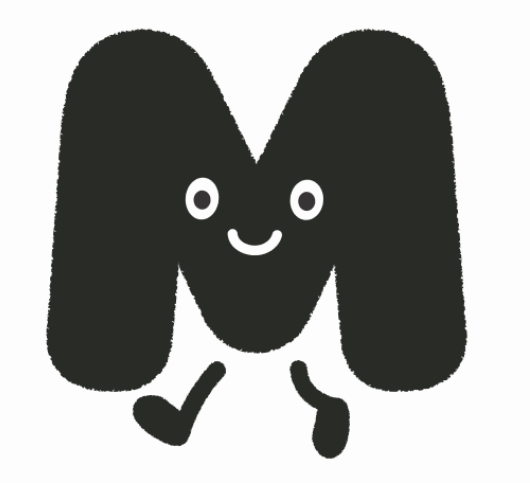
MOMENTUM

CRISP-DM :: Modeling(회귀)

팀별 활동을 통해 실습과 이론을 동시에~

목차

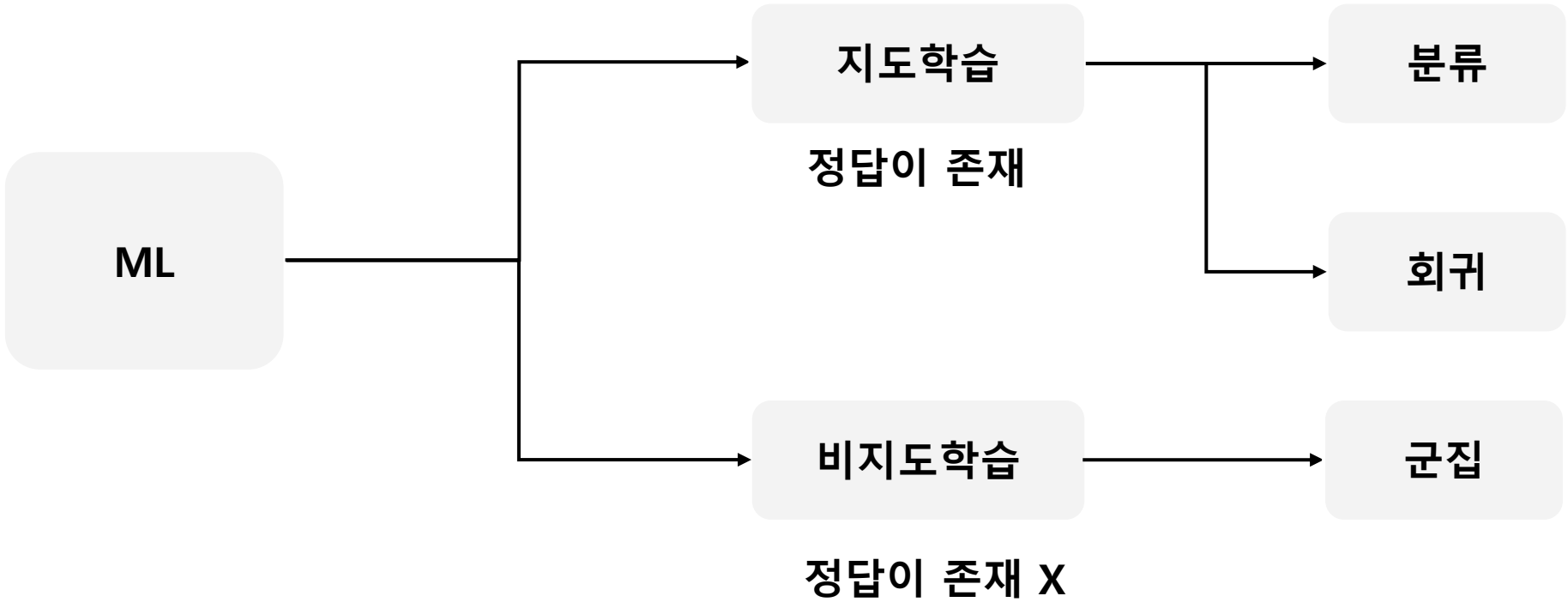
- 회귀가 뭘까...



머신러닝이 뭔가요?

