스마트인재개발원
Smart Human Resources Development

Machine Learning

chap 5. Data Scaling

김 민 수 연구원

학습목표

- 데이터 스케일링의 개념을 이해할 수 있다.
- 데이터 스케일링의 종류를 알 수 있다.

수업 흐름도

데이터 스케일링 (Data Scaling)이란?

데이터 특성(Feature)들의 값 범위를

일정한 수준으로 맞춰주는 작업

데이터 스케일링

- 특성마다 다른 범위를 가지면서 편차가 큰 데이터의 경우 모델들이 잘못된 결과를 도출할 가능성이 있음(KNN, 선형 회귀, 로지스틱 회귀 등 거리나 수치 기반 모델)

신체검사 데이터에서 시력과 키 특성을 함께 학습시킬 경우,

키의 데이터 범위가 시력에 비해 크기 때문에 데이터 거리 값을 기반으로

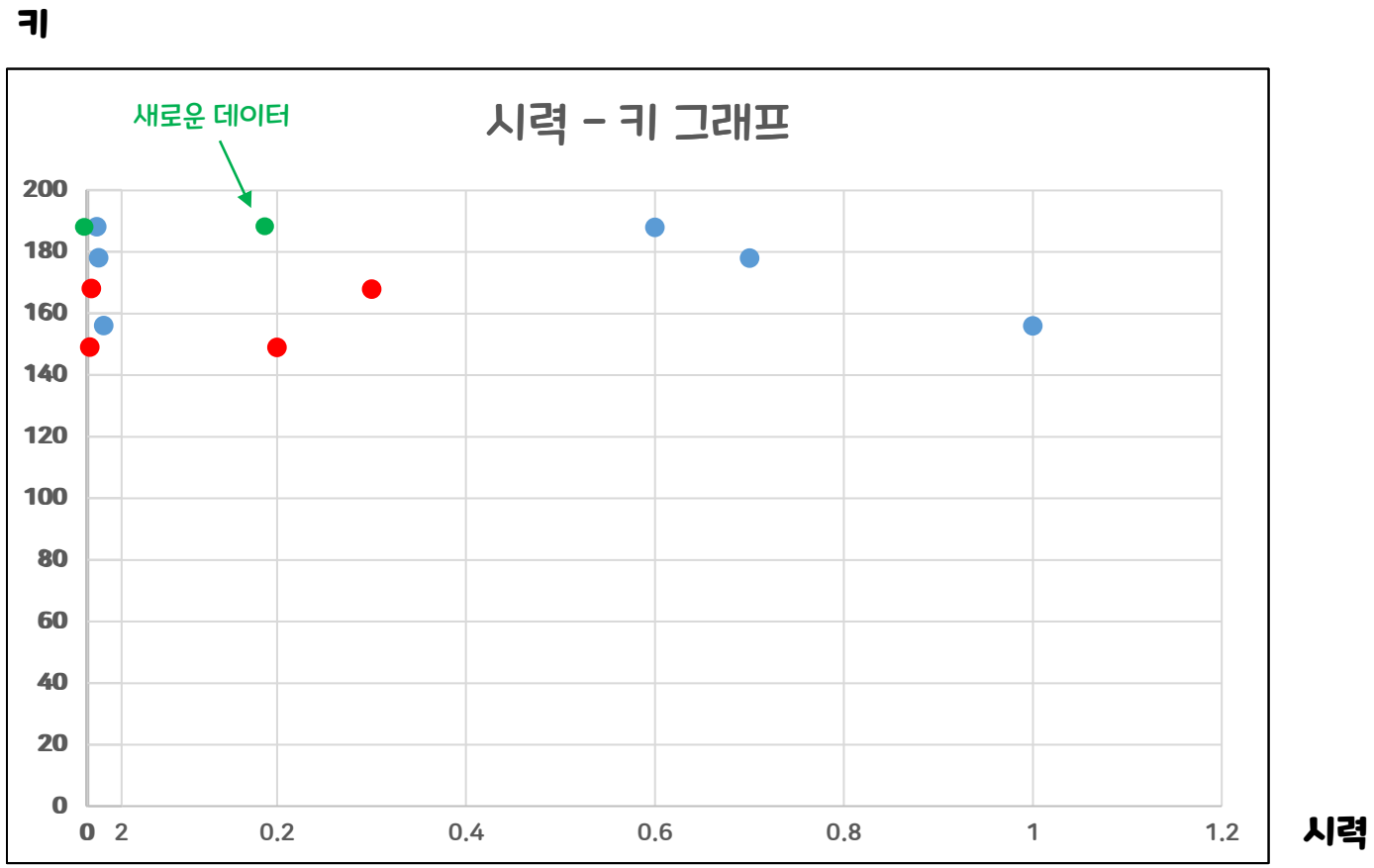
학습할 때 잘못된 영향을 끼칠 수 있음

데이터 스케일링 (Data Scaling)이란?

스마트인재개발원
Smart Human Resources Development

신체검사 데이터

시력	키	합격여부
0.7	178	합격
1.0	156	합격
0.3	168	불합격
0.6	188	합격
0.2	149	불합격

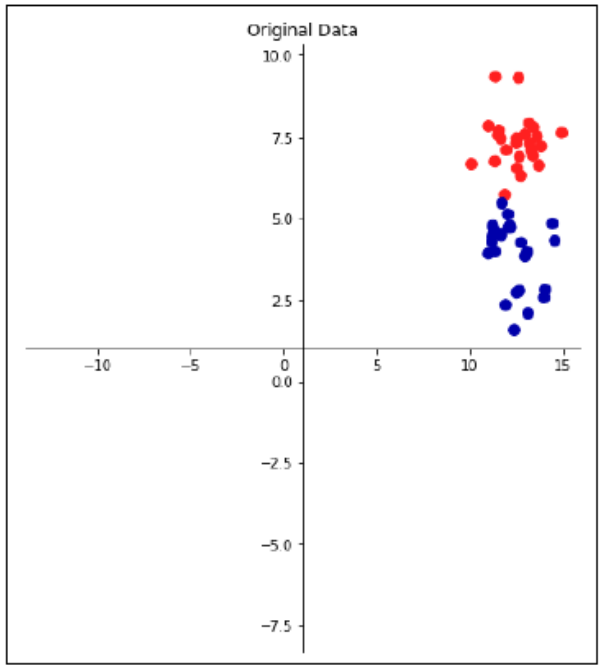


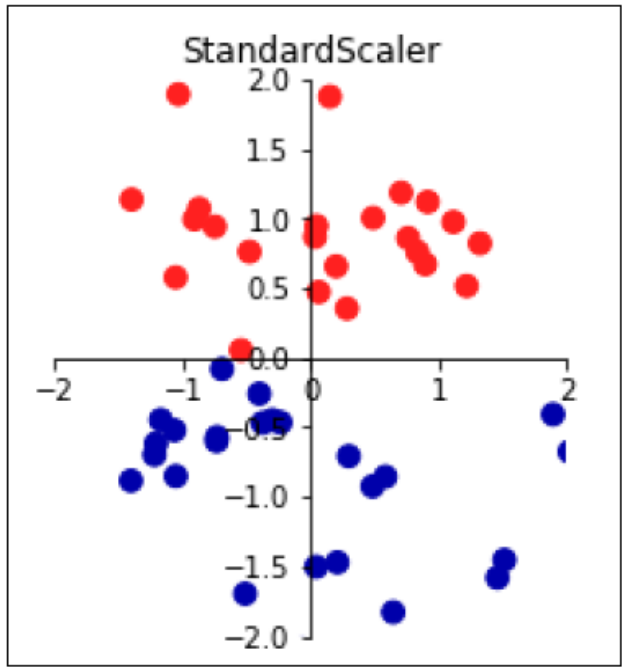
'시력'과 '키'라는 특성으로 합격여부를 판단한다면 '시력'의 중요도는 매우 떨어지게 됨

