

# Feature Scaling

# Why Feature Scaling?

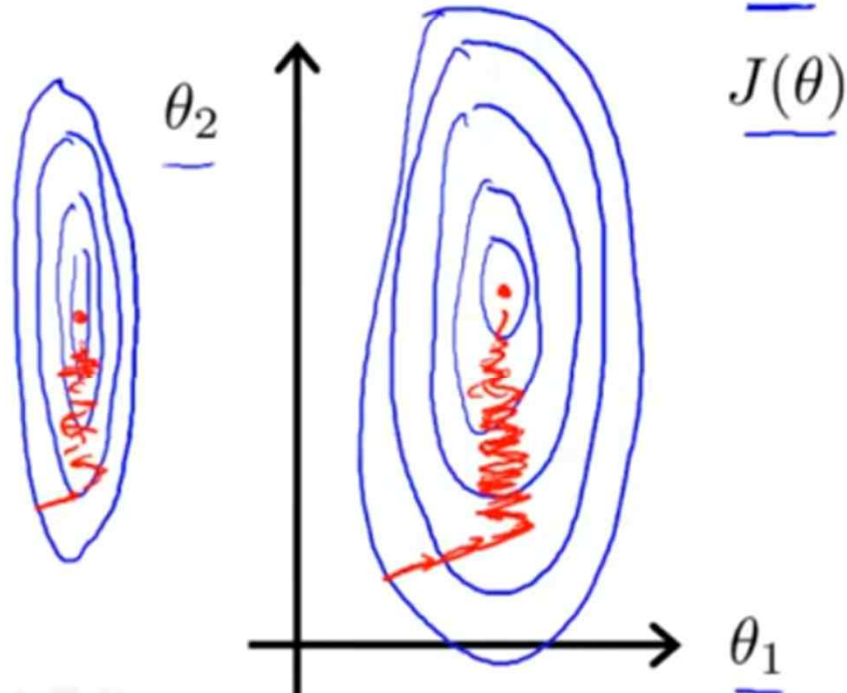
- Feature 별로 스케일이 다른 아래와 같은 상황을 예상해보자~
  - $X_1$  : 0~1사이의 값을 가짐
  - $X_2$  : 0.00001 ~ 0.00002사이의 값을 가짐
  - $X_3$  : 10000 ~ 20000000000 사이의 값을 가짐.
  - Target  $y$  : 1000000~25000000000사이의 값을 가진다고 하자..
- 이러한 상황이라면 거의  $X_2$ ,  $X_1$ 은 거의 영향을 안 미친다고 판단이 될 가능성이 큼!!
- 결론) 모든 특성의 범위를 같게 해준 상태에서 판단을 해야함!!!

## Feature Scaling

Idea: Make sure features are on a similar scale.

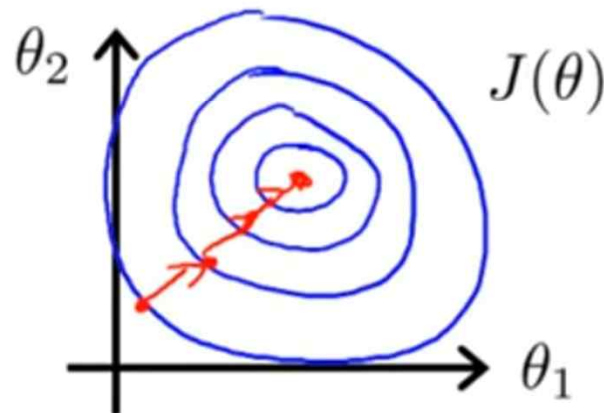
E.g.  $x_1 = \text{size (0-2000 feet}^2\text{)}$  ←

$x_2 = \text{number of bedrooms (1-5)}$  ←



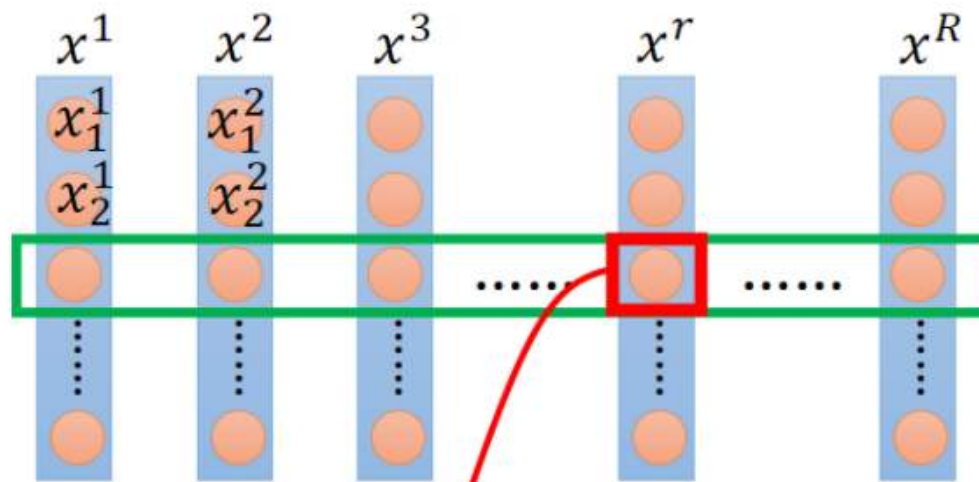
→  $x_1 = \frac{\text{size (feet}^2\text{)}}{2000}$  ←