머신러닝 모델을 이용하여 데이터의 **패턴을 학습(Fit)**하고,
미래에 대한 **판단이나 예측(Predict)**을 하는 것

머신러닝의 종류



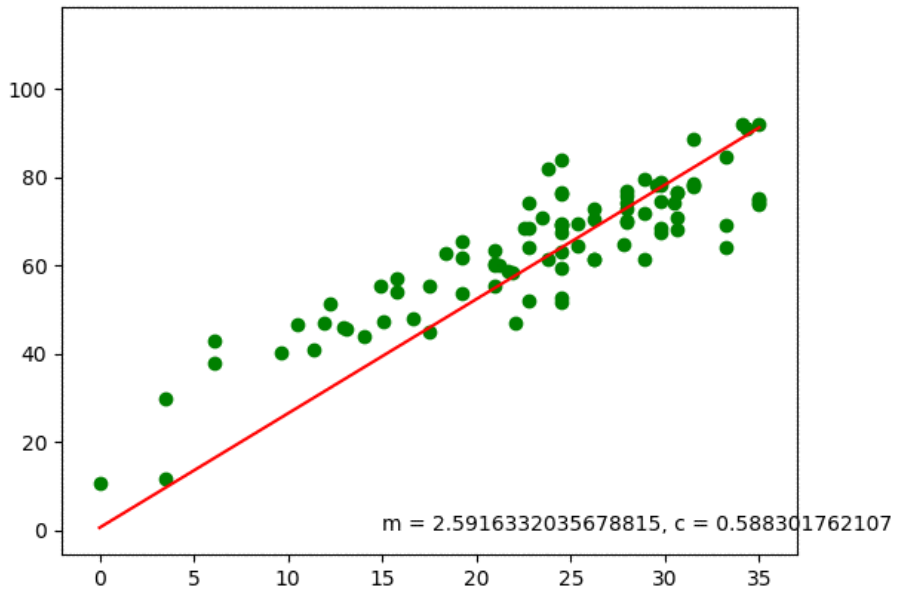
지도학습 – 회귀와 분류

회귀		분류
숫자(연속형 값)을 예측	**정의**	카테고리(범주형 값)을 예측해 분류
가격, 시각, 확률 등 수치 등을 예측	**목표**	Yes/No, A or B, 고양이 강아지 등을 분류
가격 예측, 재구매 확률 예측	**예시**	스팸 메일 분류, 질병 유무 진단, 고객 이탈 예측
선형회귀, KNN, 랜덤포레스트 라쏘, 릿지,	**주요 알고리즘**	KNN, 로지스틱 회귀, 랜덤포레스트

회귀 모델

선형 회귀 (Linear Regression) 알고리즘

- 데이터를 가장 잘 설명하는 회귀 직선을 찾는다.
- 독립 변수(X)와 종속 변수(y) 간의 선형 관계를 기반으로 예측
- 독립 변수 개수에 따라 단순 회귀(1개) / 다중 회귀(2개 이상)으로 구분



장점

- 구현하기 쉬움
- 독립 변수와 종속 변수 간의 관계 직관적으로 파악 가능
- → 기업에서 많이 쓰이는 회귀 모델

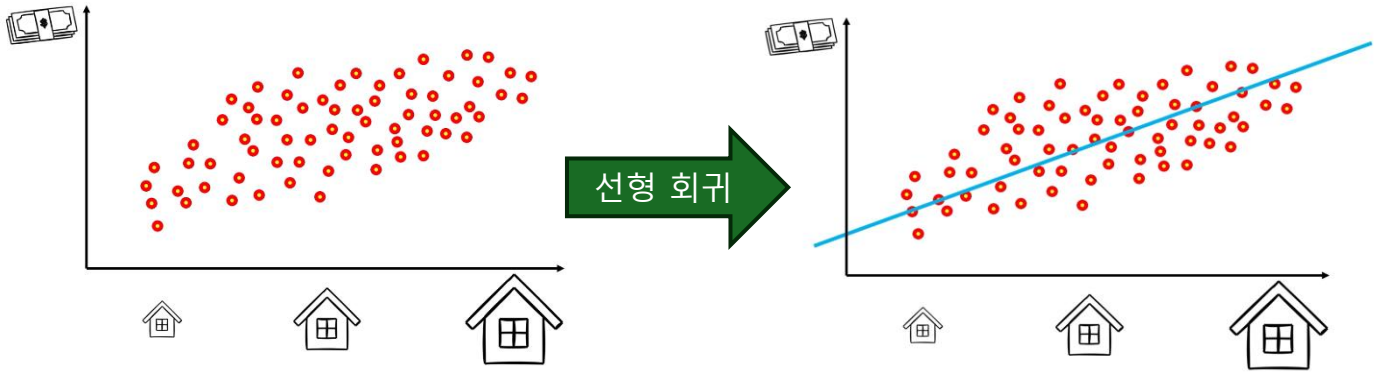
단점

- 독립 변수와 종속 변수 사이에 선형 관계가 있다는 걸 전제해야 함
- 다중공선성 문제가 발생할 수 있음 (독립 변수들 간에 강한 상관관계가 나타나는 문제)

