Machine Learning

Chapter 3 머신러닝 모델링
(Machine Learning Modeling)

START

Smart Media
스마트미디어인재개발원

학습목표

- 데이터 스케일링의 필요성을 이해 할 수 있다.
- 다양한 스케일링 방법을 알 수 있다.

데이터 스케일링
(Data scaling)

데이터 스케일링(Data scaling) 정의

데이터 스케일링 (Data scaling)

- 특성(Feature)들의 범위(range)를 정규화 해주는 작업
- 특성마다 다른 범위를 가지는 경우 머신러닝 모델들이 제대로 학습되지 않을 가능성이 있다.
  (KNN, SVM, Neural network 모델, Clustering 모델 등)

시력	키
0.2	178
1.0	156
0.5	168
0.3	188
0.6	149

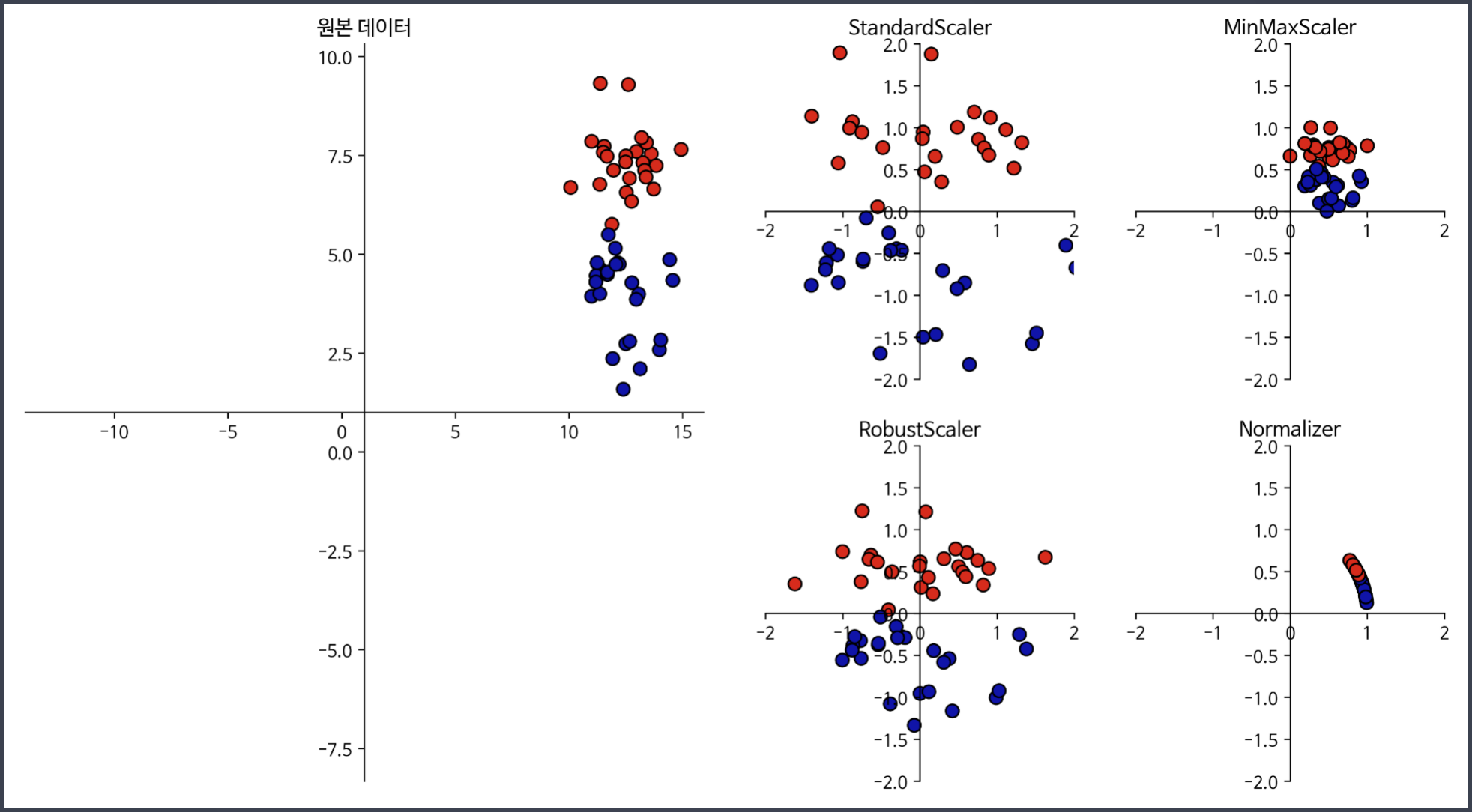
시력과 키를 함께 학습시킬 경우
키의 범위가 크기때문에 거리 값을
기반으로 학습 할 때 영향을 많이 준다.

데이터 스케일링(Data scaling) 장점

장점

- 특성들을 비교 분석하기 쉽게 만들어 준다.
- Linear Model, Neural network Model 등에서 학습의 안정성과 속도를 개선시킨다.
- 하지만 특성에 따라 원래 범위를 유지하는게 좋을 경우는 scaling을 하지 않아도 된다.

데이터 스케일링(Data scaling) 종류



데이터 스케일링(Data scaling) 종류

StandardScaler

- 변수의 평균,표준편차를 이용해 정규분포 형태로 변환 (평균 0, 분산 1)
- 이상치(Outlier)에 민감하게 영향을 받는다.

RobustScaler

- 변수의 사분위수를 이용해 변환
- 이상치(Outlier)가 있는 데이터 변환시 사용 할 수 있다.

MinMaxScaler

- 변수의 Max 값, Min 값을 이용해 변환 (0 ~ 1 사이 값으로 변환)
- 이상치(Outlier)에 민감하게 영향을 받는다.

Normalizer

- 특성 벡터의 길이가 1이 되도록 조정 (행마다 정규화 진행)
- 특성 벡터의 길이는 상관 없고 데이터의 방향(각도)만 중요할 때 사용.

데이터 스케일링(Data scaling) 주의점

주의점

- 훈련세트와 테스트세트에 같은 변환을 적용해야 한다.
- 예를 들어 StandardScaler의 경우 **훈련세트**의 **평균과 표준편차**를 이용해 **훈련세트를 변환**하고, **테스트세트**의 **평균과 표준편차**를 이용해 **테스트세트를 변환**하면 잘못된 결과가 나온다.