→  $x_2 = \frac{\text{number of bedrooms}}{5}$  ←



- Feature Scaling : 서로 다른 Feature 값의 범위를 일정한 영역대로 맞추는 작업을 의미함.
  - Feature 자체의 측정 단위가 달라서 발생하는 문제 예방
  - 학습을 하는 과정에서 특정 Feature의 편향성을 가지면서 학습을 하는 현상을 방지함. -> Bias 또한 커질 수 있어서 이에 대한 처리를 위해서도...

# Feature Scaling Method : Standardization



For each dimension  $i$ :

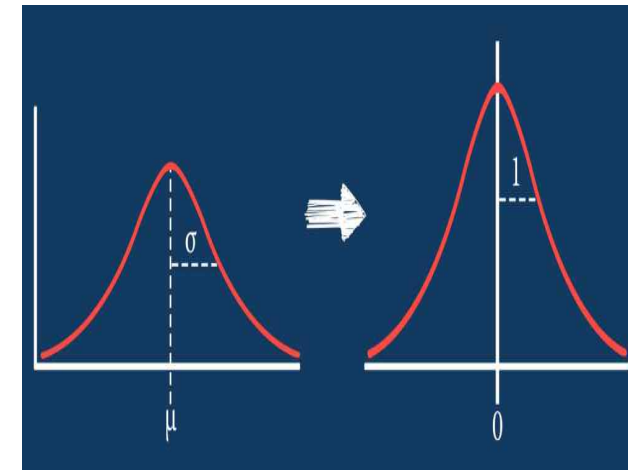
mean:  $m_i$

standard

deviation:  $\sigma_i$

$$x_i^r \leftarrow \frac{x_i^r - m_i}{\sigma_i}$$

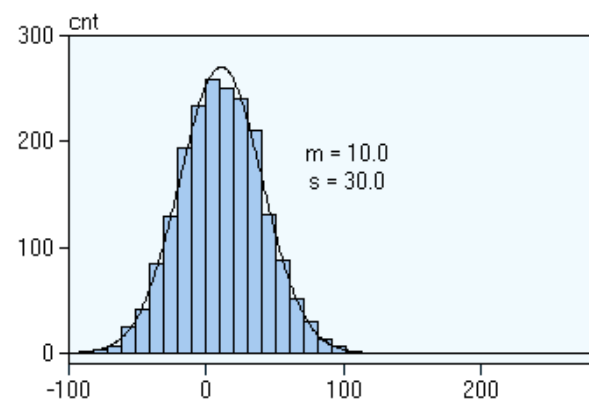
The means of all dimensions are 0, and the variances are all 1



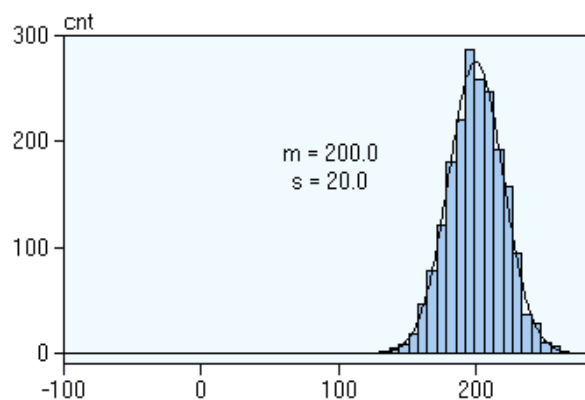
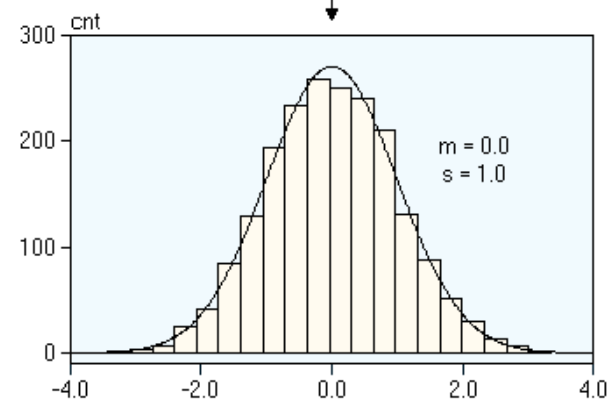
Ref) [Why feature normalization / standardization? | Develop Paper](#)

