



Isabella Lindgren

Follow

Jul 25, 2019 · 5 min read · Listen



Save



Sign in to Medium with Google



Siqi Fang

siqifang47@gmail.com

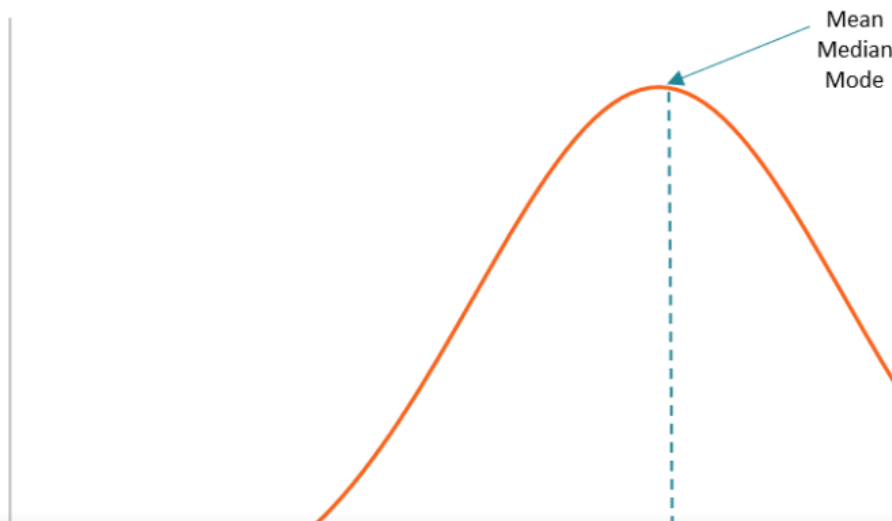
Continue as Siqi

# Transformations, Scaling and Normalization

By: Isabella Lindgren

Raw data comes in all kinds of strange distributions so sometimes it is difficult to analyze and especially to create models without some preprocessing. There are a variety of ways to shape the data into a more favorable input, so here is a quick break down of a few commonly used methods of transforming our data!

The overall goal of transforming our data is to create a more normal (\*Gaussian\*) distribution aka a bell curve. In general, normal distributions tend to produce better results in a model because there are about equal observations above and below the mean and the mean and median are the same. Models run under the assumption your data is normally distributed.





Sign in to Medium with Google



Siqi Fang

siqifang47@gmail.com

Continue as Siqi

By transforming our data we are not only as well. Normalization makes training models can better solve for coefficients. The coefficient which the changes to the value of one variable predict change to the value of another variable. Normalizing and scaling are two types of transformations that are important in data cleaning.

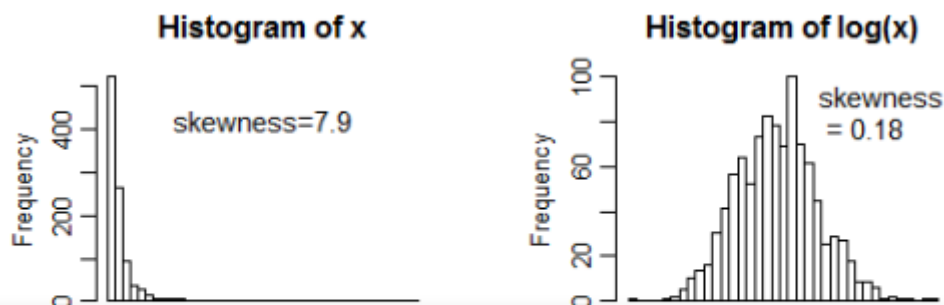
### So what is the difference between Normalizing and Scaling?

The main difference between normalizing and scaling is that in normalization you are changing the **shape of the distribution** and in scaling you are changing the **range of your data**. Normalizing is a useful method when you know the distribution *is not* Gaussian. Normalization adjusts the values of your numeric data to a common scale without changing the range whereas scaling shrinks or stretches the data to fit within a specific range.

Scaling is useful when you want to compare two different variables on equal grounds. This is especially useful with variables which use distance measures. For example, models that use Euclidean Distance are sensitive to the magnitude of distance, so scaling helps even the weight of all the features. This is important because if one variable is more heavily weighted than the other, it introduces bias into our analysis.

