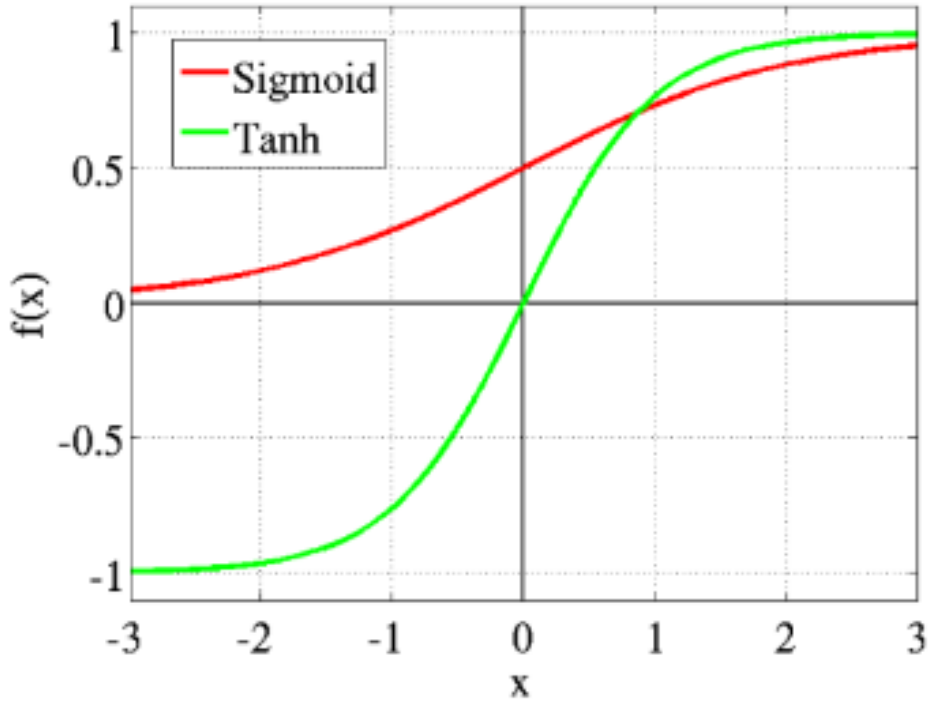
$$\begin{cases} x & x \geq 0 \\ \alpha(e^x - 1) & x < 0 \end{cases}$$



# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 활성 함수




$$h(x) = \frac{1}{1 + \exp(-x)}$$

$$h(x) = \tanh(x) = \frac{e^{x} - e^{-x}}{e^{x} + e^{-x}}$$

Sigmoid (Logistic) / tanh (Hyperbolic tangent):
기울기 소실 문제 - x값이 크거나 작은 경우 기울기가 0에 수렴

- 분류, 회귀에 모두 사용됨: 회귀에서는 은닉층에서만 사용되며 마지막 출력층은 활성 함수 사용하지 않거나 활성 함수 결과에 가중치 곱하여 사용
- 각 층마다 다른 활성 함수 활용할 수 있음

출처: UNIST IRIS Lab

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 인공신경망

성능과 정확도에 영향을 미치는 요소들

- 은닉층과 뉴런의 개수
- 활성 함수
- **학습 알고리즘**
- 학습률 등과 같은 매개변수 등

```
#ANN
from sklearn.neural_network import MLPRegressor

mymodel = MLPRegressor(hidden_layer_sizes = (100), \
                       activation = 'relu', \
                       solver = 'sgd', \
                       alpha = 0.1, \
                       learning_rate = 'constant', \
                       learning_rate_init = 0.001, \
                       tol = 1e-2, \
                       max_iter = 1000, \
                       momentum = 0.9, \
                       random_state = 42, \
                       )
```

비용 함수를 최소화 하는 가중치와 편향 찾기

- 비용함수는 일반적으로 실제값과 예측값의 제곱오차
- 전역 최대/최소를 찾기 위해 반복적으로 업데이트
- 어느 방향으로 가중치와 편향을 업데이트 해야 할지, 얼마나 움직여야 할지 중요

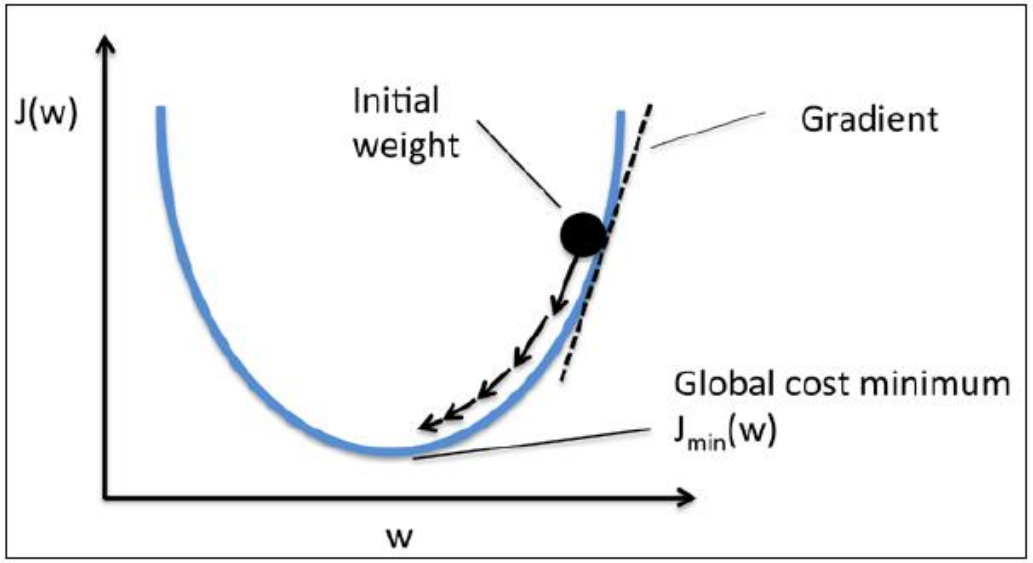
- 경사 하강법(Gradient descent)
- 다양한 경사 하강법의 변형들
  - SGD: Stochastic gradient descent (확률적 -)
  - Momentum: SGD with momentum
  - NAG: Nesterov Accelerated Gradient
  - Adagrad: Adaptive Gradient
  - Adadelta: Adaptive Delta
  - Rmsprop
- Quasi-Newton
- Levenberg-Marquardt 등

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 인공신경망

- 경사하강법(Gradient Descent)
  - 산을 내려가려는데 안개가 끼어 바로 주변만 보일때… 자기 위치에서 가장 큰 경사가 쪽으로 내려가야 함 → 이런 방식으로 주어진 곡선의 최소값을 찾아가는 방식이 경사하강법




비용함수 예시

$$J(w) = \frac{1}{2}\sum_i \left(y - \sum_j w_j x_j\right)^2$$

업데이트된 가중치

$$w_j = w_j + \Delta w_j = w_j - \eta \Delta J(w_j)$$

경사 즉 기울기

$$w_j = w_j + \eta \sum_i (y^{(i)} - y^{\wedge(i)}) x_j^{(i)}$$

출처: samsjang 블로그

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 인공신경망

성능과 정확도에 영향을 미치는 요소들

- 은닉층과 뉴런의 개수
- 활성 함수
- 학습 알고리즘
- **학습률 등과 같은 매개변수 등**

```
#ANN
from sklearn.neural_network import MLPRegressor

mymodel = MLPRegressor(hidden_layer_sizes = (100), \
                       activation = 'relu', \
                       solver = 'sgd', \
                       alpha = 0.1, \
                       learning_rate = 'constant', \
                       learning_rate_init = 0.001, \
                       tol = 1e-2, \
                       max_iter = 1000, \
                       momentum = 0.9, \
                       random_state = 42, \
                       )
```

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 트리를 이용한 기계학습 - 결정트리(Decision Trees)

- 자기 발견적인 지식기반 전문가 시스템(Expert Systems)을 이용한 문제해결 방식

전문가의 머릿속에 저장되어 있는 도메인 지식은 가설(문제), 규칙, 조건들로 구성된 지식기반의 형태로 추출


사용자 인터페이스와 추론 엔진은 지식기반 규칙들을 암호화

온라인 데이터베이스로부터 필요한 정보를 추출하여 문제를 해결하는데 사용

출처: Jensen, 2016

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 트리를 이용한 기계학습 - 결정트리(Decision Trees)