회귀 모델

단순 선형 회귀 (Simple Linear Regression) 알고리즘

- 하나의 독립 변수와 종속 변수 간의 관계를 찾는 회귀 방법
- Ex) 공부 시간 → 시험점수, 광고비 → 매출액 등
- $\hat{y} = \boldsymbol{W_1 X} + \boldsymbol{b}$

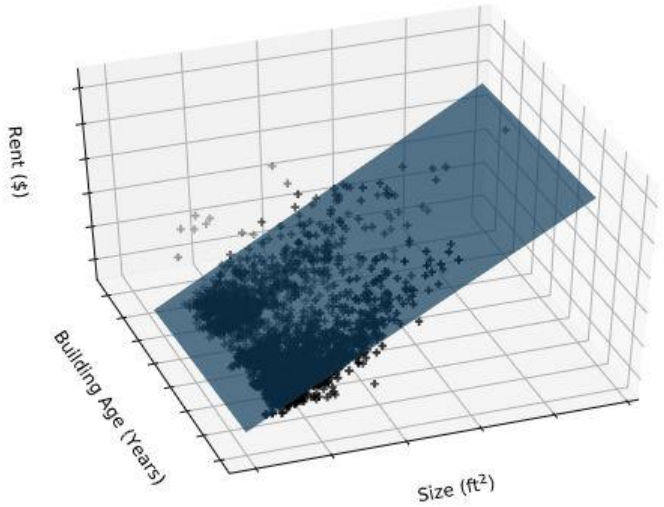


- 독립 변수: 주택 크기
- 종속 변수: 주택 가격

회귀 모델

다중 선형 회귀 (Multiple Linear Regression) 알고리즘

- 여러 개의 독립 변수와 종속 변수 간의 관계를 찾는 회귀 방법
- 현실 문제에 더 적합함 (여러 요인이 결과에 영향을 미치기 때문)
- $\hat{y} = \boldsymbol{W_1X_1} + \boldsymbol{W_2X_2} + \ldots + \boldsymbol{WnXn} + \boldsymbol{b}$

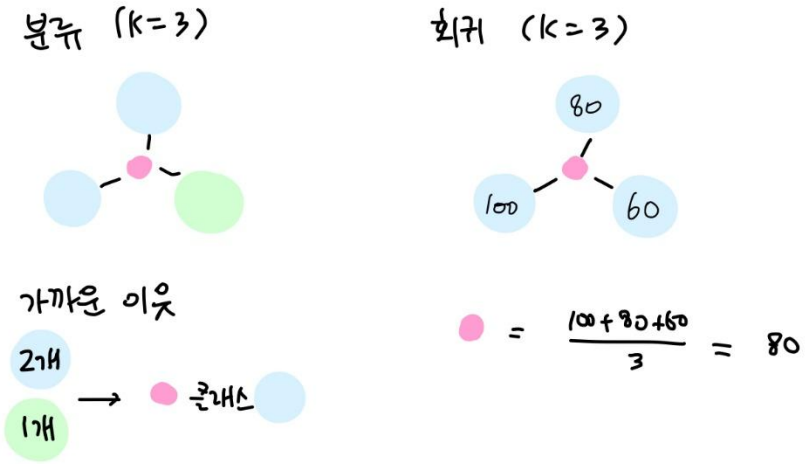


- 독립 변수: 집의 크기, 건물 나이
- 종속 변수: 임대료

회귀 모델

KNN (K-Nearest Neighbors) 회귀 알고리즘

- K개의 가장 가까운 주변 데이터의 종류를 보고 예측 하겠다.