# Feature Scaling Method : Standardization

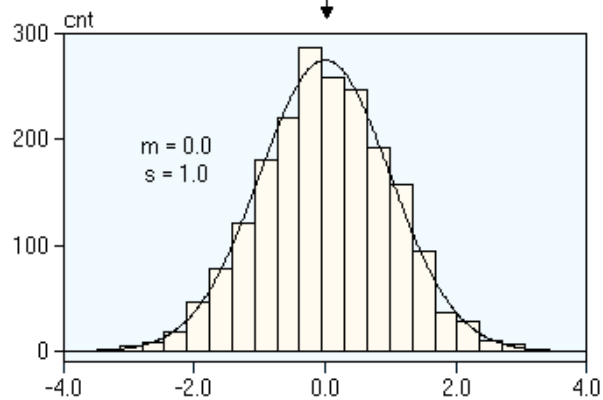
- 중요 부분 : 평균을 0, 분산이 1인 정규분포로 변환하는 작업!!!
  - 최소값, 최대값 등의 크기를 제한을 하지 않고, 정규화 변환이 가능하기에 이상치에 대한 부분을 고민을 해야함!!!
  - 이상치에 의해서 값이 상당히 변화가 심할 수 있어서, 이상치에 민감함!!!
  - 반듯이 이상치에 대한 고민을 꼭 하면서 해야함!!!!
  - 분포에 Focus!!!!
- SVM, Linear Regression, Logistic Regression 등 데이터의 분포가 정규분포를 가정하고 하는 모델에 대해서 성능을 향상시키기  
기에 적용할 수 있는 부분임!!!!



Standardisation



Standardisation



comparable distributions  
( $m = 0.0, s = 1.0$ )

# Feature Scaling Method : Normalization

## Normalization Formula

$$X_{\text{normalized}} = \frac{(X - X_{\text{minimum}})}{(X_{\text{maximum}} - X_{\text{minimum}})}$$





# Feature Scaling Method : Normalization

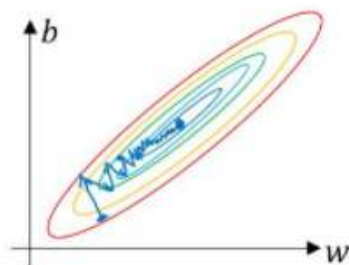
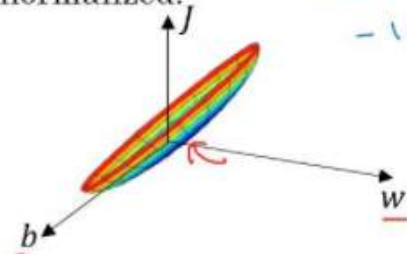
- 중요부분 : 값의 범위를  $[0,1]$ 로 옮기게 된다.
  - 앞에서 변환 수식을 보아도 이상치에 영향을 크게 받을 수 밖에 없는 구조임!!!
  - min, Max를 사용해서 하기에 이 기준 값들이 바로 경우에 따라서는 이상치가 될 수 있기 때문임!!!
  - 앞의 Standardization의 경우에 있어서는 다른 데이터에 의해서 이상치(min, max)가 좀 조절이 될 여지가 있을 수 있지만, Normalization의 경우에는 바로 그 값이 변환 기준이 되기에 더 유의를 해야함!!!

정리) Standardization 은 상치를 포함해서 변환을 하고, 이상치를 제거를 해도 괜찮을 수 있지만, Normalization 은 이상치를 제거를 하고 해야 좀 더 괜찮은 경향이 있다!!!

# Feature Scaling Method : Normalization

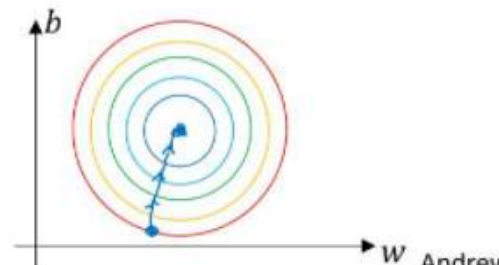
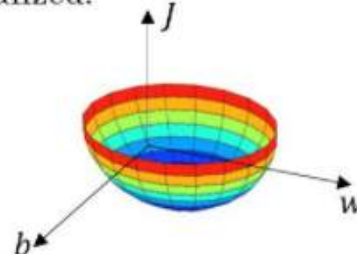
Why normalize inputs?