- 인간 도메인 전문가와 지식공학자에 기반한 전문가 시스템

"그 지형은 태양에너지를 최대한 이용하는 주거지 개발에 적합하다, 즉 지붕에 태양패널을 설치할 수 있을 것이다"

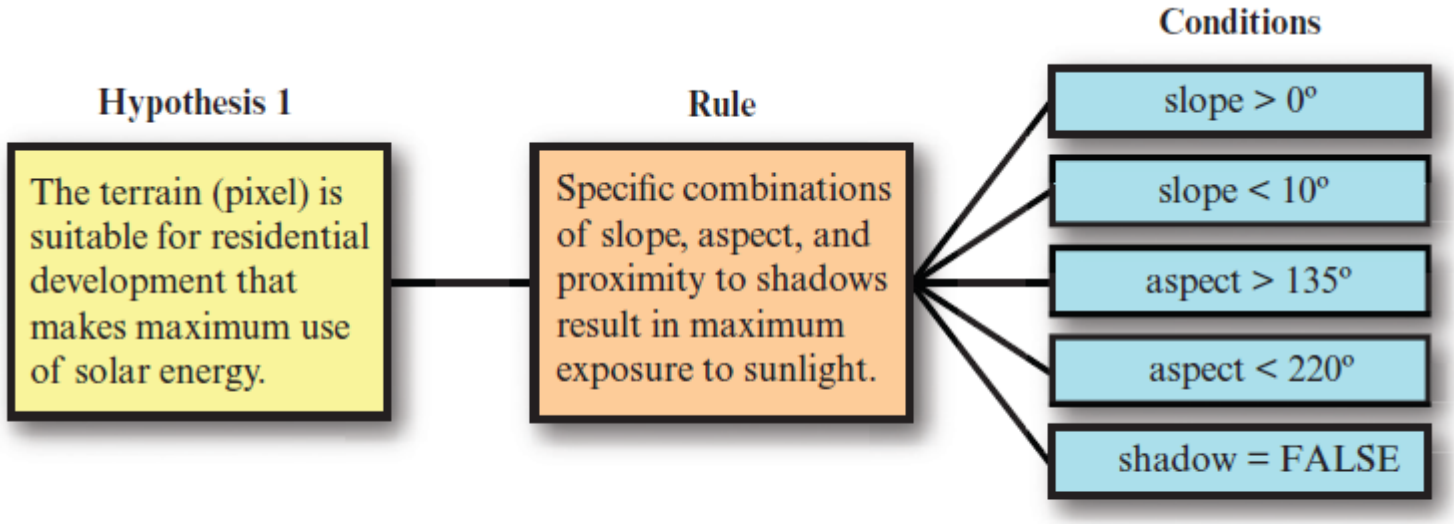



FIGURE 10–4 A human-derived decision-tree expert system with a rule and conditions to be investigated by an inference engine to test hypothesis 1.

출처: Jensen, 2016

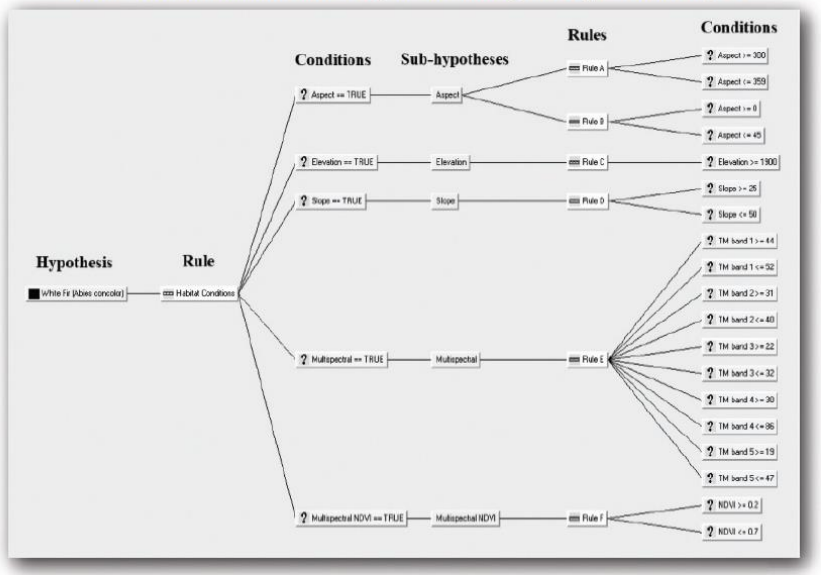
# II. 기계학습을 이용한 장기예측

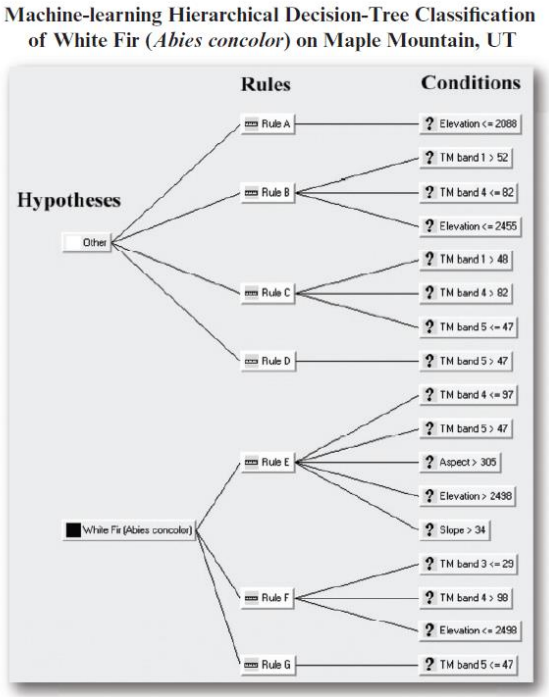
## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 지식기반 구축의 자동화: 기계학습

- 기계학습에서는 귀납적 학습과정을 통해 전문가 시스템을 위한 지식기반을 구축
- “좋은 훈련자료가 있다면 인간 전문가로부터 일반 이론을 명확히 추출하는 것 보다 훨씬 쉽다.”

북미 전나무 분류:
인간 전문가




북미 전나무 분류:
결정트리

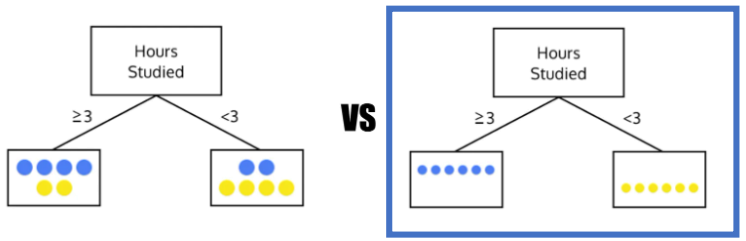
출처: Jensen, 2016

# II. 기계학습을 이용한 장기예측

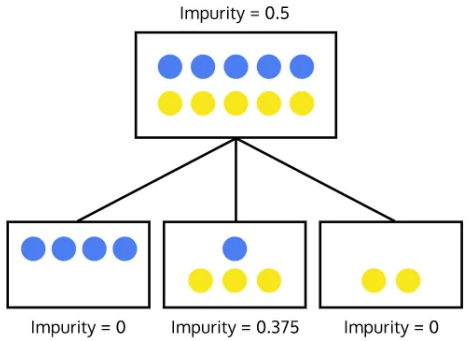
## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 결정트리(Decision Trees)의 학습 원리

- 순도(homogeneity)가 증가하고 불순도(impurity) 혹은 불확실성(uncertainty)이 최대한 감소하는 방향으로 학습을 진행
- 순도가 증가 / 불확실성이 감소 = 정보획득(information gain)




불순도가 감소한 방향




Information Gain = Mother - Children
: 0.5 − (0 + 0.375 + 0) = 0.125

정보획득량(Information Gain)
이 커지는 방향으로 가지치기

출처: UNIST IRIS Lab

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 트리를 이용한 기계학습 - 결정트리(Decision Trees)

전문가 시스템을 위한 지식기반을 자동적으로 구축하기 위해 귀납적 학습 기법을 적용

성능과 정확도에 영향을 미치는 요소들

- 트리의 최대 깊이(None인 경우 최대로 성장)
- 가지가 나누어질 수 있는 잎의 수
- 가지를 나눌 때 사용할 입력변수의 최대 개수 ('sqrt ' 인 경우 전체 입력변수 수의 제곱근)