- KNN 분류와 회귀의 구조는 동일하지만 분류는 범주형(클래스)을 예측하는 것이고 회귀는 수치 값을 예측하는 것

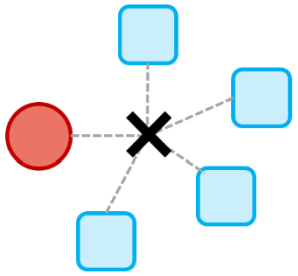
- KNN 회귀는 K개의 이웃의 종속 변수 값을 평균하여 예측값을 산출

회귀 모델

KNN (K-Nearest Neighbors) 회귀 알고리즘

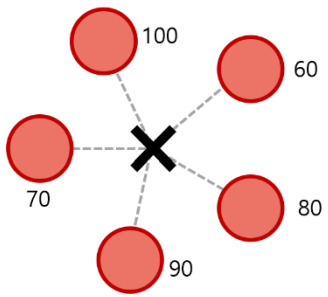
- K개의 가장 가까운 주변 데이터의 종류를 보고 예측 하겠다.

<Classification>



∴ ✕의 클래스는 □

<Regression>



∴ ✕의 예측 타깃 값은 $\frac{60+70+80+90+100}{50}=80$

장점

- 알고리즘이 간단하고 직관적
- 비선형적 데이터에도 잘 작동함

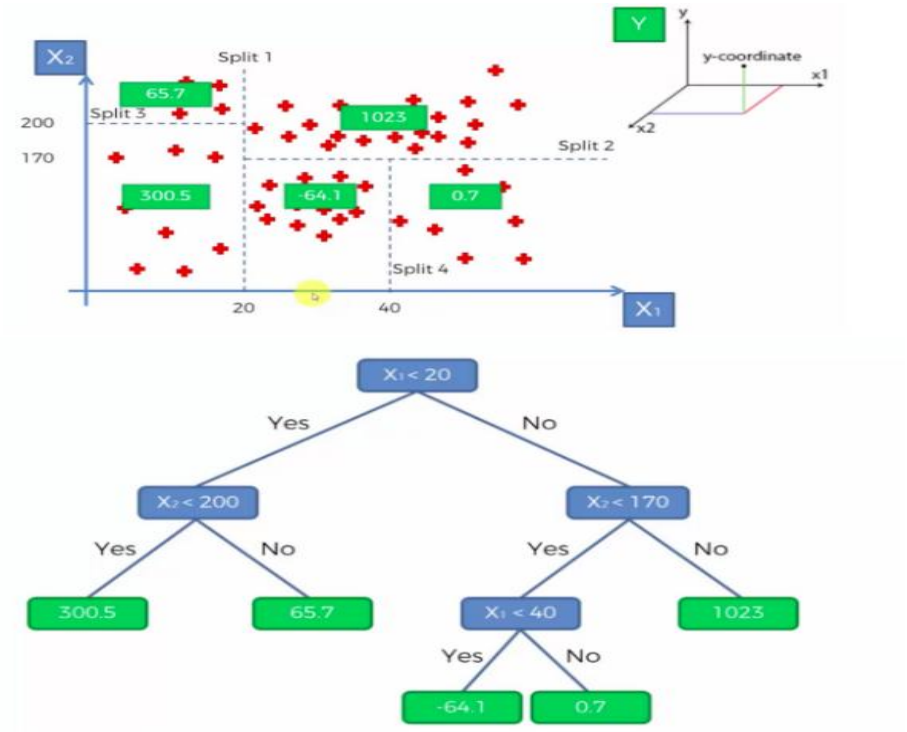
단점

- 대규모 데이터면 계산 비용이 높음
- K값이 적절하지 않으면 과대/과소적합이 발생
- 거리기반이므로 전처리가 필수

회귀 모델

랜덤 포레스트 회귀 알고리즘 (앙상블 중 배깅)