Unnormalized:  
 $w_1$   $x_1: 1 \dots 1000 \leftarrow$   
 $w_2$   $x_2: 0 \dots 1 \leftarrow$   
 $-1 \dots 1$



Normalized:

$$J(w, b) = \frac{1}{m} \sum_{i=1}^m \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

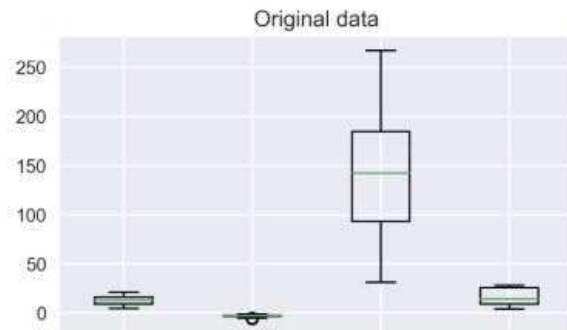


normalize input

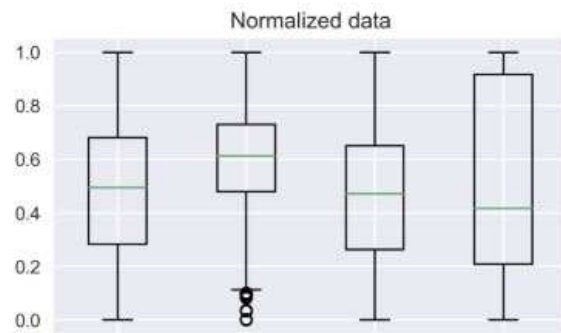
Normalization을 해주게 되면, Loss Function을 찾아가는 과정에 있어서도 좀 빠르게 찾아갈 수 있다.

왼쪽의 경우에는 처리가 안되어서 Gradient Descent로 할 때 이리저리 왔다갔다 하는데 (oscillation), 오른쪽의 경우는 보다 편하게 접근이 가능함.

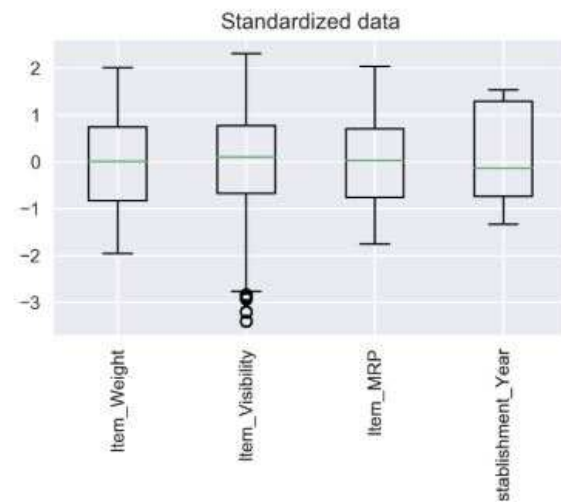
→ Gradient Descent 기반에서는 필요함!!!!