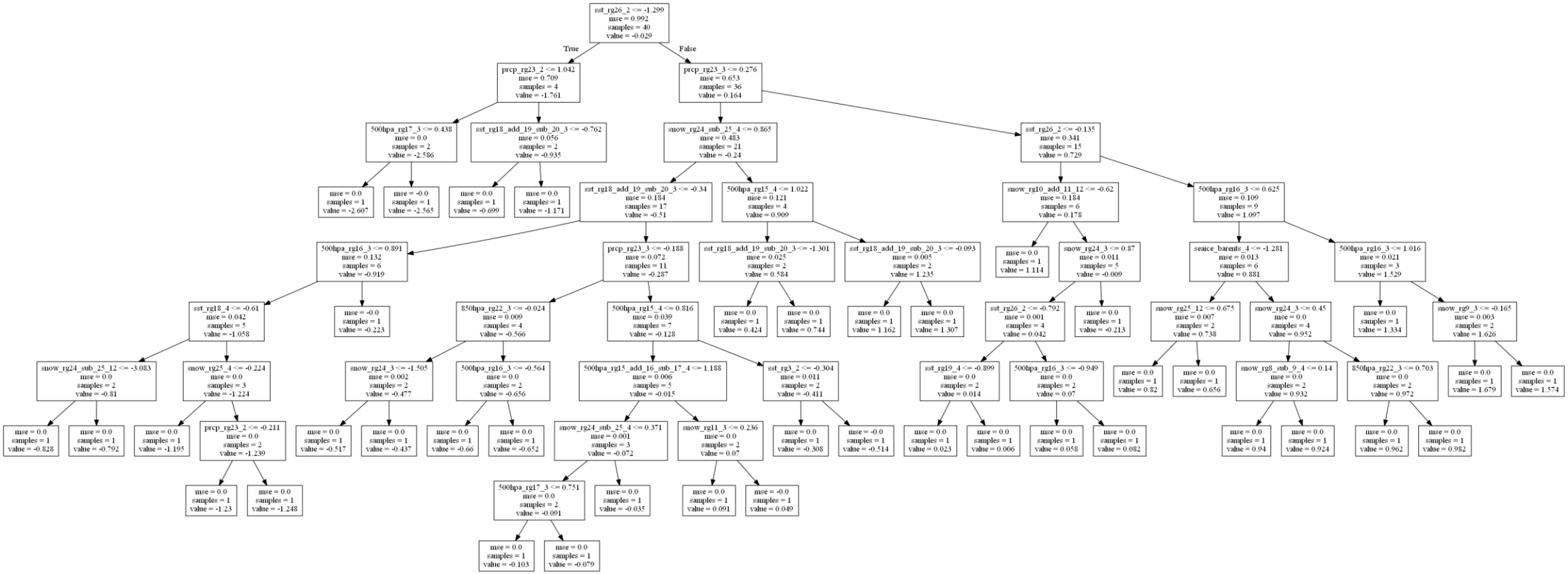
```
#DT
from sklearn.tree import DecisionTreeRegressor

mymodel = DecisionTreeRegressor(criterion = 'mse', \
                                max_depth = None, \
                                min_samples_split = 2, \
                                max_features = None, \
                                random_state = 42, \
                                )
```

출처: Jensen, 2016

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

2020년 8월 평균기온 예측 트리



# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 트리를 이용한 기계학습 - 결정트리(Decision Trees)

결정트리의 장점

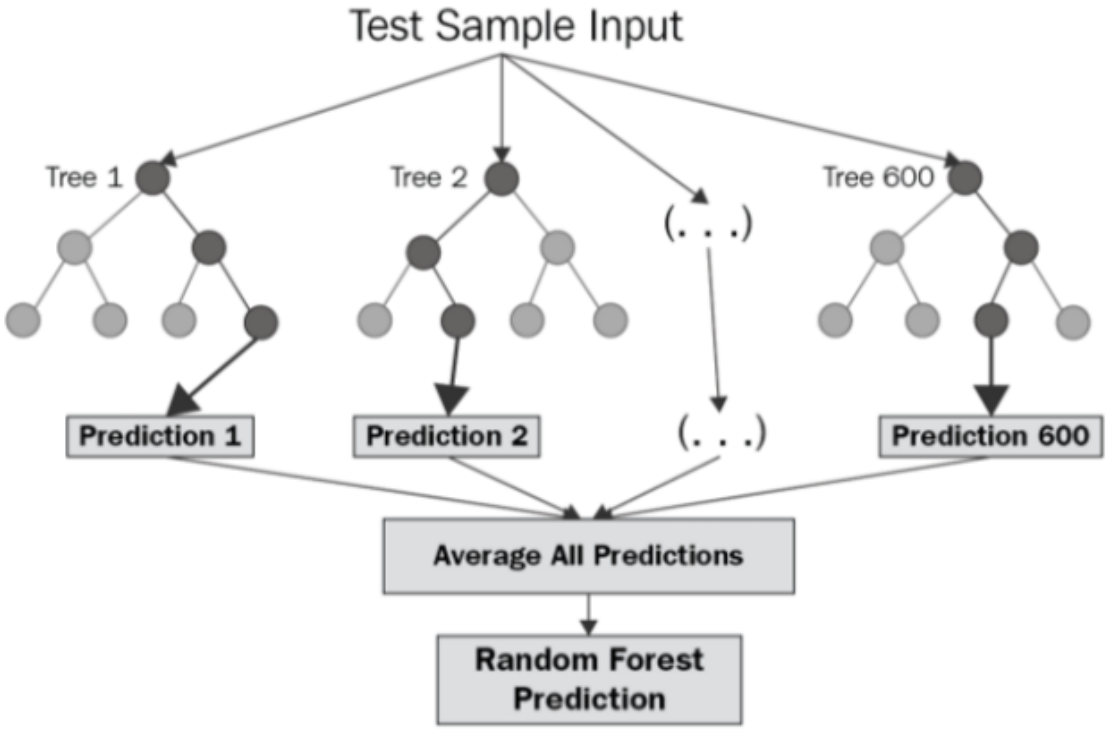
- 훈련 자료로부터 직접 규칙과 조건들을 생성하여 전문가 시스템을 훈련
- 전문가 시스템의 결과를 평가할 수 있고 결론에 어떻게 도달했는지 역으로 확인할 수 있음
- 입력변수의 분포에 대한 어떠한 가정도 필요 없음
- 입력변수 사이의 비선형적 및 계층적 관계를 밝힐 수 있음

출처: Jensen, 2016

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 앙상블 기법 - 랜덤포레스트(Random Forests)

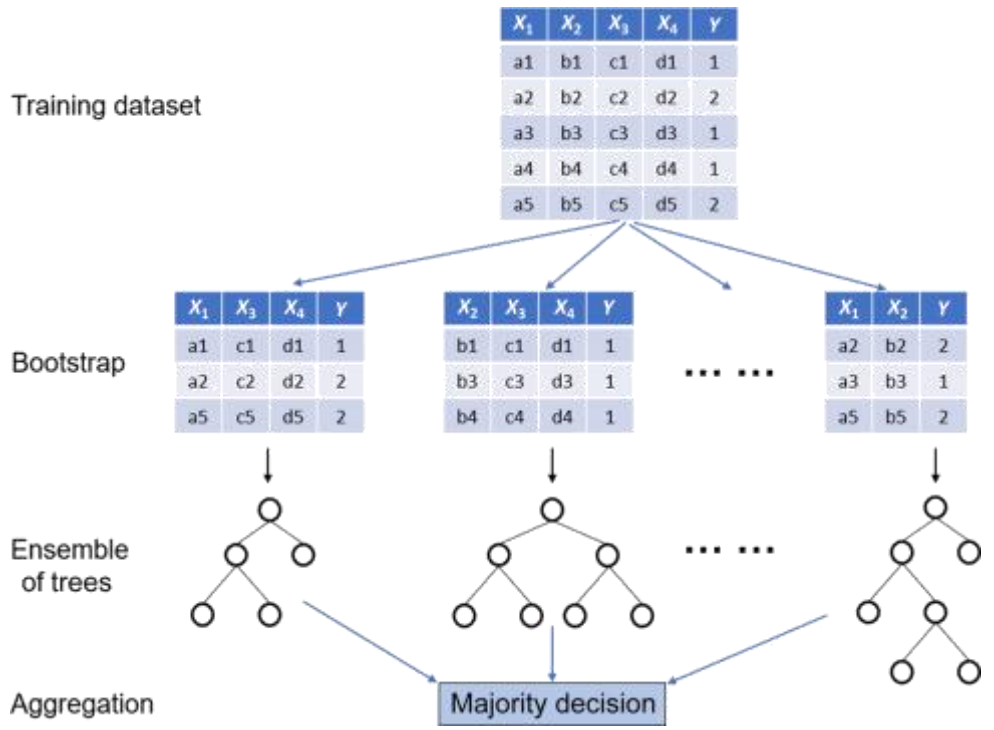
여러 결정트리를 취합하여 학습 성능을 높이는 앙상블 기법



# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 앙상블 기법 - 랜덤포레스트(Random Forests)