- 의사결정 나무(Decision Tree)를 여러 개 만들어, 각 나무의 결과를 합치는 방식

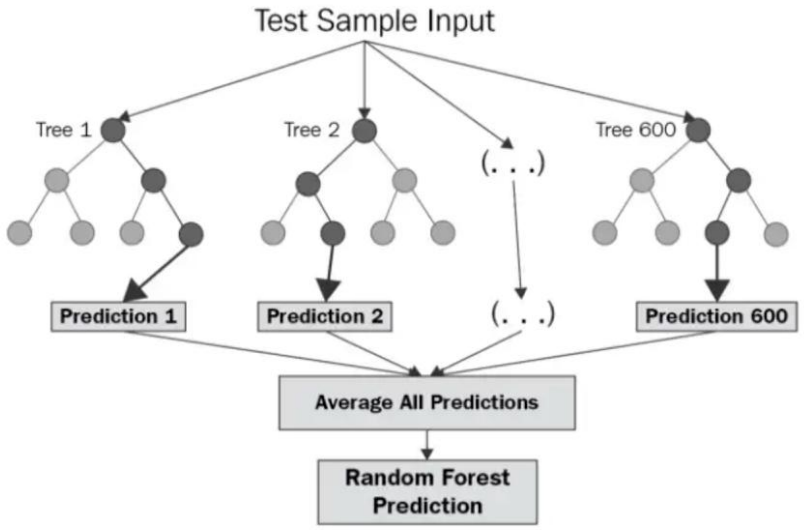


- 속성 값 테스트(질문)을 통해 데이터를 하위 집합으로 분할하여 값을 예측

회귀 모델

랜덤 포레스트 회귀 알고리즘 (앙상블 중 배깅)

- 의사결정 나무(Decision Tree)를 여러 개 만들어, 각 나무의 결과를 합치는 방식



-표본 추출을 통해 의사결정 나무를 여러개 만들어, 해당 결과를 평균으로 합치는 방식

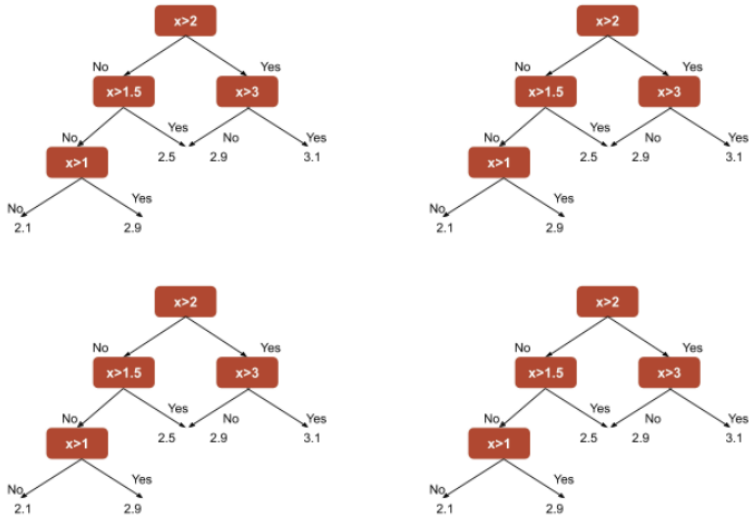
- 4개의 결정트리 회귀 모델이 각각
12.3, 13.0, 11.7, 12.5 라고 예측

-> 최종 예측값 = 12.375(4개의 평균)

회귀 모델

랜덤 포레스트 알고리즘 (앙상블 중 배깅)

- 의사결정 나무(Decision Tree)를 여러 개 만들어, 각 나무의 결과를 합치는 방식



장점

- 높은 예측 성능
- 복잡한 패턴도 모델링 가능
- 전처리 필요 없음
- 과적합 방지

단점

- 해석이 어려움
- 훈련 시간이 느림

오늘의 실습 데이터 Insurance

변수명	데이터 타입	설명
age	int	나이 (18세 이상)
sex	object	성별 (male, female)
bmi	float	체질량지수 (Body Mass Index)의료상 과체중 여부를 판단하는 지표 (kg/m²)
children	int	자녀 수
smoker	object	흡연 여부 (yes, no)
region	object	거주 지역 (southwest, southeast, northwest, northeast)
charges	**float**	**보험료 (target 변수) – 예측 대상**

ML 프로세스

요구사항 정의 | 데이터 수집 및 전처리 | 모델 정의하기 | 평가하기

무엇을 예측할 것인가?

무엇으로 예측할 것인가?

종속 변수(예측 할 대상) : 고객의 보험료 청구 비용예측

독립 변수(갖고 있는 정보) : 나이, BMI, 흡연 여부 등

ML 프로세스

요구사항 정의 | 데이터 수집 및 전처리 | 모델 정의하기 | 평가하기

어떤 데이터를 수집할 것인가?

