스케일링과 변수변환

- 회귀분석에 사용되는 데이터는 그 자체로 사용하기 보다는 스케일이나 함수변환 등의 전처리 과정을 거치는 경우가 많다. 전처리 과정은 공분산 행렬의 조건 행렬이나 데이터 간의 관계를 선형회귀에 맞게 바꾸기 위해서 사용된다.

- 조건수 (Condition number)

  - 조건수는 공분산행렬의 X^TX의 가장 큰 고유치와 가장 작은 고유치의 비율을 뜻한다.

$$\text{Condition number} = \frac{\lambda_{max}}{\lambda_{min}}$$

  - 조건수가 크면 역행렬을 계산할 때 오차가 이차적 영향이 커진다.

예) $Ax = b$

조건수가 작은경우

*행렬 A가 단위행렬이면 조건수 가장 작은 값으로 조건수의 값이 1이다.

$Cond(I) = 1$

조건수가 큰경우

*조건수가 크면 약간의 오차만 있어도 해가 전혀 다른 값을 가진다. 따라서 조건수가 크면 회귀분석을 사용한 예측값의 오차도 커지게 된다.

- 회귀분석과 조건수

  - 회귀분석에서 조건수가 커지는 경우는 크게 두가지가 있다.

    1. 변수들의 단위 차이로 인해 특정의 스케일이 크게 달라지는 경우. 이경우에는 스케일링(Scaling)으로 해결한다.

    2. 다중공선성 즉, 상관관계가 큰 독립변수들이 있는 경우, 이 경우에도 변수 선택이나 PCA를 사용한 차원축소등으로 해결한다.

스케일링
독립변수의 단위 제곱수이므로 scale 함수를 통해 쉽게 비교하게 계산시켜줄 수 있다.
scale 함수는 표준편차와 평균을 사용한다.

변수변환
독립변수나 종속변수가 심하게 한쪽으로 치우친 분포를 보이는 경우 / 독립변수나 종속변수와의 관계가 곡선형태 혹은 나팔형으로 퍼진 경우 / 종속변수와 예측치가 비선형 관계를 보이는 경우
로그함수 및 제곱근을 이용하면 회귀분석이 향상될 수 있다.