자료의 부트스트랩(Bootstrap) 랜덤 샘플링
- 각 트리에 대해 사용할 자료를 N개 무작위 복원 추출
- 예를 들어 전체 자료의 대략 70% 사용; 나머지 30%는 OOB (out-of-bag)

입력변수의 랜덤 샘플링
- 각 트리에서 가지를 나눌 때마다 입력변수 중 m개를 무작위 복원 추출
- 전체 입력변수가 M개라면 예를 들어 M의 제곱근을 사용

다수의 트리에 대해 수행하여 결과 취합 (voting, 평균)

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 앙상블 기법 - 랜덤포레스트(Random Forests)

성능과 정확도에 영향을 미치는 요소들

- Base Estimator는 결정트리
  - 트리의 최대 깊이(None인 경우 최대로 성장)
  - 가지가 나누어질 수 있는 잎의 수
  - 가지를 나눌 때 사용할 입력변수의 최대 개수 ('sqrt ' 인 경우 전체 입력변수 수의 제곱근)
- 트리의 개수

```
#RF
from sklearn.ensemble import RandomForestRegressor

mymodel = RandomForestRegressor(criterion = 'mse', \
                                n_estimators = 100, \
                                max_depth = None, \
                                min_samples_split = 2, \
                                max_features = None, \
                                bootstrap = True, \
                                random_state = 42
                                )
```

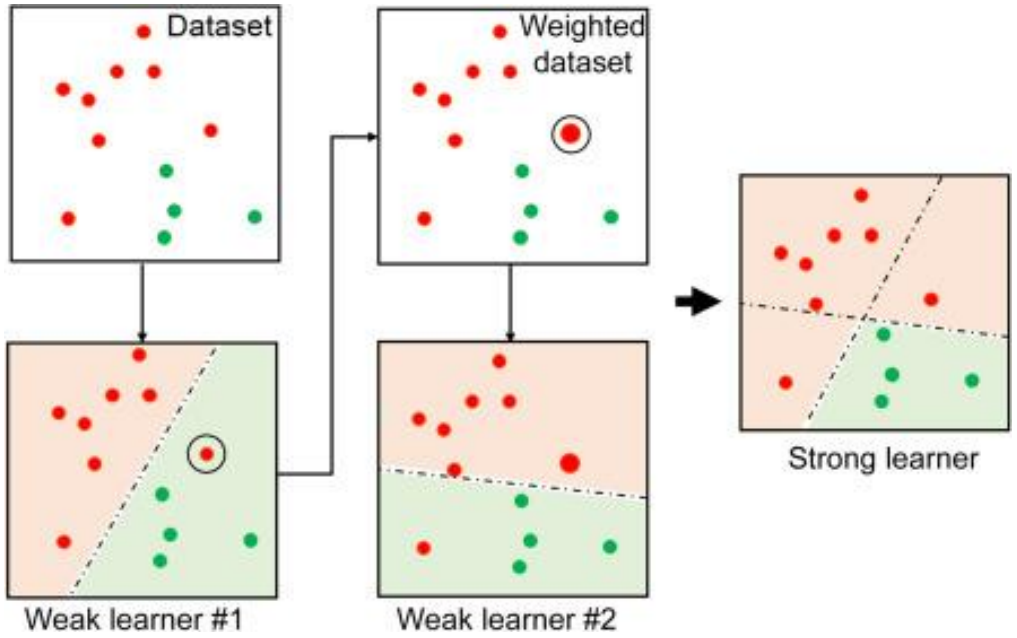
## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 부스팅 기법 - 에이다부스트(Adaptive Boosting; AdaBoost)

여러 학습 알고리즘의 결과물에 가중치를 주어
학습 성능을 높이는 가속화(부스팅) 기법

메타 알고리즘

에이다부스트는 약한 학습기(weak learner)의
결과물에 가중치를 주어 성능을 향상시킴




출처: Misra and Li, 2020

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 부스팅 기법 - 에이다부스트(Adaptive Boosting; AdaBoost)

성능과 정확도에 영향을 미치는 요소들

- Base Estimator는 결정트리
  - 트리의 최대 깊이(None인 경우 최대로 성장)
  - 가지가 나누어질 수 있는 잎의 수
  - 가지를 나눌 때 사용할 입력변수의 최대 개수('sqrt ' 인 경우 전체 입력변수 수의 제곱근)
- 트리의 개수
- 학습률: 새로운 학습자가 얼마나 기여하도록 할지 정함(0~1)
- 손실 함수: 가중치 업데이트할 때 사용

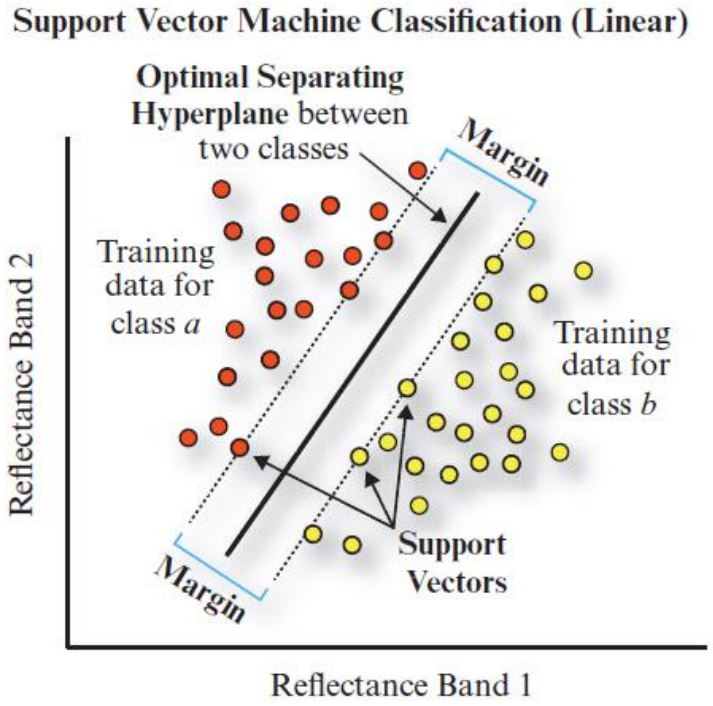
```
#Adaboost
from sklearn.tree import DecisionTreeRegressor
from sklearn.ensemble import AdaBoostRegressor

mydt = DecisionTreeRegressor(criterion = 'mse', \
                             max_depth = 5, \
                             min_samples_split = 2, \
                             max_features = None, \
                             random_state = 42, \
                             )
mymodel = AdaBoostRegressor(base_estimator = mydt, \
                             n_estimators = 500, \
                             learning_rate = 1, \
                             loss = 'square', \
                             random_state = 42
                             )
```

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

### 서포트 벡터 머신(Support Vector Machine)

- 훈련 자료를 이용해 클래스 간을 분리하는 최적의 초평면(hyperplane)을 찾음: 클래스의 가장 가까운 훈련 샘플들 사이의 마진을 최대화
- 경계에 놓여 있는 포인트들이 **서포트 벡터**라고 불리며 그 마진의 중간이 클래스를 분리하는 최적의 초평면
- 잘못 분리된 훈련 자료들에 음의 가중치를 주어 영향을 줄임



