데이터 사이언스

머신러닝: 회귀분석

박재완 교수

숭실대학교
Soongsil University

■ 목표설정

- 목표: 보스턴 주택 가격 데이터에 머신러닝 기반의 회귀 분석을 수행
  주택 가격에 영향을 미치는 변수를 확인하고 그 값에 따른 주택 가격을 예측

■ 핵심 개념 이해

▪ 머신러닝

- 1959년 아서 사무엘: '컴퓨터에 명시적인 프로그램 없이 스스로 학습할 수 있는 능력을 부여하는 연구 분야'로 정의
- 인간이 지식과 경험을 학습하는 방법을 적용하여 컴퓨터에 입력된 데이터에서 스스로 패턴을 찾아 학습하여 새로운 지식을 만들고 예측하는 통찰을 제공하는 AI의 한 분야

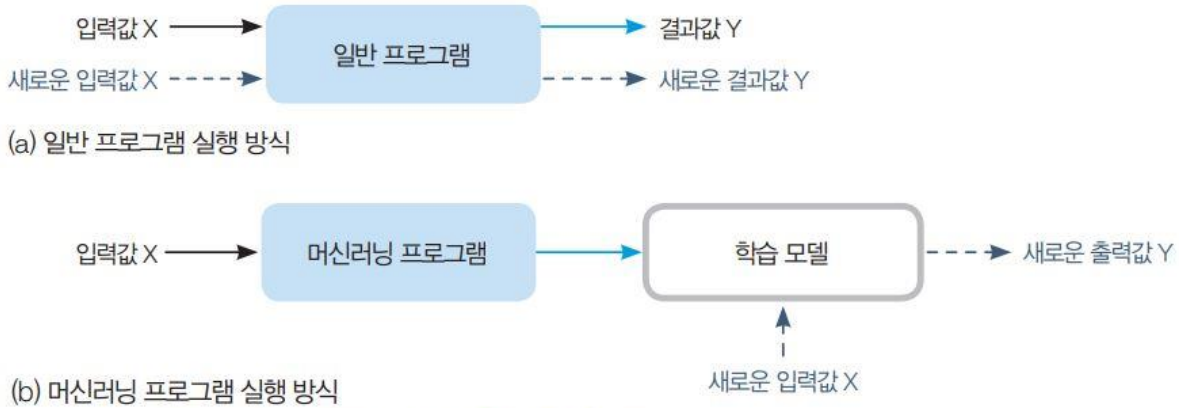


그림 10-1 일반 프로그램과 머신러닝 프로그램 실행 방식 비교

■ 핵심 개념 이해

▪ 머신러닝 프로세스

데이터 수집 → 데이터 전처리 및 훈련/테스트 데이터 분할 → 모델 구축 및 학습 → 모델 평가 → 예측

▪ 지도 학습

- 학습을 하기 위한 훈련 데이터에 입력과 출력을 같이 제공
- 문제(입력)에 대한 답(출력, 결과값)을 아는 상태에서 학습하는 방식
- 입력: 예측 변수, 속성, 특징
- 출력: 반응 변수, 목표 변수, 클래스, 레이블

