정규화(Normalization) ※ '정규화·표준화'가 '이상치 제거'를 의미하지 않음

$$x_{new} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

※ 표준화와 정규화를 이를 구분해서 쓰지 않을 때도 있다. (맨 마지막 페이지 참고)

- 데이터를 특정 구간으로 바꾸는 척도법이다 (ex. 0~1 or 0~100).
- 식 : (측정값 - 최소값) / (최대값 - 최소값)
- 데이터 군 내에서 특정 데이터가 가지는 위치를 볼 때 사용된다.
- 주가와 같은 주기를 띄는 데이터의 경우 과거에 비해서 현재 데이터의 위치가 어느정도 인지 파악하기에 좋아진다.

이를 '정규화'라고 표현할 때도 있다.

표준화(Standardization)

<특정 데이터 분포를 정규이 '0'이고 분산이 '1'인 분포로 변환함.>

$$z_i = \frac{x_i - \bar{x}}{s}$$

- 데이터를 0을 중심으로 양쪽으로 데이터를 분포시키는 방법이다. 표준화를 하게 되면 각 데이터들은 평균을 기준으로 얼마나 떨여져 있는지를 나타내는 값으로 변환된다.
- 식 (Z-score 표준화) : (측정값 - 평균) / 표준편차
- 변환된 데이터는 다소 평평하게 만드는 특성을 가진다 (진폭의 감소). 진폭의 감소로 각 데이터의 간격이 감소하게 된다 (ex. 10000의 단위에서 0.1 단위로 감소).

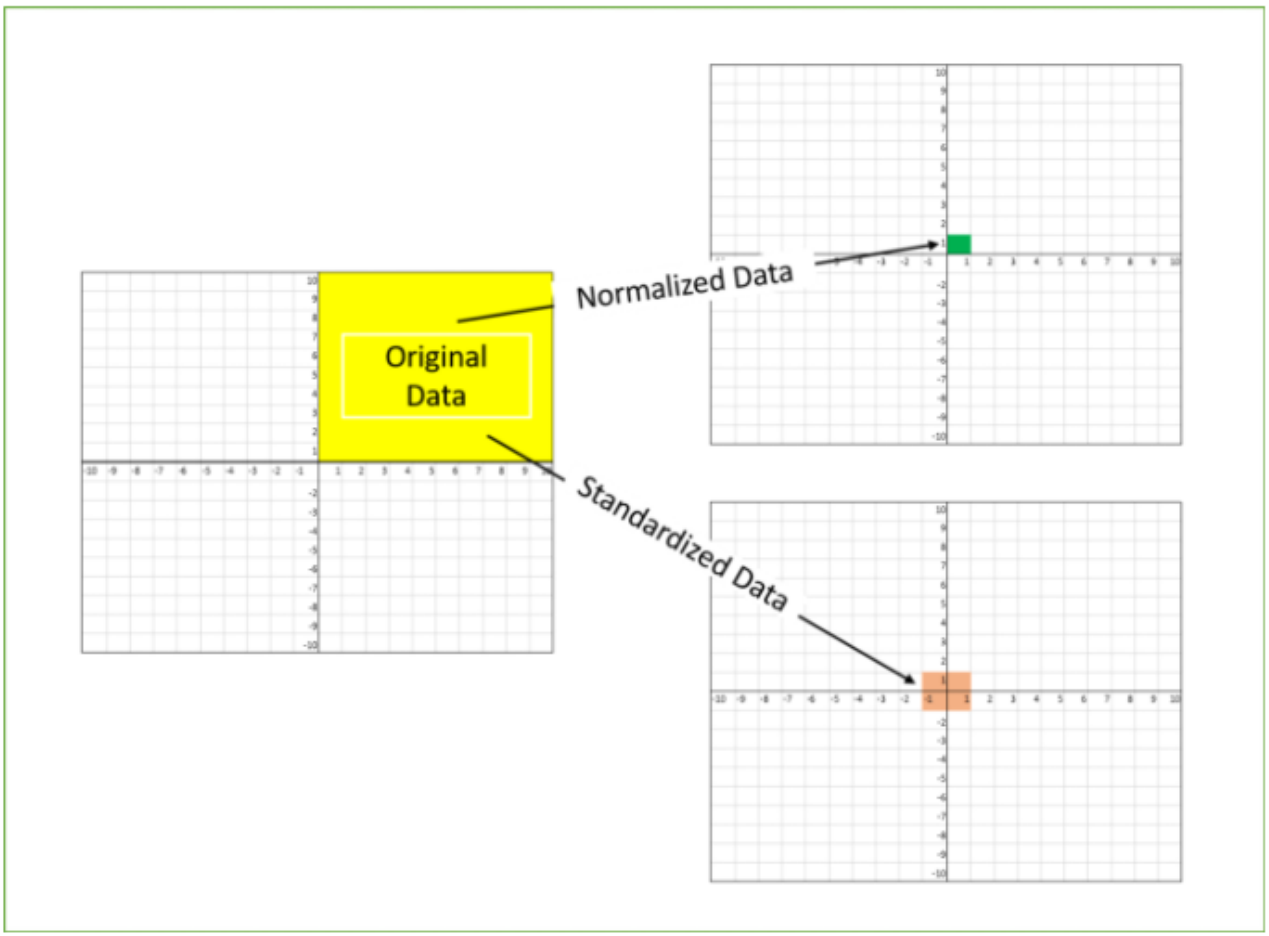
※ 신경망 훈련을 할 때, 입력 데이터들을 정규화(or 표준화)하면, 2차함수의 모형이 symmetric 하게 된다 (좌우)대칭적인, 정형이 잡힌