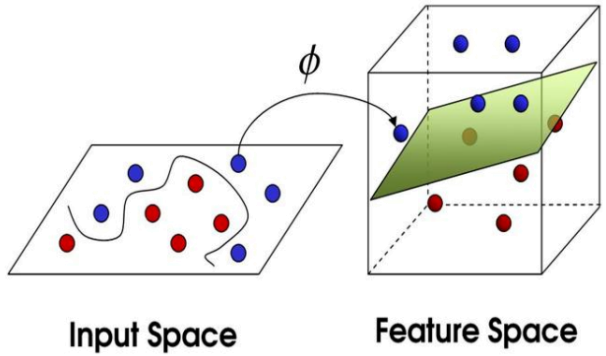

출처: Jensen, 2016

# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고 모델 학습을 위한 최적화 방법을 정하자

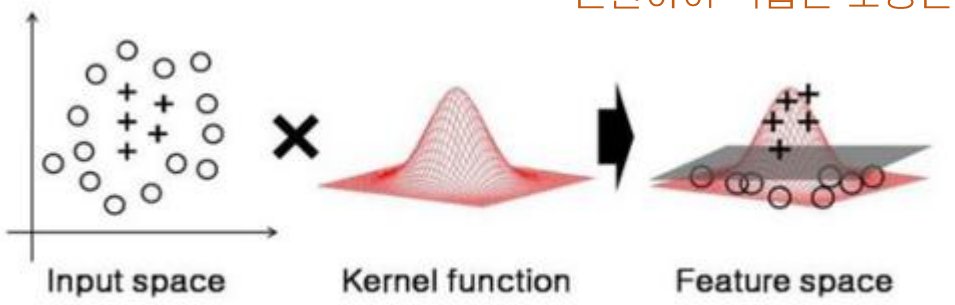
### 서포트 벡터 머신(Support Vector Machine)

- 선형의 초평면을 찾는 것이 불가능할 경우 커널 함수를 이용하여 고차원 공간으로 변환하여 적합한 초평면을 찾음
- 널리 이용되는 SVM 커널 함수는 선형, 다항식, 가우시안(RBF), 시그모이드 등이 있음
- 클래스의 경계에서 선택된 적은 훈련 자료만으로도 꽤 정확한 분류를 수행함




출처: UNIST IRIS Lab

RBF Kernel을 이용하여 3차원 공간으로 변환하여 적합한 초평면을 찾음




출처: bskyvision 블로그

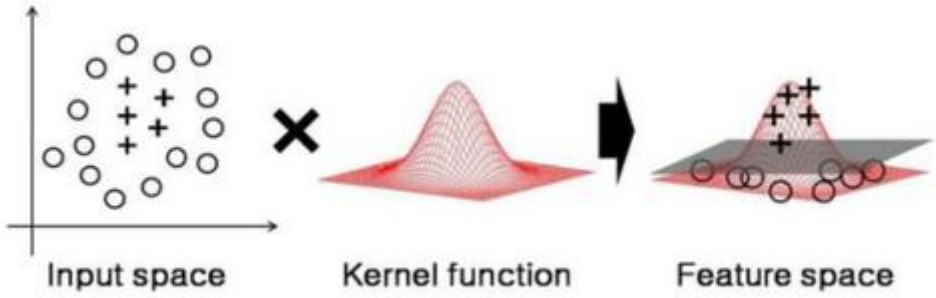
# II. 기계학습을 이용한 장기예측

## 활용할 기법을 선정하고
## 모델 학습을 위한 최적화 방법을 정하자

### 서포트 벡터 회귀(Support Vector Regression)

```
#SVR
from sklearn.svm import SVR

mymodel = SVR(kernel = 'rbf', \
              gamma = 'scale', \
              C = 1.0, \
              epsilon = 0.1, \
              )
```

- Kernel: 알고리즘에서 사용하는 커널 타입
- Gamma: 커널의 계수
  - 하나의 자료 샘플이 영향력을 행사하는 거리를 결정
  - RBF의 경우 표준편차와 반비례하는 계수로 gamma가 크면 영향력을 행사하는 거리가 짧아지고 gamma가 낮으면 커짐
  - 'scale ' 로 지정하면 1 / (입력 변수의 개수 * X의 분산) 이용
- C: 정규화 매개변수
  - 정규화 강도는 C에 반비례
  - C가 너무 낮으면 과소적합, 너무 크면 과대적합 가능성
- Epsilon: 실제값과 예측값 차이 허용 범위




출처: bskyvision 블로그

# II. 기계학습을 이용한 장기예측

## 모델 평가를 위한 기준을 정하자

### 단정예측을 위한 모델 평가

$$r_{XY} = \frac{\frac{\sum_i^n (X_i - \overline{X})(Y_i - \overline{Y})}{n-1}}{\sqrt{\frac{\sum_i^n (X_i - \overline{X})^2}{n-1}} \sqrt{\frac{\sum_i^n (Y_i - \overline{Y})^2}{n-1}}}$$

- 피어슨 상관계수
- 평균절대오차(Mean Absolute Error; MAE)

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} |y_i - x_i|}{n} = \frac{\sum_{i=1}^{n} |e_i|}{n}$$

TABLE B1. Contingency table for tercile forecasts.

| | Observed category | | | |
|---|---|---|---|---|
| Predicted category | AN | NN | BN | |
| AN | *a* | *b* | *c* | $\mathrm{PRED}_{AN}$ |
| NN | *d* | *e* | *f* | $\mathrm{PRED}_{NN}$ |
| BN | *g* | *h* | *i* | $\mathrm{PRED}_{BN}$ |
| | $\mathrm{OBS}_{AN}$ | $\mathrm{OBS}_{NN}$ | $\mathrm{OBS}_{BN}$ | Total, *n* |

### 삼분위로 나타냈을 때

- Heidke Skill Score (HSS)
- Total Accuracy = Proportion Correct
- Hit Rate = Probability of Detection
- False Alarm Rate = Probability of False Detection

$$\mathrm{HSS} = \frac{\sum_k p(\mathrm{PRED}_k, \mathrm{OBS}_k) - \sum_k p(\mathrm{PRED}_k) p(\mathrm{OBS}_k)}{1 - \sum_k p(\mathrm{PRED}_k) p(\mathrm{OBS}_k)}$$

$$PC = \sum_k p(\mathrm{PRED}_k, \mathrm{OBS}_k)$$

# II. 기계학습을 이용한 장기예측

## Toy Example 단정예측 결과

### 인공신경망

| Month | HSS | Total Accuracy | r | MAE | HR (AN) | HR (NN) | HR (BN) |
|---|---|---|---|---|---|---|---|
| Jun | 0.40 | 0.61 | 0.46 | 0.45 | 0.50 | 0.56 | 0.73 |
| Jul | 0.23 | 0.49 | 0.42 | 0.85 | 0.33 | 0.54 | 0.56 |
| Aug | 0.16 | 0.44 | 0.33 | 0.91 | 0.43 | 0.54 | 0.36 |

8월 평균기온 예측결과(단정예측을 삼분위에 적용)

| YEAR | TRUE | PRED |
|---|---|---|
| 2016 | AN | AN |
| 2017 | NN | NN |
| 2018 | AN | NN |
| 2019 | AN | BN |
| 2020 | AN | AN |

### 트리를 이용한 기계학습 - 결정트리(Decision Trees)

| Month | HSS | Total Accuracy | r | MAE | HR (AN) | HR (NN) | HR (BN) |
|---|---|---|---|---|---|---|---|
| Jun | -0.01 | 0.34 | 0.27 | 0.63 | 0.00 | 0.38 | 0.53 |
| Jul | 0.19 | 0.46 | 0.28 | 1.18 | 0.50 | 0.38 | 0.50 |
| Aug | 0.27 | 0.51 | 0.23 | 1.07 | 0.43 | 0.62 | 0.50 |

| YEAR | TRUE | PRED |
|---|---|---|
| 2016 | AN | NN |
| 2017 | NN | NN |
| 2018 | AN | AN |
| 2019 | AN | AN |
| 2020 | AN | BN |

# II. 기계학습을 이용한 장기예측

## Toy Example 단정예측 결과

### 앙상블 기법 - 랜덤포레스트(Random Forests)

| Month | HSS | Total Accuracy | r | MAE | HR (AN) | HR (NN) | HR (BN) |
|---|---|---|---|---|---|---|---|
| Jun | 0.09 | 0.44 | 0.40 | 0.47 | 0.00 | 0.81 | 0.33 |
| Jul | 0.30 | 0.54 | 0.41 | 0.85 | 0.33 | 0.69 | 0.56 |
| Aug | 0.24 | 0.49 | 0.51 | 0.81 | 0.29 | 0.77 | 0.43 |