각 특성의 스케일이 다른 원본 데이터

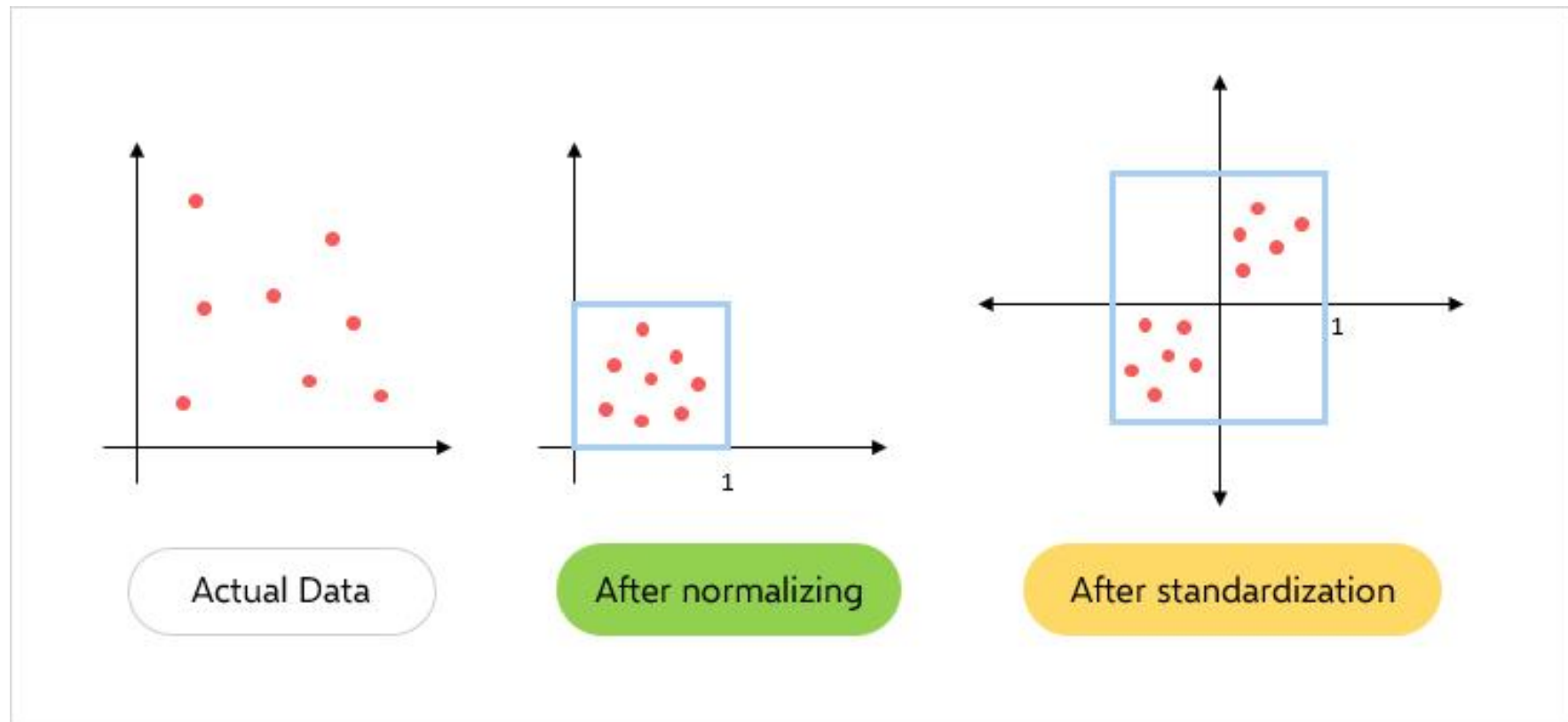


Normalization을 하여서 값을 0~1로 bound를 함!



Standardization을 하여서 평균을 0으로 만들어서 분포를 시킴!

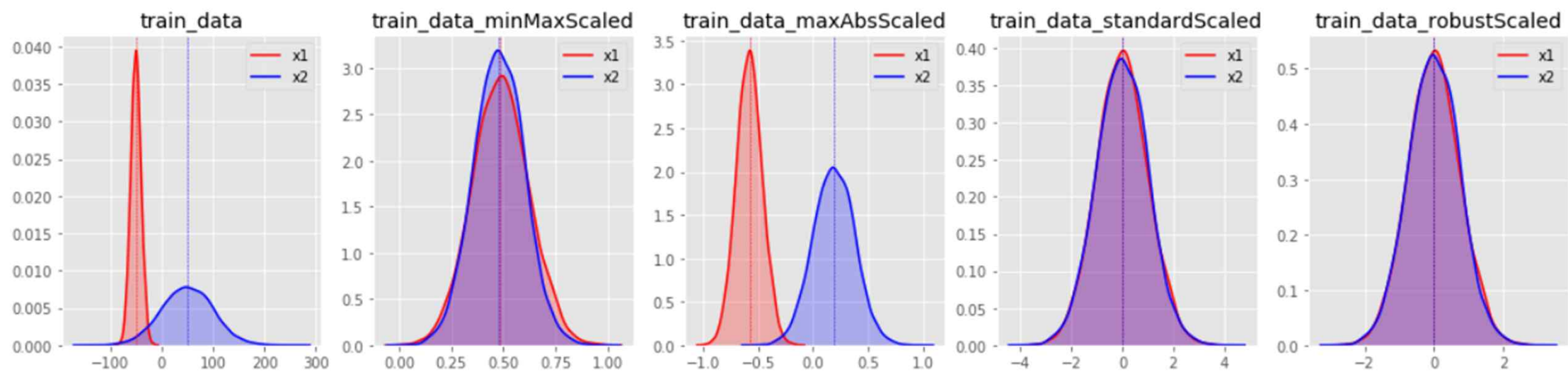
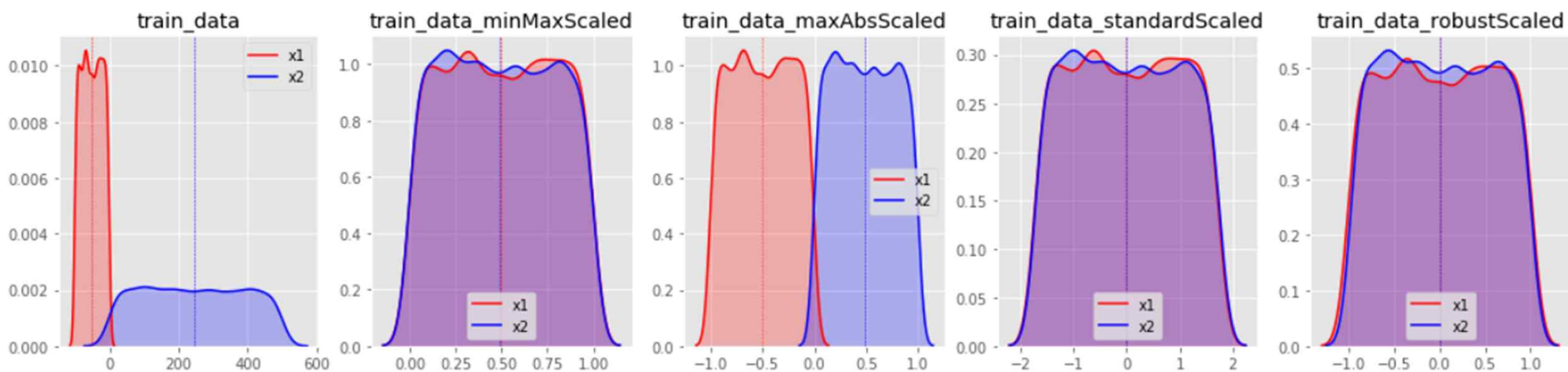
결론은  
어떤 데이터냐  
어떤 모델을 사용할  
것인가에 따라서  
각기 다르게 된다!!!