↕

elongated : 비정상적으로 가늘고 긴.

Normalization is used when we want to bound our values between two numbers, typically, between [0,1] or [-1,1]. While Standardization transforms the data to have zero mean and a variance of 1, they make our data "**unitless**". Refer to the below diagram, which shows how data looks after scaling in the X-Y plane.

단위가 없



02. 정규화 (Normalization)

- 수식: (요소값 - 최소값) / (최대값 - 최소값)
- 정규화는 전체 구간을 0~100으로 설정하여 데이터를 관찰하는 방법입니다.

이 방법은 데이터 군 내에서 특정 데이터가 가지는 위치를 볼 때 사용합니다.

시세와 같이 주기를 띄는 데이터의 경우 과거 대비 현재 데이터의 위치를 파악하기에 용이합니다.

일반 가격 데이터에 정규화 적용

정규화/표준화는 데이터를 보는 유일한 값은 아니지만, 데이터를 볼 때 중간단계 가공 방법 혹은 대략적으로 형태를 볼 때 유용하게 쓰이며 수식 또한 어렵지 않습니다.

여기까지 우선 데이터를 살펴보기 위한 기본적인 함수/방법을 마치고, 다음 단원에선 데이터 전처리에 대해 기록하도록 하겠습니다. 데이터 전처리는 정해져 있는 것이 아니라서 경험 위주의 기술이며, 생각을 더듬어가며 기록해야 하기에 약간의 시일이 소요될 듯 합니다.

01. 표준화 (Standardization) = '큰 정규화'