### 부스팅 기법 - 에이다부스트(Adaptive Boosting; AdaBoost)

| Month | HSS | Total Accuracy | r | MAE | HR (AN) | HR (NN) | HR (BN) |
|---|---|---|---|---|---|---|---|
| Jun | 0.10 | 0.44 | 0.45 | 0.48 | 0.00 | 0.69 | 0.47 |
| Jul | 0.27 | 0.51 | 0.37 | 0.89 | 0.42 | 0.54 | 0.56 |
| Aug | 0.20 | 0.46 | 0.45 | 0.84 | 0.29 | 0.69 | 0.43 |

### 서포트 벡터 회귀(Support Vector Regression)

| Month | HSS | Total Accuracy | r | MAE | HR (AN) | HR (NN) | HR (BN) |
|---|---|---|---|---|---|---|---|
| Jun | 0.11 | 0.44 | 0.49 | 0.43 | 0.10 | 0.63 | 0.47 |
| Jul | 0.27 | 0.51 | 0.42 | 0.80 | 0.33 | 0.69 | 0.50 |
| Aug | 0.20 | 0.46 | 0.49 | 0.81 | 0.21 | 0.77 | 0.43 |

8월 평균기온 예측결과(단정예측을 삼분위에 적용)

| YEAR | TRUE | PRED |
|---|---|---|
| 2016 | AN | AN |
| 2017 | NN | NN |
| 2018 | AN | NN |
| 2019 | AN | NN |
| 2020 | AN | NN |

| YEAR | TRUE | PRED |
|---|---|---|
| 2016 | AN | AN |
| 2017 | NN | NN |
| 2018 | AN | NN |
| 2019 | AN | BN |
| 2020 | AN | NN |

| YEAR | TRUE | PRED |
|---|---|---|
| 2016 | AN | NN |
| 2017 | NN | NN |
| 2018 | AN | NN |
| 2019 | AN | NN |
| 2020 | AN | AN |

# II. 기계학습을 이용한 장기예측

## 풀고 싶은 문제를 정의하자: 확률예측


확률분포 예측(베이지안)
→ 각 카테고리별 확률

**5월(ex. 5/15)에 예측한 6, 7, 8월 평균 기온 확률 예측**

기온 값 자체를 예측
→ 단정예측

5월(ex. 5/15)에 예측한 6, 7, 8월 평균 기온 예측

삼분위 카테고리로 분류
→ 각 카테고리별 확률
→ 확률예측

# II. 기계학습을 이용한 장기예측

## 확률과 불확실성

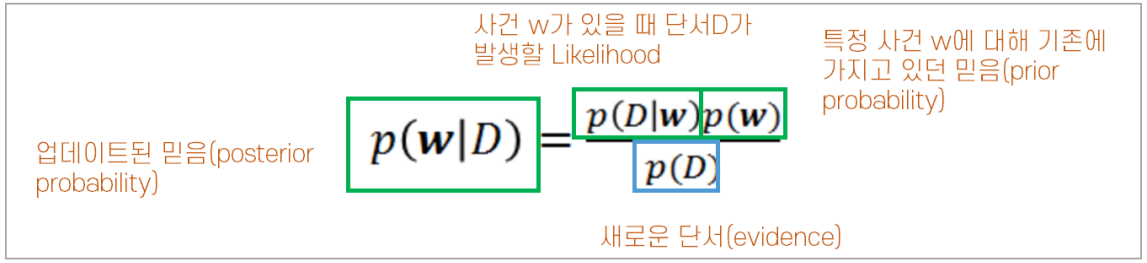
### 경험적 확률(빈도주의) (여러 번 되풀이 할 수 있는) 사건의 빈도

동전던지기
앞면이 나올 확률은?

$$P(\text{앞면}) = \frac{\text{앞면이 나온 수}}{\text{지금까지 던진 수}}$$

### 베이지안 확률 (여러 번 되풀이 할 수 없는 경우도 포함하여) 사건에 대한 믿음

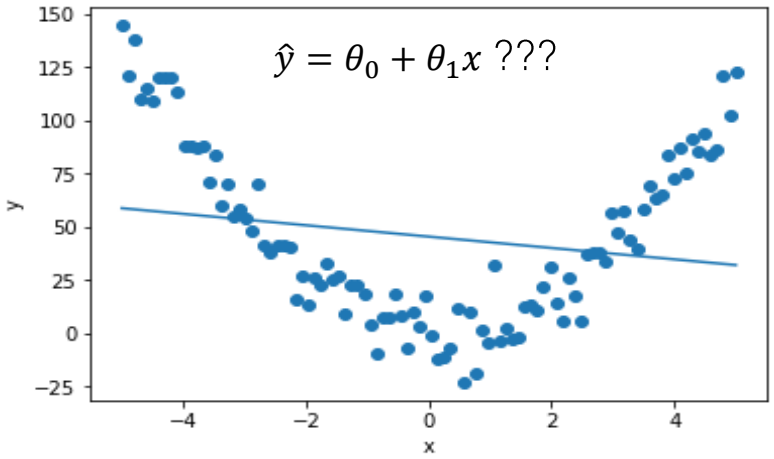
백두산 화산 폭발의 확률은?
확률 = 믿음

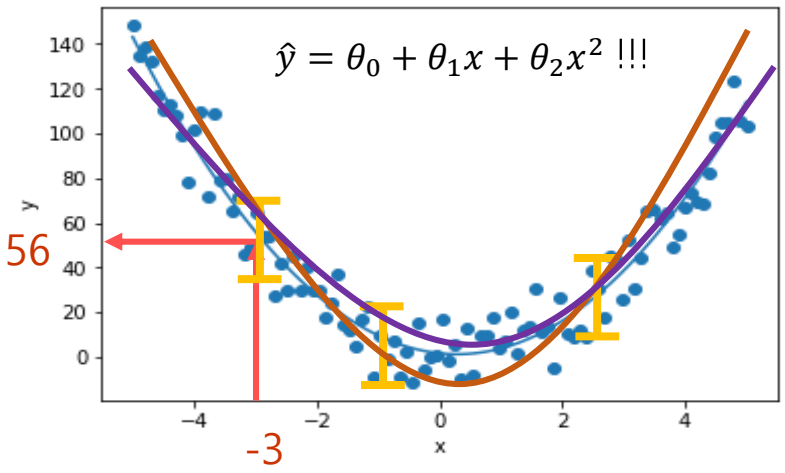



사건과 관련된 여러 확률을 이용해 새로 일어날 사건의 확률을 추정
이러한 단서들이 추가되면 사건에 대한 불확실성을 정량화

## 베이지안 추론

### 전통적 추론(Classical Inference)



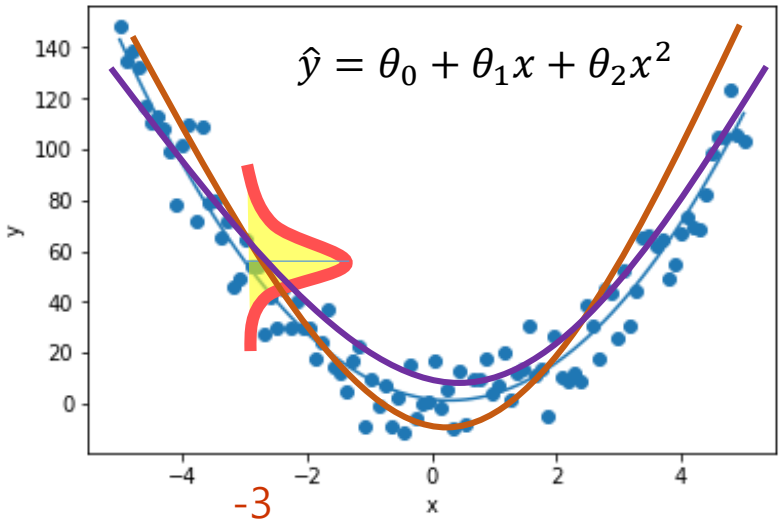




- 우리의 목적: 새로운 $X_*$에 대한 y 값 추정하도록 → 모집단의 특성을 파악(즉 적합한 함수 도출)
- 전통적 추론: 관측된 자료(훈련 자료) 에 관한 함수를 사용하여 그 함수가 가지는 미지의 모수를 추론
- 하지만 전통적 추론에 의한 결과는 자료에 대한 불확실성 정보를 알려주지 않음!