### What are some methods we can use to transform our data?

#### *Log Transformation:*





This is one of the most commonly used transformations for data to reduce variability and make your data easier to interpret the data and it helps meet the

### ***Min-Max Scaling:***

The objective of Min-Max scaling is to shift the values closer to the mean of the column. This method scales the data to a fixed range, usually [0, 1] or [-1, 1]. A drawback of bounding this data to a small fixed range is that we will, in turn, end up with smaller standard deviations, which suppresses the weight of outliers in our data.

$$z = \frac{x - \min(x)}{[\max(x) - \min(x)]}$$

### ***Standardization (Z-Score Normalization):***

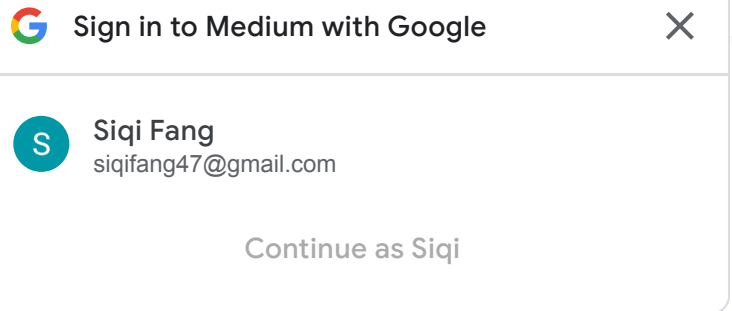
Standardization is used to compare features that have different units or scales. This is done by subtracting a measure of location ( $x - \bar{x}$ ) and dividing by a measure of scale ( $\sigma$ ).

$$x' = \frac{x - \bar{x}}{\sigma}$$

This transforms your data so the resulting distribution has a mean of 0 and a standard deviation of 1. This method is useful (in comparison to normalization) when we have important outliers in our data and we don't want to remove them and lose their impact.

### ***Unit Vector Transformation:***

This method uses the Pythagorean Theorem ( $v_x^2 + v_y^2 = v^2$ ) in order to determine the





$$||\vec{v}|| = \sqrt{v}$$

Formula to deter



Sign in to Medium with Google



Siqi Fang

siqifang47@gmail.com

Continue as Siqi

To normalize the vector, we divide each component by the magnitude of the vector in order to scale down to 1. For example, a vector with value 10 divided by 10 equals 1. To scale down to vector size 1, all other components need to be divided by the same amount, 10, as well. So using this method, we can change the length of the vector without affecting the direction.

When performing unit vector transformations, you can create a new variable  $x'$  with a range  $[0,1]$ .

$$x' = \frac{x}{||x||}$$

### ***Mean Normalization:***

This normalization will create the distribution of features between  $[-1, 1]$  by dividing by the standard deviation.

$$x' = \frac{x - \text{mean}(x)}{\text{max}(x) - \text{min}(x)}$$

### ***Box Cox Transformation:***

Box Cox is used to stabilize the variance (eliminate heteroskedasticity) and transform non-normal dependent variables to a normal shape.





$$(4.3) \quad y_i^{(\lambda)}$$

Form



Sign in to Medium with Google



Siqi Fang

siqifang47@gmail.com

Continue as Siqi

Any value of  $\lambda$  when our datapoint ( $y$ ) is equal to 1 evaluates to 1. Therefore, when we subtract our datapoint ( $y^\lambda$ ) from 1, we center our transformed data around 0. By dividing by  $\lambda$ , we are normalizing the exponential increase of  $\lambda$  from the numerator.

The boxcox function in Scipy tests a range of  $\lambda$  values and returns the value that makes your data look the most normal. It is also important to note that boxcox only works if all the data is positive and greater than 0, which can be easily achieved by adding a constant ( $c$ ) to all data before transforming.

These are just a few of the ways we can transform our data. These methods become especially useful and necessary when using machine learning algorithms.

### References:

1. [https://sebastianraschka.com/Articles/2014\\_about\\_feature\\_scaling.html](https://sebastianraschka.com/Articles/2014_about_feature_scaling.html)
2. <https://www.statisticshowto.datasciencecentral.com/box-cox-transformation/>
3. <https://www.khanacademy.org/computing/computer-programming/programming-natural-simulations/programming-vectors/a/vector-magnitude-normalization>
4. [https://en.wikipedia.org/wiki/Feature\\_scaling](https://en.wikipedia.org/wiki/Feature_scaling)