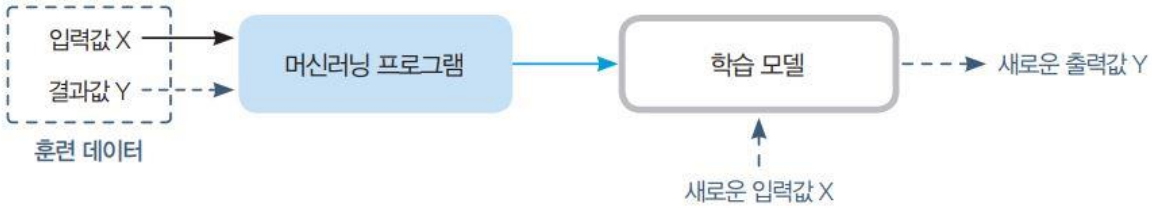
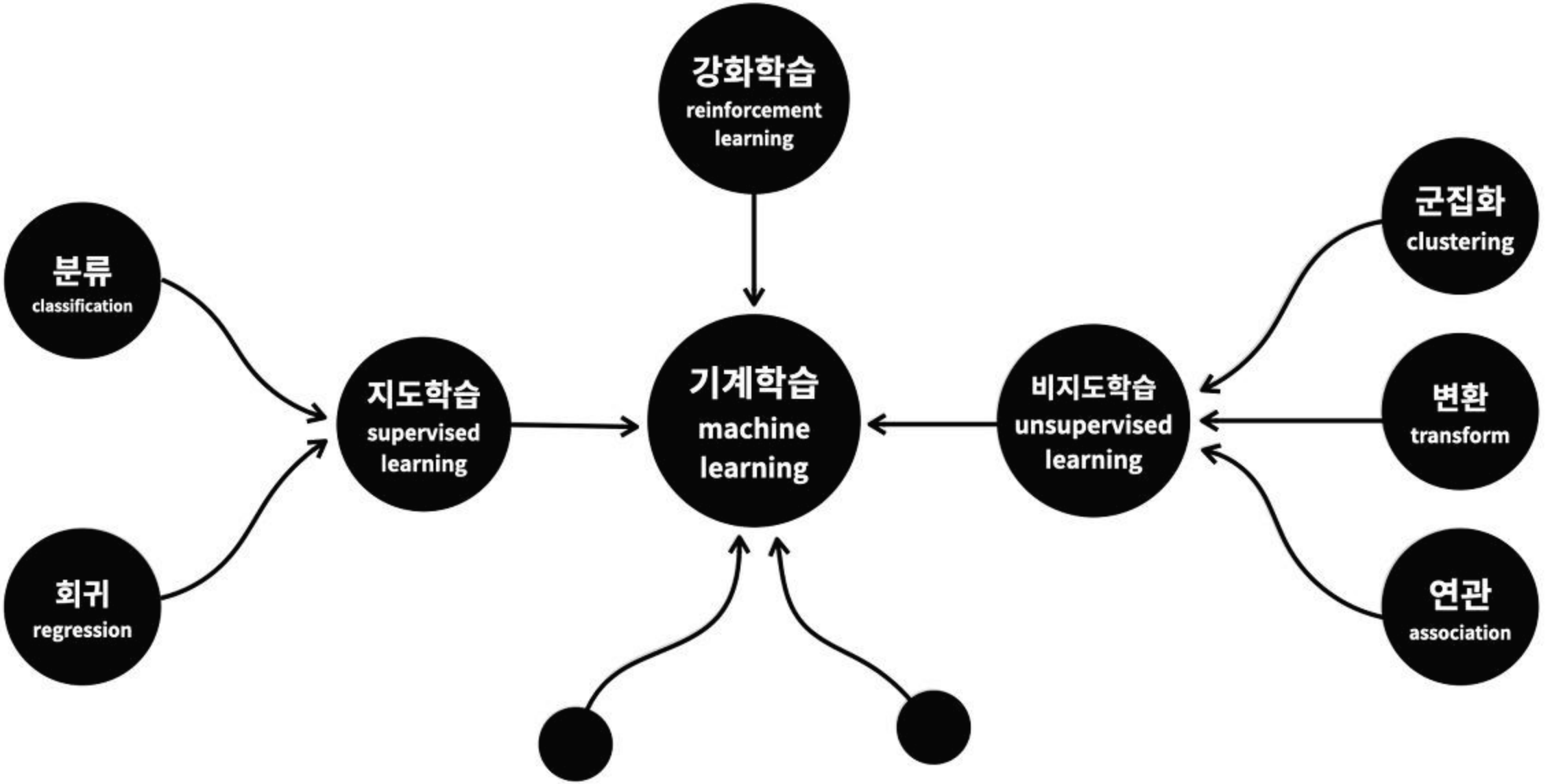


그림 10-2 머신러닝의 지도 학습 방식

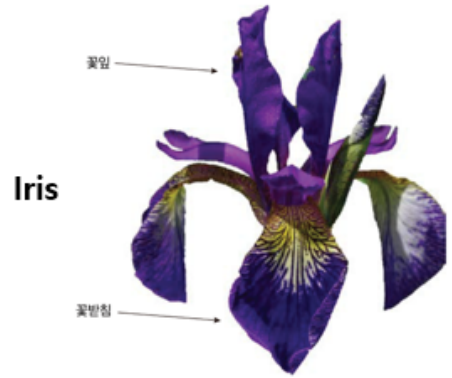
▪ 사이킷런

- 파이썬으로 머신러닝을 수행하기 위한 쉽고 효율적인 개발 라이브러리를 제공
- 보스톤 주택 가격 데이터, 붓꽃 데이터 등과 같은 머신러닝 분석용 데이터셋 을 제공
- 전체 n개의 컬럼 중 앞에서 (n-1)개의 컬럼은 독립 변수 X를 의미
- 마지막 컬럼 은 종속 변수 Y이며, 데이터셋 객체의 target 배열로 관리