# II. 기계학습을 이용한 장기예측

## 베이지안 추론

### 베이지안 추론(Bayesian Inference)




- 우리가 추론하고자 하는 모수는 불확실하며 이 불확실성의 정도를 확률모델로 표현하고 싶음
- 모수 각각에 대한 확률모델과 관측된 자료를 사용
- 확률예측을 통해 불확실성 정보를 제공

관측된 자료 $y$ 에 대한 확률모델

사후확률 $\mathbf{p}(\boldsymbol{\theta}|\boldsymbol{y}) = \dfrac{\boldsymbol{p}(\boldsymbol{y}|\boldsymbol{\theta})\,\boldsymbol{p}(\boldsymbol{\theta})}{\boldsymbol{p}(\boldsymbol{y})}$ 사전확률

- 모수에 관한 과거의 경험이나 사전 지식 같은 주관적 견해를 수량화한 모수의 특성(사전확률분포)을
- 자료로부터 얻은 모수에 관한 정보(관측된 자료)와 결합하여
- 사후확률분포를 얻음(확률예측결과를 제공)

$\theta_0, \theta_1, \theta_2$ 각각의 사전확률분포
입력 $X$ 대해 관측된 자료 $y$

↓

$\theta_0, \theta_1, \theta_2$ 각각의 사후확률분포

↓

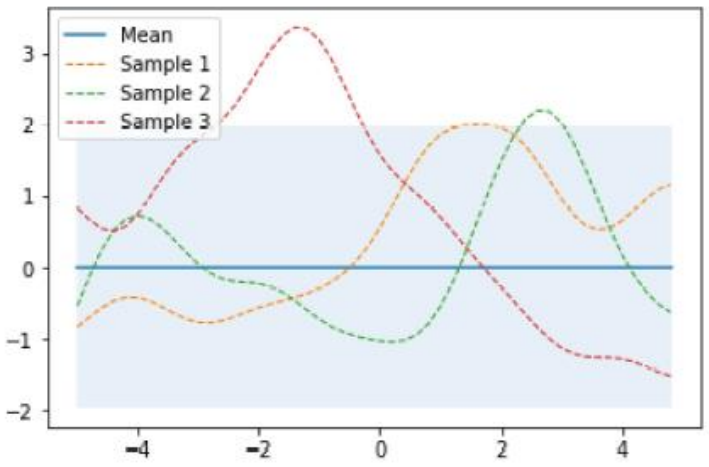
새로운 $X_*$에 대한 예측 값 $\hat{y}$의 분포

# II. 기계학습을 이용한 장기예측

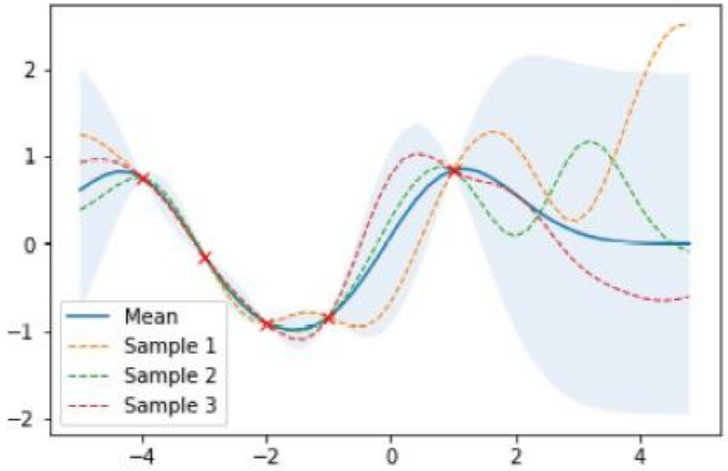
## 가우시안 프로세스: 비모수 베이지안 모델

- 앞서 본 예시처럼 어떤 함수를 사용할지(모수가 몇 개인지, 모수가 어떤 분포를 따를지) 규정하는 대신
- "모든 가능한 함수"에 사전확률을 주고 더 그럴듯한 함수에 높은 사후확률을 주면 어떨까?

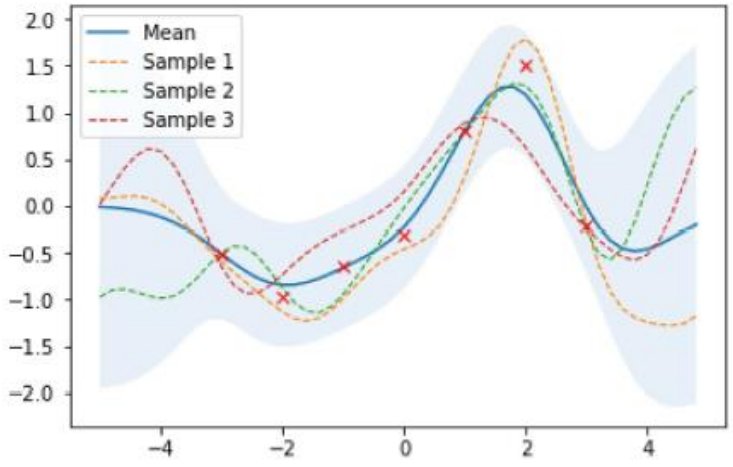
- 각각의 함수를 특정한 입력 x 에서 함수값 f(x)를 가지는 굉장히 긴 벡터로 생각하고,
- 무한한 모든 함수를 고려할 필요 없이 관측된 유한한 지점에서의 함수값의 특성만을 보면 됨




Three samples from prior




Three samples from posterior
Noise-free data



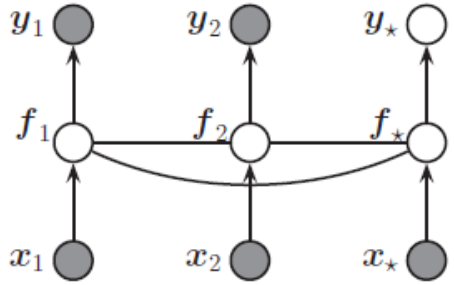

Three samples from posterior
Noisy data

출처: martin Krasser github

# II. 기계학습을 이용한 장기예측

## 가우시안 프로세스 모델

Gaussian Process는 확률과정(stochastic process)으로 임의의 점 $x \in R^d$ 에서 확률변수 $f(x)$를 가지며 유한한 확률변수들의 결합확률분포 $p(f(x_1), \ldots, f(x_N))$ 는 Gaussian




$$p(f|X) = N(f|\mu, \mathrm{K})$$

$\mu(x)$ 평균 함수

$K_{ij} = \kappa(x_i, x_j)$ 공분산함수(커널)

$x_i$ 와 $x_j$ 가 가까우면 $f(x_i)$와 $f(x_j)$도 가까울 것이라 가정 ➡ 공분산함수(커널)에 의해 결정

# II. 기계학습을 이용한 장기예측

## 가우시안 프로세스 모델

이미 관측한 자료 y와 예측 $f_*$ 의 결합분포 $\begin{pmatrix} \mathbf{y} \\ \mathbf{f}_* \end{pmatrix} \sim N\left[0, \begin{pmatrix} \mathbf{K}_y & \mathbf{K}_* \\ \mathbf{K}_*^{\mathrm{T}} & \mathbf{K}_{**} \end{pmatrix}\right]$

GP 사전분포 $p(\mathbf{f}|\mathbf{X}) = N(\mathbf{f}|\boldsymbol{\mu}, \mathbf{K})$ GP 사후분포 $p(\mathbf{f}|\mathbf{X}, \mathbf{y})$

$\mathbf{K}_y = \kappa(\mathbf{X}, \mathbf{X}) = \mathbf{K}$

$\mathbf{K}_* = \kappa(\mathbf{X}, \mathbf{X}_*)$

$\mathbf{K}_{**} = \kappa(\mathbf{X}_*, \mathbf{X}_*)$

사후 예측분포 $p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y}) = \int p(\mathbf{f}_*|\mathbf{f}, \mathbf{X}_*) p(\mathbf{f}|\mathbf{X}, \mathbf{y}) d\mathbf{f}.$