Standard Scaler

분산 : 데이터가 퍼져 있는 정도
→ 클수록 들쭉날쭉 불안정함

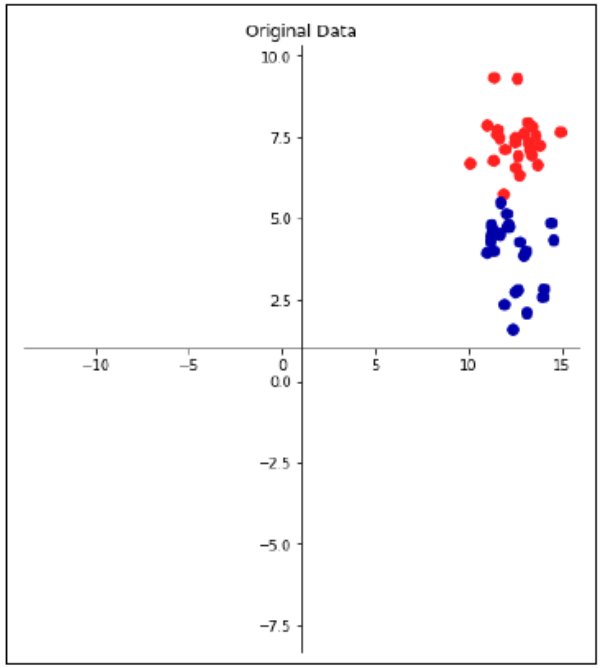
- **변수의 평균, 분산을 이용해 정규분포 형태로 변환 (평균 0, 분산 1)**
- 이상치가 있다면 평균과 분산에 영향을 미쳐 변환된 데이터의 분포는 매우 달라짐

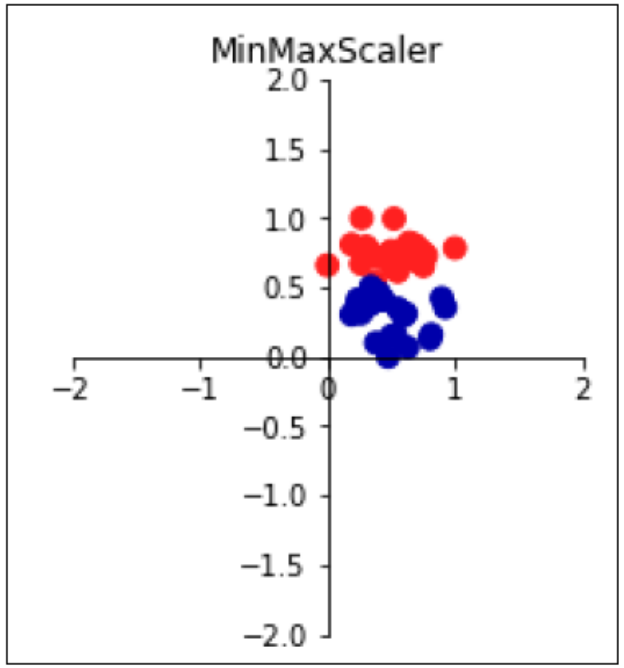




MinMax Scaler

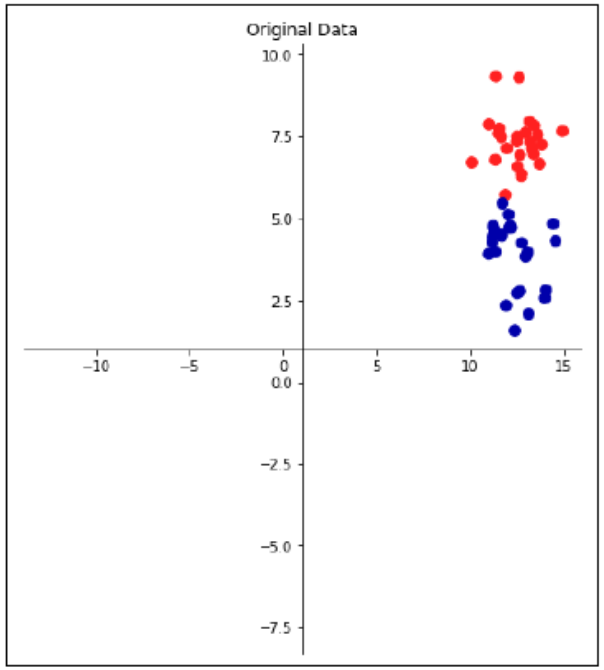
- 데이터를 0 ~ 1 사이 값으로 변환 (-값이 있다면 -1 ~ 1 사이로 변환)
- 이상치가 있다면 사용하기 힘듦