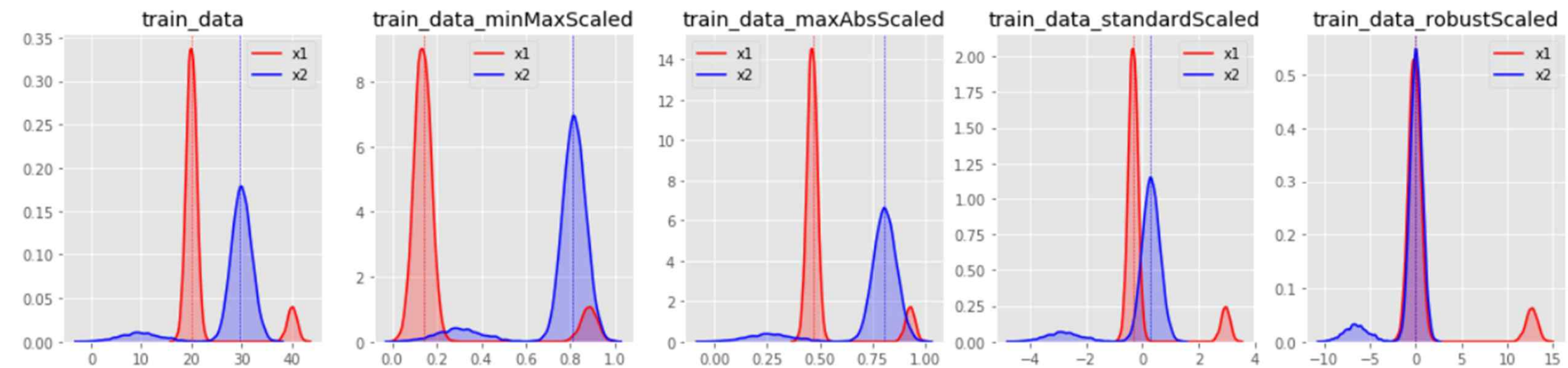
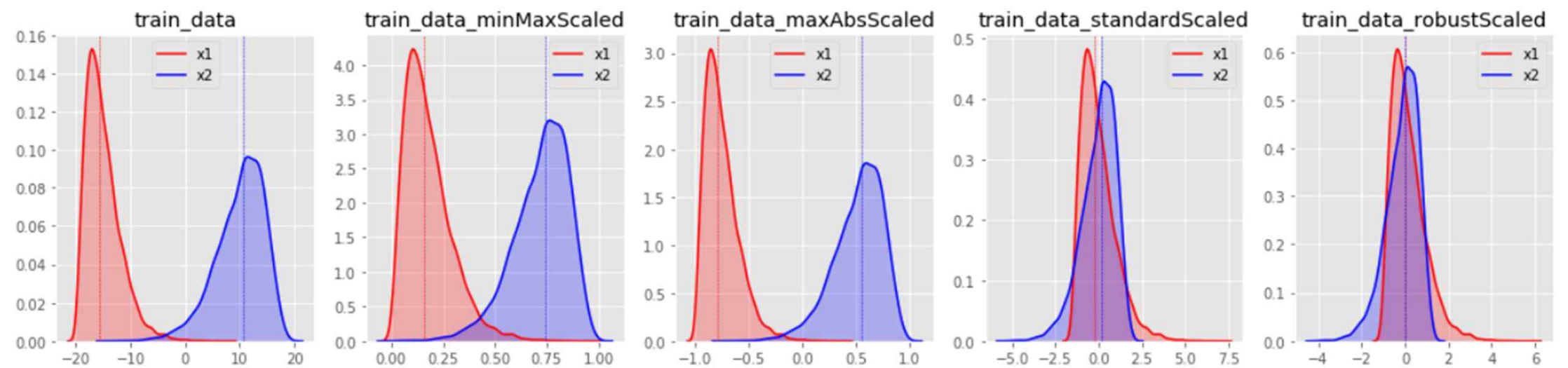