강화학습
reinforcement learning
분류
classification
회귀
regression
지도학습
supervised learning
기계학습
machine learning
비지도학습
unsupervised learning
군집화
clustering
변환
transform
연관
association

이미지 출처: 생활코딩

- **독립변수(Independent Variable)**: 다른 변수에서 영향을 받지 않는 독립적인 변수
- **종속변수(Dependent Variable)**: 종속적인 의존적인 변수, 독립변수에 영향을 받아서 변화하는 변수



연속형 자료: 공변량(Covariance)

ID (Order)	나이 (Age)	성별 (Gender)	경력 (Career)	연봉 (Salary)
1	33	남	5	4000
2	31	여	3	5000
3	39	여	1	3800
4	32	남	7	4000
5	24	여	3.5	6000
6	28	남	4	4500
⋮	⋮	⋮	⋮	⋮
1000	35	여	7	7000

범주형 자료: 요인(Factor)

Target 클래스

회귀: 숫자(Numeric)

종속변수

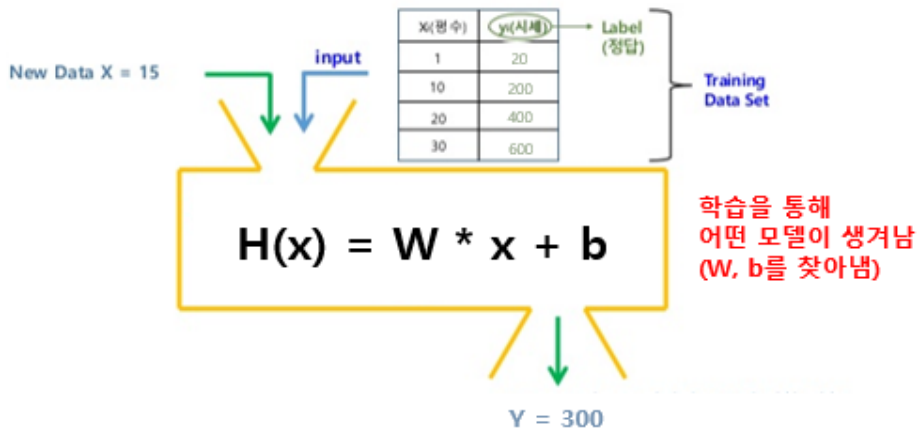
ID (Order)	꽃받침 길이 (Sepal length)	꽃받침 너비 (Sepal width)	꽃잎 길이 (Petal length)	꽃잎 너비 (Petal width)	종의 이름 (Species)
1	5.1	3.5	1.4	0.2	Setosa
2	4.9	3.0	1.4	0.2	Setosa
3	4.7	3.2	1.3	0.2	Setosa
⋮	⋮	⋮	⋮	⋮	⋮
51	6.4	3,5	4.5	1.2	Versicolor
⋮	⋮	⋮	⋮	⋮	⋮
150	5.9	3.0	5.0	1.8	Virginica

Target 클래스

분류: 범주형(Categorical)

독립변수

모델



학습을 통해
어떤 모델이 생겨남
(W, b를 찾아냄)

ID (Order)	나이 (Age)	성별 (Gender)	경력 (Career)	연봉 (Salary)
1	33	남	5	4000
2	31	여	3	5000
3	39	여	1	3800
4	32	남	7	4000
5	24	여	3.5	6000
6	28	남	4	4500
. . .	. . .	. . .	. . .	. . .
1000	35	여	7	7000

Target

숫자(Numeric)