데이터를 종류별로 어떻게 전처리할 것인가?

age	sex	bmi	children	smoker	region	charges
19	female	27.9	0	yes	southwest	16884.92
18	male	33.77	1	no	southeast	1725.552

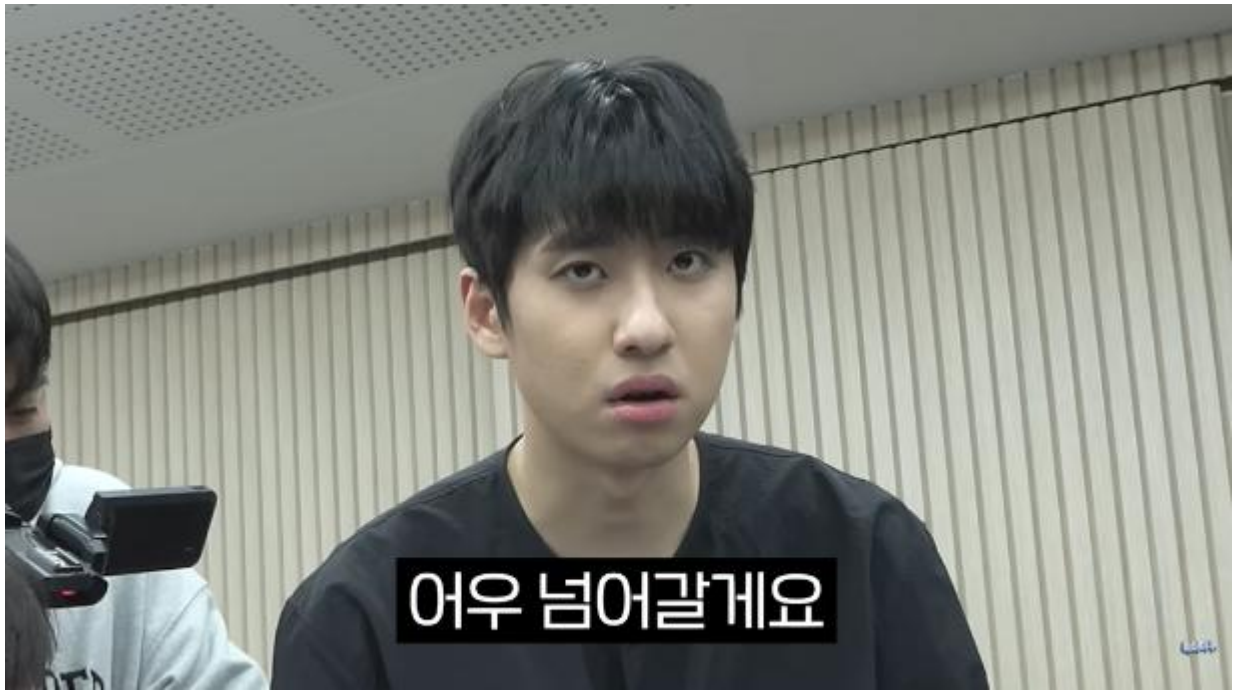
age	sex	bmi	children	smoker	region	charges
19	0	27.9	0	1	3	16884.92
18	1	33.77	1	0	2	1725.552

컴퓨터가 이해할 수 있도록 데이터 처리하기

스케일링, 인코딩, 결측치 처리

데이터 수집 및 전처리

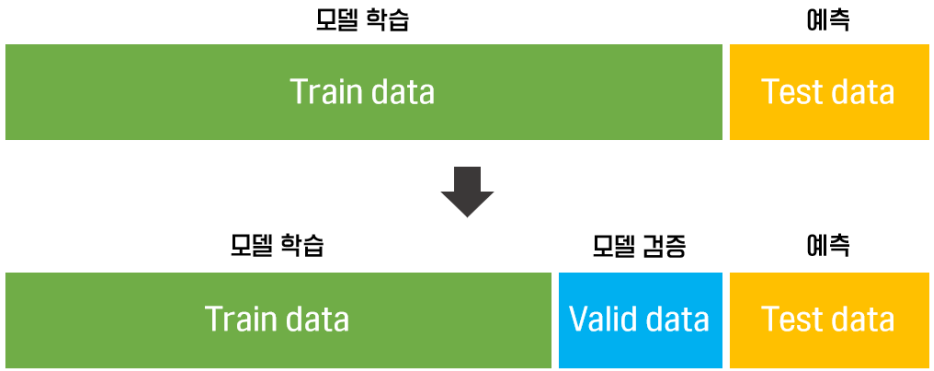
데이터 전처리(스케일링, 인코딩, 결측치 처리)



데이터 수집 및 전처리

데이터셋 분할(train test split)

- 예측에 필요한 데이터를 모델에 적용시키기 위해 모델을 train 시켜야함
- 모델이 학습을 위해 사용되는 데이터가 train data
- 선정한 모델이 잘 작동하는지 평가하는 데이터가 test data
- Test로 넘어가기 전, 모델을 fine tuning하는 데이터가 valid data



교재, 문제집 (train) → 모의고사 (vaild) → 실제시험 (test)

ML 프로세스

요구사항 정의 | 데이터 수집 및 전처리 | 모델 정의하기 | 평가하기

어떤 모델을 사용할 것인가?

