기계 학습 (Machine Learning)

본 문서는 기계 학습 과정을 모델(model)이라는 학생을 가르치는 것에 비유합니다.

교사가 학생을 가르치는 예시를 먼저 들겠습니다. 교사는 문제 은행의 문제들 중 일부는 연습 문제로 사용하고 나머지는 시험 문제로 출제합니다. 학생에게 연습 문제와 그에 대한 답지를 주고 학습하도록 합니다. 학생은 문제와 답지를 보며 공부합니다. 어느 정도 학습이 완료되면 시험을 봅니다. 시험을 볼 때는 문제만 주어야 하며, 연습 문제와 같은 문제가 출제되어서는 안 됩니다. 학생은 문제를 풀고 답안지를 제출합니다. 교사는 학생이 제출한 답안지와 실제 답을 비교하여 학생의 문제 풀이 능력을 평가합니다. 간단히 나타내면 아래와 같습니다:

1. 문제 은행의 문제들을 연습 문제와 시험 문제로 나눕니다.
2. 학생에게 연습 문제와 그에 대한 답을 주고 학습시킵니다.
3. 학생에게 시험 문제를 풀립니다.
4. 학생이 제출한 답안지와 실제 답을 비교합니다.

기계 학습도 교사가 학생을 가르치는 것과 다르지 않습니다. 우선 데이터를 데이터프레임으로 불러들입니다. 이것이 우리의 문제 은행입니다. 데이터프레임의 특성 중 알고 싶은 특성에 대한 열을 변수 `y` 에 저장합니다. 학습에 사용될 특성만을 가지는 데이터프레임을 변수 `X` 에 저장합니다. `y` 는 답, `X` 는 문제인 것입니다. 그 후에는 데이터를 훈련 데이터와 검증 데이터로 나눕니다. `X`, `y` 가 총 네 개의 데이터셋으로 나뉘는데, 훈련 `X`, 검증 `X`, 훈련 `y`, 검증 `y` 로 나뉩니다. 이제 모델을 데이터에 피팅(fitting)합니다. 피팅은 훈련 `X` 와 훈련 `y` 를 이용합니다. 그 후에는 검증 `X` 에 대한 예측 값을 모델로부터 받아 검증 `y` 와 비교하고 평가합니다. 간단히 나타내면 아래와 같습니다:

1. 데이터셋을 불러옵니다.
2. 데이터셋을 문제와 답으로 나눕니다.
3. 문제와 답을 훈련용과 검증용으로 나눕니다.
4. 모델을 정의합니다.
5. 훈련용 문제와 답을 이용해 모델을 피팅합니다.
6. 검증용 문제에 대한 모델의 예측 값을 얻습니다.
7. 검증용 답과 예측 값을 비교합니다.
8. 평가 지표를 이용해 모델을 평가합니다.

```python
import pandas as pd
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestRegressor
from sklearn.metrics import mean_absolute_error

# Step 1
path = '../data/melb_data.csv'
df = pd.read_csv(path)

# Step 2
features = ['Rooms', 'Bathroom', 'Landsize',
            'BuildingArea', 'YearBuilt',
            'Lattitude', 'Longtitude']
X = df[features]
y = df['Price']

# Step 3
tr_X, val_X, tr_y, val_y = train_test_split(X, y)

# Step 4
model = RandomForestRegressor(random_state=0)

# Step 5
model.fit(tr_X, tr_y)

# Step 6
preds = model.predict(val_X)

# Step 7, 8
error = mean_absolute_error(val_y, preds)
```