

# Data Transformation

# Standardization (Scaling)

Standardization refers to shifting the distribution of each attribute to have a mean of zero and a standard deviation of one (unit variance).

Scale function is used for standardization

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

# Scaling features to a range

Another standardization is scaling features to lie between a given minimum and maximum value.

MinMaxScaler is used for this.

Min and Max will be taken from the input

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

# Data Normalization

Normalization refers to rescaling real valued numeric attributes into the range 0 and 1.

It is useful to scale the input attributes for a model that relies on the magnitude of values, such as distance measures used in k-nearest neighbors and in the preparation of coefficients in regression.

# When to use it?

Any algorithm that computes distance or assumes normality, **scale your features!!!**

It is hard to know whether rescaling your data will improve the performance of your algorithms before you apply them. It often can, but not always.

A good tip is to create rescaled copies of your dataset and test the performance.

**Tree based models** are not distance based models and can handle varying ranges of features. Hence, Scaling is not required while modelling trees.

**Naive Bayes** are by design equipped to handle the data as such so scaling is not required.