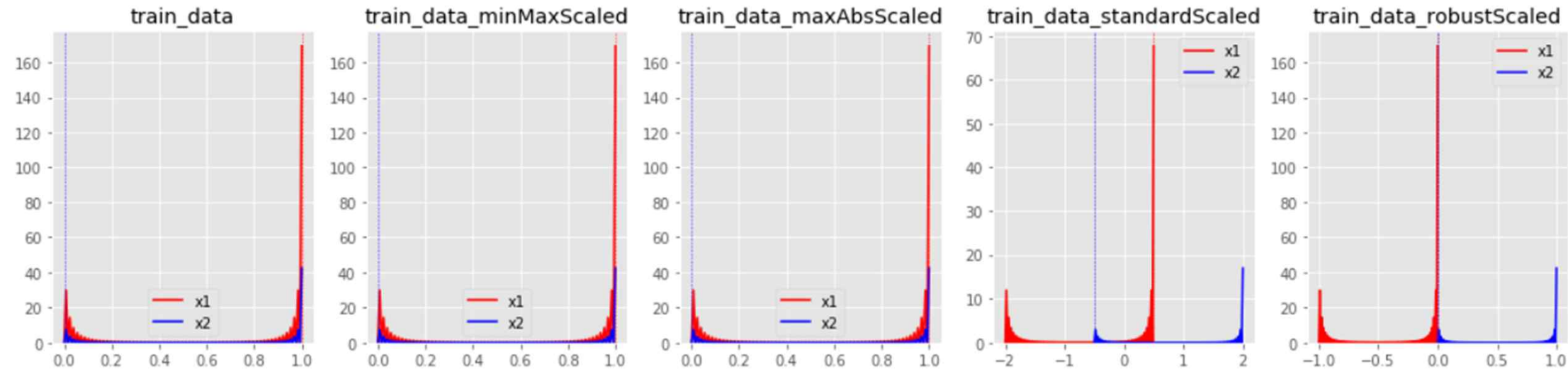
Ref) [What Do Normalization and Standardization Mean? When to Normalize Data and When to Standardize Data? | by Akalbir Singh Chadha | Becoming Human: Artificial Intelligence Magazine](#)

# ScikitLearn : Scaler들에 대해서...

- StandardScaler()
  - 특성들을 평균0, 분산1로 scaling 작업
  - 정규 분포를 유지하게 변환을 수행
  - 최소, 최대를 제한하지 않아서 이상치에 영향을 받을 수는 있음!
- MinMaxScaler()
  - 가장 작은 값을0, 가장 큰 값은 1로 bound
  - 이상치(min, max)에 대해서 기준이 되어서 이상치에 아주 크게 영향을 받음!!!
- MaxAbsScaler()
  - 각 특성의 절대값이 1, 0의 값이 0이 되도록 -> -1 ~ 1 사이로 bound
  - 특징이 양수로만 되어 있다면 MinMaxScaler와 동일한 역할을 함.
  - 당연히 이상치(min, max)에 대해서 기준이 되어서 이상치에 아주 크게 영향을 받음!!!!
- RobustScaler()
  - 평균과 분산을 사용하는 것 대신에 중앙값 & 사분위값을 사용!!!
  - 이상치의 영향을 최소화할 수 있음!!!!
- Normalizer()
  - 위의 방식들은 모두 세로 컬럼별로 수행을 하였으나, 이 방법은 1개의 sample에 대한 가로줄이 기준!!!
  - 한 1개의 sample에서 값을 모두 L2 norm = 1이 되도록 변환
  - 보통 학습에서 보다는 딥러닝 등에서 학습 벡터에 주로 사용하며, 일반적인 피쳐에서는 잘 사용하지 않음.