지도학습인가?

비지도학습인가?

분류인가?

회귀인가?

age	sex	bmi	children	smoker	region	charges
19	0	27.9	0	1	3	16884.92
18	1	33.77	1	0	2	1725.552

X : 독립변수 Y : 종속변수

X와 Y를 모두 학습하니 지도 학습이며, 회귀 모델을 사용해야함

X에 따른 Y(보험료)를 모델에 학습 (훈련데이터) **Fit한다**

age	sex	bmi	children	smoker	region	charges
28	1	33	3	0	2	?
33	1	22.705	0	0	1	?

X : 독립변수 Y : 종속변수

훈련시킨 모델을 통해 다른 고객의 보험료를 예측한다.

같은 독립 변수를 갖고 있는 고객들의 데이터를 X로 넣고 Y를 **Predict한다**

ML 프로세스

요구사항 정의 | 데이터 수집 및 전처리 | 모델 정의하기 | **평가하기**

그래서 잘 예측이 되었는가?

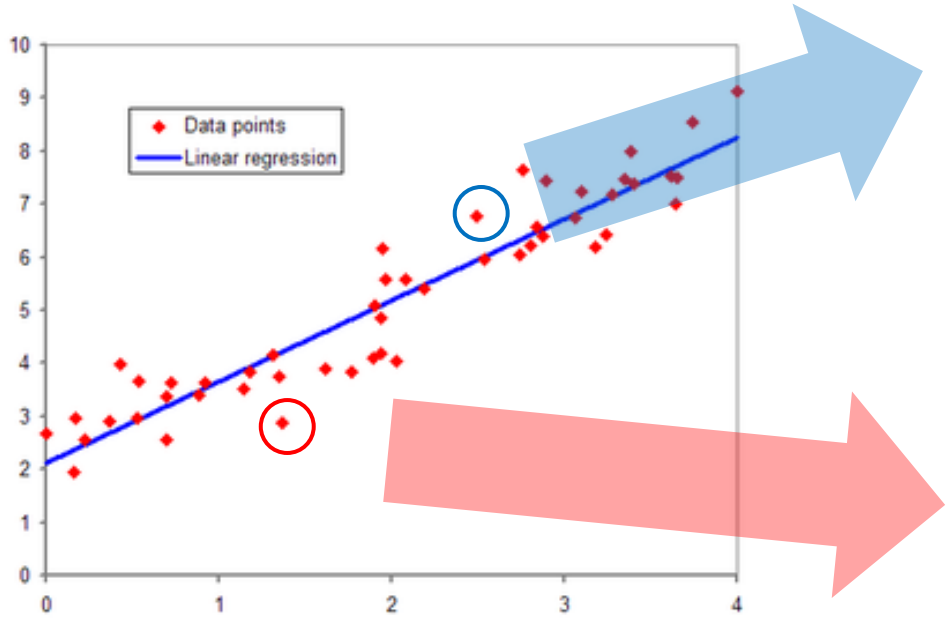
예측	실제
1276번 : 10892.14$	1276번 : 10959.33$
1337번 : 2066.54$	1337번 : 2007.95$

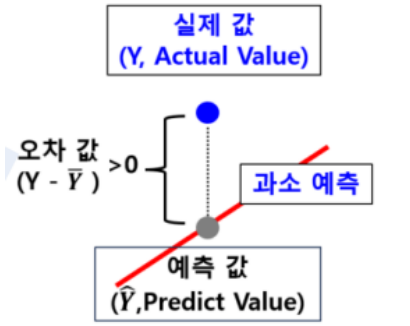
실제값과 예측값을 비교해
모델의 성능을 평가한다.