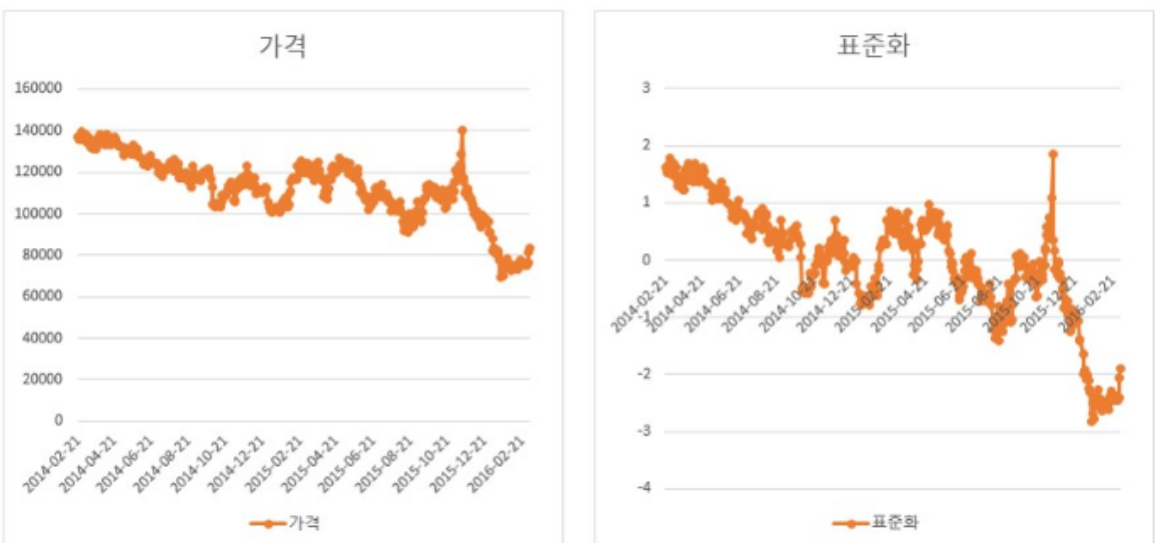
- 수식: (요소값 - 평균) / 표준편차

- 평균을 기준으로 얼마나 떨어져 있는지를 나타내는 값으로, 이 방법을 적용하려는 때는 "2개 이상의 대상"이 "단위"가 다를 때 대상 데이터를 같은 기준으로 볼 수 있게 합니다.

예를 들어, 삼성전자와 현대차의 주식시세에 대해 동일 기간에 대해 표준화를 수행하면 두 종목이 어떤 특징을 가지고 움직이는 지 관찰하는데 도움이 됩니다.

또 다른 예시로 몸무게와 키에 대해 표준화를 수행한 후 데이터를 보았을 때 몸무게는 음수, 키는 양수 값이 나온다면 그 사람은 평균보다 키가 크고 몸이 마른 편이라 볼 수 있습니다.

- 또한 이 방법은 데이터를 다소 평평하게 하는(로그보다는 덜하지만 데이터의 진폭을 줄이는) 특성을 가집니다. 이 방법을 적용하면 간극이 줄어드는 효과가 발생하여 고객별 매출금액과 같이 간극이 큰 데이터의 간극을 줄이는 결과를 얻게 됩니다. 그 결과 분석 대상 고객군을 정하는 데 (약간의) 편의성을 제공하게 됩니다.



일별 가격 데이터에 표준화 적용

1. Definition

There are different types of data normalization. Assume you have a dataset X, which has N rows(entries) and D columns(features). $X[:,i]$ represent feature i and X[j,:] represent entry j. *We have:*

Z Normalization(Standardization):

$$\hat{X}[:,i] = \frac{X[:,i]-\mu_i}{\sigma_i}, (\mu_i = \frac{1}{N} * \sum_{k=1}^{N} X[k,i], \sigma_i = \sqrt{\frac{1}{N-1} * \sum_{k=1}^{N}(X[k,i]-\mu_i)^2}\ )$$

I used to falsely think this method somehow yields a standard Gaussian result. In fact, standardization does **not change** the type of distribution:

$$\hat{X} = aX + b \rightarrow f_{\hat{X}}(x) = \frac{1}{|a|} f\left(\frac{x-b}{a}\right)$$

pdf of standardized data

This transformation sets the mean of data to 0 and the standard deviation to 1. In most cases, standardization is used feature-wise

Min-Max Normalization:

$$\hat{X}[:,i] = \frac{X[:,i]-\min(X[:,i])}{\max(X[:,i])-\min(X[:,i])}$$