RBF 커널 및 매개변수 $\kappa(\mathbf{x}_i, \mathbf{x}_j) = \sigma_{\mathbf{f}}^2 \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^{\mathrm{T}} \mathbf{M} (\mathbf{x}_i - \mathbf{x}_j)\right] + \sigma_y^2 \delta_{ij}$

Vertical variation parameter

Length scale parameter

Noise parameter

새로운 $X_*$에 대한 예측 $f_*$

$p(\mathbf{f}_*|\mathbf{X}_*, \mathbf{X}, \mathbf{y}) = N(\mathbf{f}_*|\boldsymbol{\mu}_*, \boldsymbol{\Sigma}_*)$ $\boldsymbol{\mu}_* = \mathbf{K}_*^{\mathrm{T}} \mathbf{K}_y^{-1} \mathbf{y}$

$\boldsymbol{\Sigma}_* = \mathbf{K}_{**} - \mathbf{K}_*^{\mathrm{T}} \mathbf{K}_y^{-1} \mathbf{K}_*$

커널의 매개변수는 LML (로그 주변 우도함수) 최대화하도록 결정

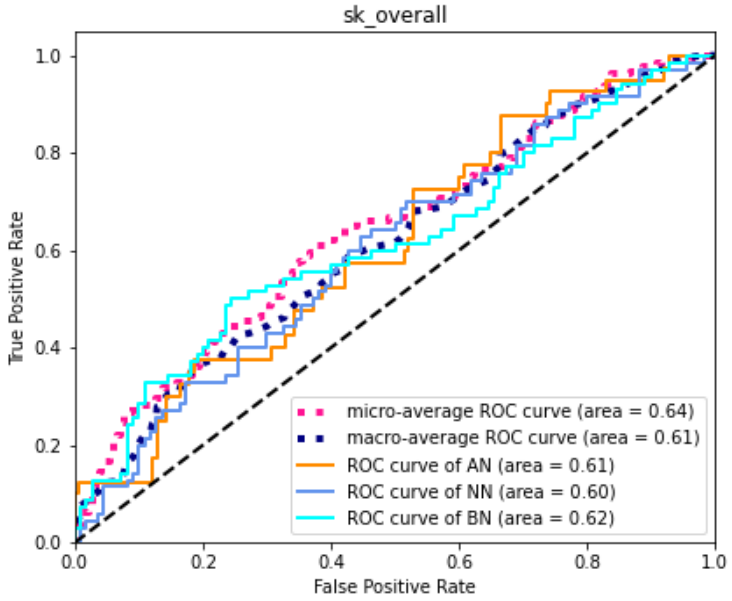
$$\log p(\mathbf{y}|\mathbf{X}) = \log \mathcal{N}(\mathbf{y}|\mathbf{0}, \mathbf{K}_y) = -\frac{1}{2}\mathbf{y}^T \mathbf{K}_y^{-1} \mathbf{y} - \frac{1}{2}\log|\mathbf{K}_y| - \frac{N}{2}\log(2\pi)$$

# II. 기계학습을 이용한 장기예측

## 모델 평가를 위한 기준을 정하자

### 확률예측을 위한 모델 평가

- AUC (Area Under the Receiver Operating Characteristic Curve)
  - False positive rate을 x축에, True positive rate을 y축에 둔 ROC 곡선 아래 면적
  - Perfect Forecast: 1.0
  - Random Forecast: 0.5
- RPSS (Rank Probability Skill Score)




$$\mathrm{RPS} = \sum_k (Y_k - O_k)^2,$$

$$\mathrm{RPS_{Ref}} = \sum_k (P_k - O_k)^2,$$

$$\mathrm{RPSS} = 1 - \frac{\mathrm{RPS}}{\mathrm{RPS_{Ref}}},$$

where $k$ refers to categories of AN, NN, and BN; $Y_k$, $O_k$, and $P_k$ denote cumulative forecast probability, cumulative observation probability, and cumulative climatological probability for category $k$, respectively. A perfect set of forecasts would be scored as 1.0, and a set of random forecasts would be scored as zero.

# II. 기계학습을 이용한 장기예측

## Toy Example 확률예측 결과

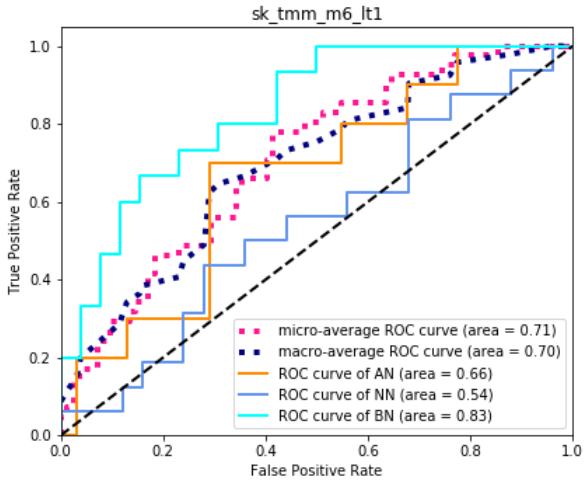
### 가우시안 프로세스

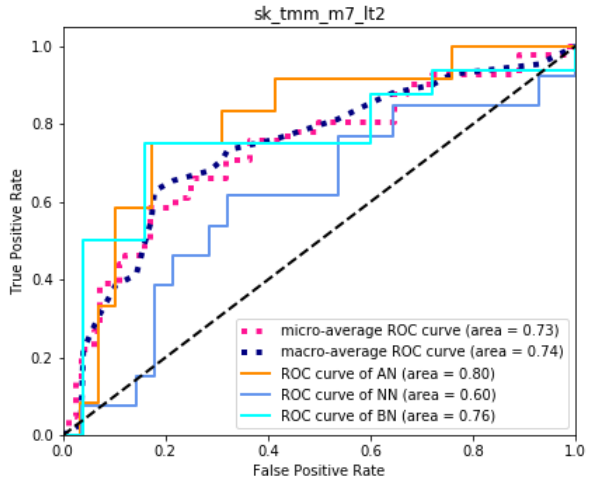
| Month | HSS | Total Accuracy | r | MAE | HR (AN) | HR (NN) | HR (BN) |
|---|---|---|---|---|---|---|---|
| Jun | 0.25 | 0.50 | 0.50 | 0.45 | 0.48 | 0.44 | 0.59 |
| Jul | 0.33 | 0.56 | 0.42 | 0.82 | 0.49 | 0.49 | 0.68 |
| Aug | 0.28 | 0.52 | 0.46 | 0.84 | 0.57 | 0.31 | 0.67 |

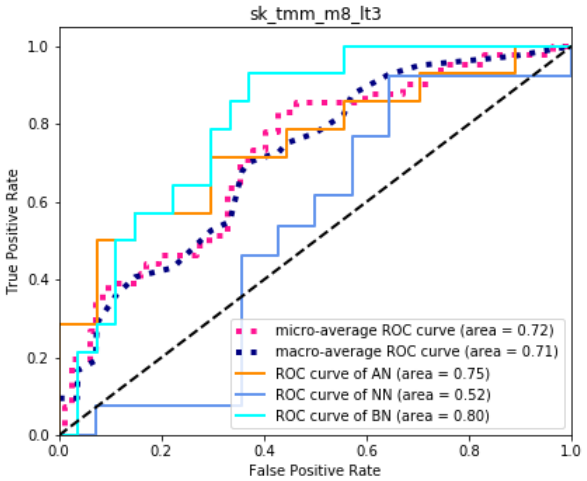
| Month | AUC | RPSS |
|---|---|---|
| Jun | 0.70 | 0.10 |
| Jul | 0.74 | 0.13 |
| Aug | 0.71 | 0.13 |

8월 평균기온 예측결과 (확률예측을 삼분위에 적용)

| YEAR | TRUE | PRED |
|---|---|---|
| 2016 | AN | AN |
| 2017 | NN | NN |
| 2018 | AN | NN |
| 2019 | AN | BN |
| 2020 | AN | AN |




# 오늘의 학습

## I. 기계학습이란

- 인공지능이란
- 기계학습이란
- 기계학습의 종류

## II. 기계학습을 이용한 장기예측

- 장기예측을 위한 준비
- 다양한 기계학습을 이용한 기온예측 – Toy Examples


APCC
APEC 기후센터
감사합니다.