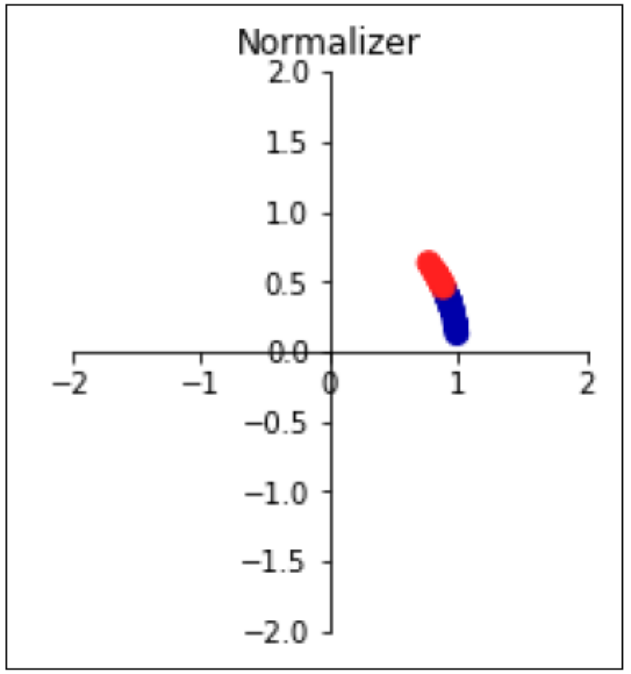




Normalizer

- 데이터를 지름이 1인 원에 투영시킴
- 데이터의 거리는 상관없고 **방향(각도)**만 중요할 때 사용(ex 단어의 유사도 판단)





데이터 스케일링 (Data Scaling)이란?

특징

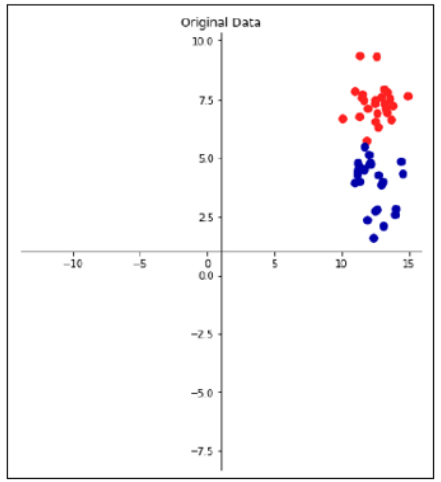
- 거리, 수치 기반 모델 적용시 특성들을 비교 분석하기 쉽게 만들어 예측에 도움을 줌
- 특히 회귀 모델(연속적인 실수를 예측)에서 학습의 안정성과 속도를 개선시킴
- 트리기반 모델 등 거리값에 관계없는 모델들은 굳이 Scaling을 해줄 필요가 없음

  (데이터의 분포가 고르다면 반드시 Scaling을 해야하는 것은 아님)

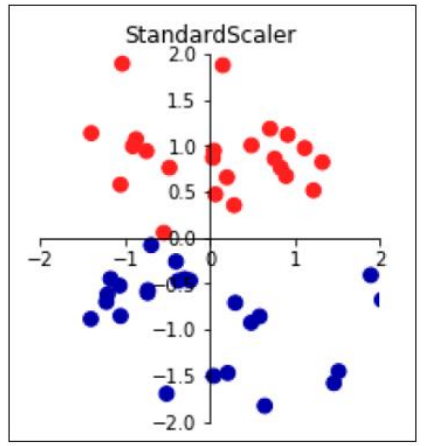
데이터 스케일링 (Data Scaling)이란?

주의점

- 훈련(train) 데이터와 평가(test) 데이터에 같은 변환을 적용해야 함 (예측값의 범위가 달라질 수 있음)



train data



test data