Dacon_nuclear_study_1st_3조

3 조 2 주차 발표자료_피처스케일링

피처 스케일링

서로 다른 변수의 값 범위를 일정한 수준으로 맞추는 작업

표준화와 정규화

표준화

데이터의 피처 각각이 평균이 0이고 분산이 1인 가우시안 정규 분포를 가진 값으로 변환하는 것

표준화를 통해 변환될 피처 x의 새로운 i번째 데이터를 x_new라고 하면

➔ 원래 값에서 피처 x의 평균을 뺀 값을 피처 x의 표준편차로 나눈 값

정규화

서로 다른 피터의 크기를 통일하기 위해 크기를 변환해주는 개념

개별 데이터의 크기를 모두 똑 같은 단위로 변경

새로운 데이터 x_new는

➔ 원래 값에서 피처 x의 최솟값을 뺀 값을 피처 x의 최댓값과 최솟값의 차이로 나눈 값

〈예시〉

Normalizer 모듈

선형대수에서 정규화 개념이 적용

개별 벡터의 크기를 맞추기 위해 변환하는 것을 의미

개별 벡터를 모든 피처 벡터의 크기로 나눔

StandardScaler

표준화를 쉽게 지원하기 위한 클래스

개별 피처를 평균이 0이고 분산이 1인 값으로 변환 ->가우시안 정규 분포를 가질 수 있도록

데이터가 가우시안 분포를 가지고 있다고 가정하고 구현된 선형회기, 로지스틱 회귀, 소프트 벡터 머신 등 적용 가능

예시 코드

모든 칼럼 값의 평균이 0에 가까운 값으로, 분산은 1에 가까운 값으로 변환

MinMax

데이터값을 0과 1사이의 범위 값으로 변환 (음수값의 경우 -1~1)

데이터 분포가 가우시안 분포가 아닐 경우 적용 가능

예시 코드

정리

주로 사용하는 메소드는 fit(), transform(), fit_transform()

학습 데이터와 테스트 데이터 세트로 분리하기 전에 먼저 전체 데이터 세트에 스케일링을 적용한 뒤 학습과 테스트 데이터 세트로 분리하는 것이 바람직