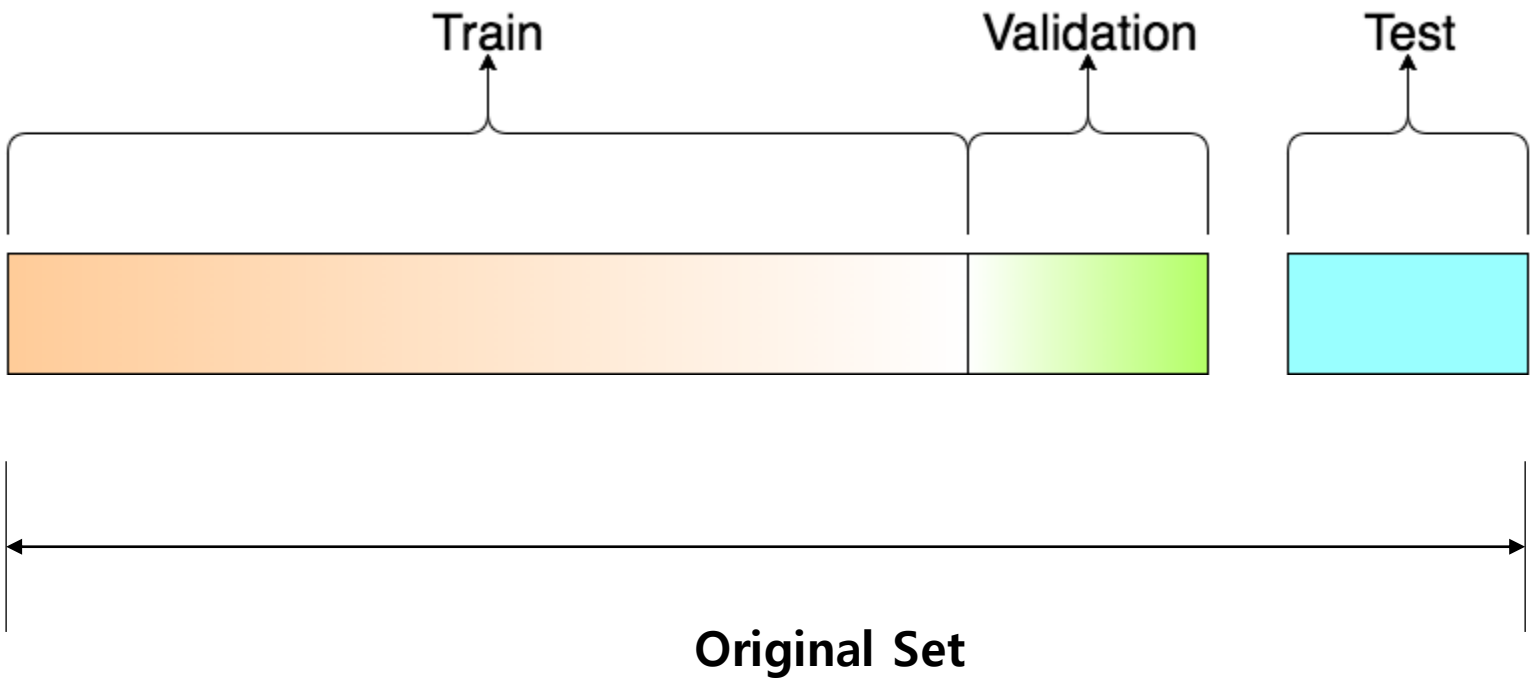
$$y = ax + b$$

$$f(x) = ax + b$$

$$H(x) = wx + b$$

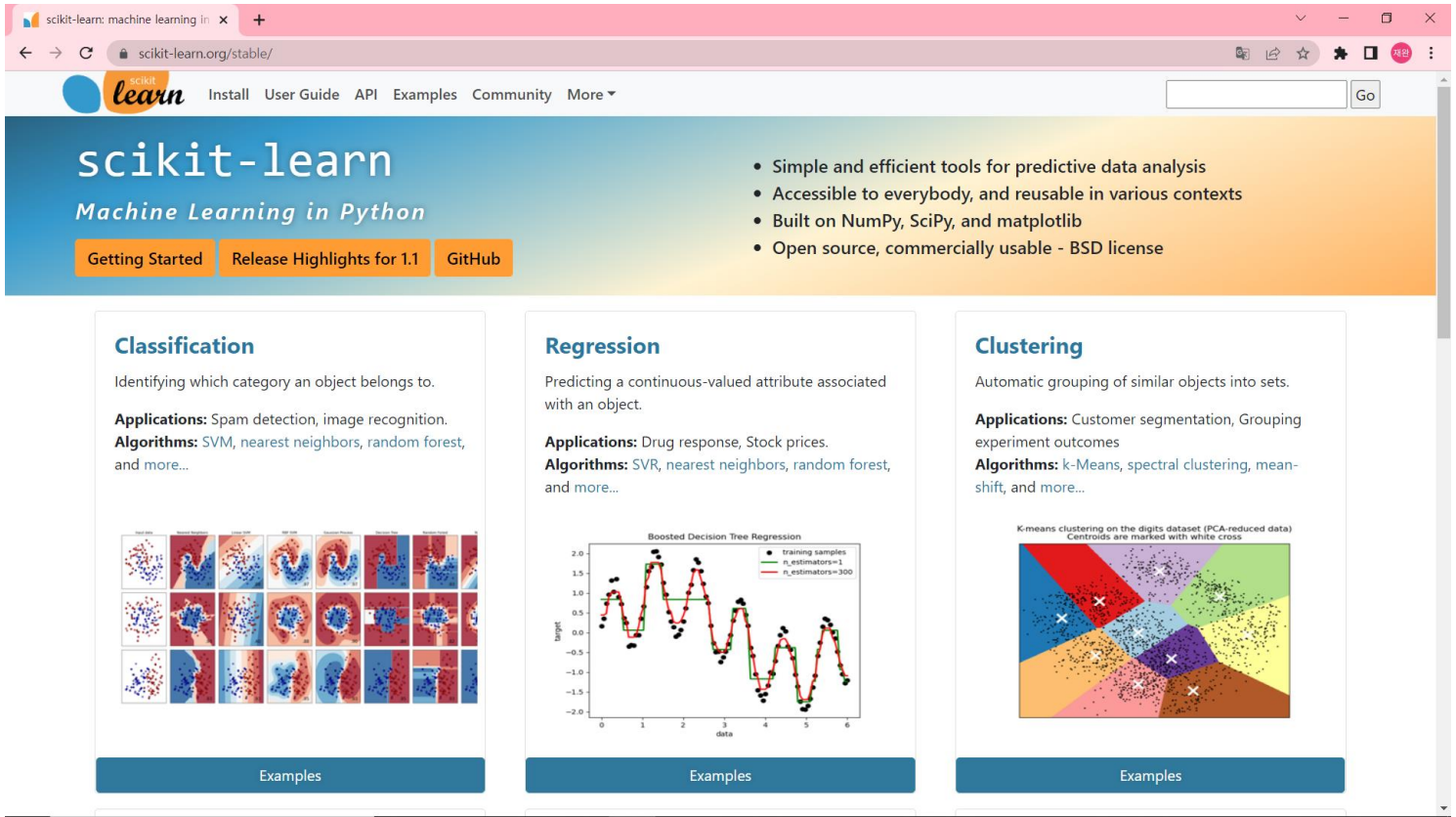
Hypothesis Weight Bias

$$y = w_1x_1 + w_2x_2 + w_3x_3 + \cdots + w_nx_n + b \Rightarrow Y = \sum_{i=1}^{N} w_iX_i + b$$


Train
Validation
Test
Original Set

■ 핵심 개념 이해

- 사이킷런: https://scikit-learn.org/stable/
  - 파이썬으로 머신러닝을 수행하기 위한 쉽고 효율적인 개발 라이브러리를 제공
  - 보스톤 주택 가격 데이터, 붓꽃 데이터 등과 같은 머신러닝 분석용 데이터셋 을 제공
  - 전체 n개의 컬럼 중 앞에서 (n-1)개의 컬럼은 독립 변수 X를 의미
  - 마지막 컬럼 은 종속 변수 Y이며, 데이터셋 객체의 target 배열로 관리



■ 핵심 개념 이해

▪ 분석 평가 지표

- 회귀 분석 결과에 대한 평가 지표는 예측값과 실제값의 차이인 오류의 크기가 됨
- 정확한 평가를 위해 오류의 절대값 평균이나 제곱의 평균, 제곱 평균의 제곱근 또는 분산 비율 을 사용

표 10-1 회귀 분석 결과에 대한 평가 지표

평가 지표	수식	사이킷런 라이브러리
MAE: Mean Absolute Error	$\frac{1}{n}\sum_{i=1}^{n}\lvert Y_i-\hat{Y}_i\rvert$	metrics.mean_absolute_error()
MSE: Mean Squared Error	$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2$	metrics.mean_squared_error()
RMSE: Root Mean Squared Error	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}$	없음
R^2: Variance score, 결정 계수coefficient of determination	$\frac{\text{예측값의 분산}}{\text{실제값의 분산}}$	metrics.r2_score()