[ 앞의 샘플에 대해서 각기 보면]

- Standard Scaler, Robust Scaler가 표준화된 분포로 변환을 해주고 있다.
- Minmax, maxAbs의 경우에는 특정한 값이 몰려있는 경우에는 그 형태가 그대로 많이 유지되고 있음을 보이고 있다.
- 마지막의 outlier의 경우에는 모든 방법들에 있어서 심각하게 왜곡현상을 야기하고 있음!

- 모든 컬럼을 동일한 기준으로 수행할 필요도 없으며!!!!
- 모든 컬럼에 대해서 적용할 필요도 없으며,
- 특정 컬럼의 값을 펴야할지
- 특정 컬럼의 값의 형태를 유지하면서 값만 조정할지

여러 상황에 따라서 결정을 해야하지, 절대적인 부분이 있는 것은 아님!!!!!!

+ 참고로 수치값의 단위가 너무 크거나 그러면 log도 취하는 것도 같이 사용!!!



# When Scaling (Standardization, Normalization )

- **Gradient Descent** 기반의 최적화를 사용하는 알고리즘 ( Linear Regression, Logistic Regression, Neural Network etc) 에 있어서는 Scaling 작업이 필요하다!!!

$$\theta_j := \theta_j - \alpha \frac{1}{m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

위의 수식에서 보듯이 서로 다른 크기의 Feature의 경우에는 서로 다른 Step Size를 계산을 하게 되고, 그 결과 Gradient Descent의 Oscillation이 발생을 하고, 그 결과 최적화 가는 과정이 Slow하게 된다!

# When Scaling (Standardization, Normalization )

- Tree Based Algorithm들은 Scaling 작업에 대해서 영향이 거의 없음!!!!!!!
- 이유는 Tree based algo들은 중요한 해당 피처로 나뉘었을 때 여러가 어떻게 되는지 보면서 하나씩 기준을 정해나가기 때문에 전체적인 피처들의 크기가 전혀 학습에 영향을 미치지 않음!!!  
→ 정보의 불순도만 신경쓰면 되니!!!!
- 따라서 일반적으로 Tree Based Algo의 경우에는 Scaling 작업이 영향을 미치지 못한다!!!!!!

# When Scaling (Standardization, Normalization )

- Distance Based Algorithm 기반(KNN, K-means, SVM etc)은 거리가 상당히 중요한 부분이므로, 이 계산에 있어서 영역대가 다르게 되면 다르게 거리가 계산이 되어서 왜곡현상이 벌어질 수 있음.

	Student	CGPA	Salary '000
0	1	3.0	60
1	2	3.0	40
2	3	4.0	40
3	4	4.5	50
4	5	4.2	52

원본 데이터

	Student	CGPA	Salary '000
0	1	-1.184341	1.520013
1	2	-1.184341	-1.100699
2	3	0.416120	-1.100699
3	4	1.216350	0.209657
4	5	0.736212	0.471728

Scaling 작업 후

- Distance AB before scaling =>  $\sqrt{(40 - 60)^2 + (3 - 3)^2} = 20$
- Distance BC before scaling =>  $\sqrt{(40 - 40)^2 + (4 - 3)^2} = 1$
- Distance AB after scaling =>  $\sqrt{(1.1 + 1.5)^2 + (1.18 - 1.18)^2} = 2.6$
- Distance BC after scaling =>  $\sqrt{(1.1 - 1.1)^2 + (0.41 + 1.18)^2} = 1.59$

When Scaling (

When Standardization ?

When Normalization ? )

- Normalization : 수집한 데이터가 정규분포를 따르지 않아도 되는 KNN, Neural Network 같은 알고리즘을 사용하려면, Gradient Descent 이유 및 거리 계산을 적정 수준으로 Bound 해야하니 Normalization을 수행해보도록 추천
- Standardization의 경우에는 Normalization 같이 값이 bound는 되지 않기에, 정규 분포 중심의 가정을 통한 알고리즘에 적용
- 실제는 그냥 raw data로 해보기, normalization 해서 해보기, standardization 해서 해보기, 각기 선택적(일부 컬럼은 스케일링 안 하기, 일부 컬럼은 normalization, 일부 컬럼은 standardization 하기)등으로 직접 해서 평가하기!!!!