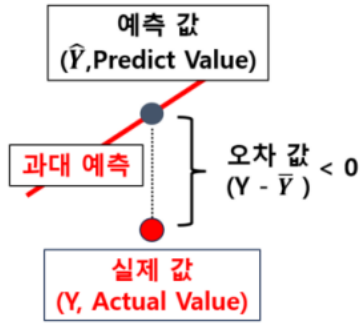
평가하기

오차

- 오차: 모델이 예측한 값과 실제 값의 차이
- 오차 $= y_i - \hat{y}_i$

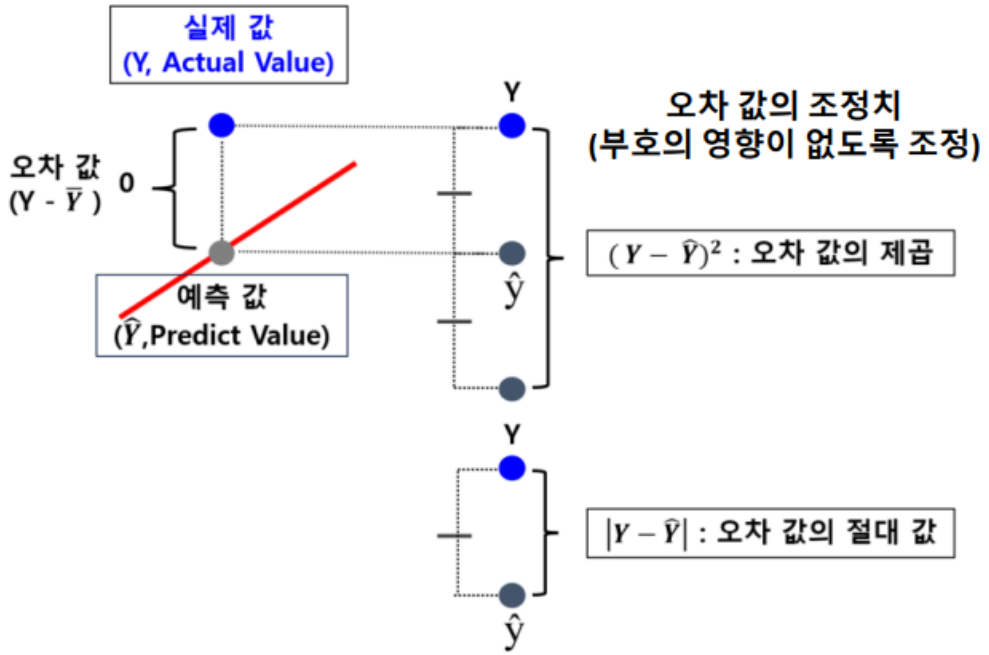






평가하기

오차 값의 조정치



모든 실제 값과 예측 값의 "오차의 제곱값" 평균으로 모델의 성능(정확도) 계산 : MSE

$$\frac{1}{N}\sum_{i=1}^{N}(y_i-\hat{y}_i)^2$$

모든 실제 값과 예측 값의 "오차의 절대값" 평균으로 모델의 성능(정확도) 계산 : MAE

$$\frac{1}{N}\sum_{i=1}^{N}|y_i-\hat{y}_i|$$

평가하기

MSE (Mean Squared Error : 평균 제곱 오차)

- 실제값과 예측값의 차이를 제곱해 평균한 것

- 제곱했기 때문에 MSE는 항상 양수이며, 0에 가까울수록 모델의 예측 정확도가 높음

$$\mathrm{MSE} = \frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{y}_i)^2$$

평가하기

RMSE (Root Mean Squared Error : 평균 제곱근 오차)

- MSE에 루트를 씌운 값
- 원본 데이터와 같은 단위를 가져, 직관적인 해석 가능

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

평가하기

MAE (Mean Absolute Error : 평균 절대값 오차)

- 실제값과 예측값의 차이를 절대값을 씌워 평균한 것
- 오차의 방향성을 제거하고 크기만을 고려함
- 이상치에 덜 민감하다는 장점

$$MAE = \frac{1}{N}\sum_{i=1}^{N} | y_i - \hat{y}_i |$$

평가하기

R^2score (결정계수)

- 회귀모델에서 독립 변수가 종속 변수를 얼마나 잘 설명해주는 지 보여주는 지표

- 1에 가까울수록 모델의 설명력이 높다는 것

$$R^2score = \frac{SSE}{SST} = 1 - \frac{SSR}{SST}$$

$$SST = SSE + SSR$$

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- SST(전체제곱합) = ∑(실제값과 평균값의 차이)²

$$SSE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- SSE(오차제곱합)= ∑(예측값과 평균값의 차이)²

$$SSR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- SSR(회귀제곱합) = ∑(실제값과 예측값의 차이)²