INTRODUCTION TO DATA TRANSFORMATION I

Data transformation techniques converts given data into a new, standard format. The format effects only the values of features not anything else. Data transformation is done for following two main reasons:

1. It removes the domination caused by the set of features present in the data set. After data transformation all features comes on same level irrespective of their values.
2. The transformed data improves the end result of the Machine learning models.

x_1	x_2
23	123.89
45	134.23
12	137.10
45	165.67
67	166.78

Figure 16: x_2 is dominating feature over x_1

x_1	x_2
-0.71	-1.1
0.30	-0.5
-1.23	-0.4
0.30	1.03
1.33	1.08

Figure 17: x_1, x_2 after data transformation

INTRODUCTION TO DATA TRANSFORMATION II

Data Transformation Techniques

There are basically two types of transformation techniques (meant only for quantitative variables).

1. Standardisation:

- It uses mean and standard deviation of the feature to scale down. The resulting value lies in range of $[-\infty, +\infty]$.
- **The most common technique of standardisation is Z-score.** This technique scales down the feature to follow standard normal distribution (mean zero and standard deviation of one).
- In Figure 18, process of standardisation is illustrated. The feature x_1 is inputted with some $\mu = 3.0$ and $\sigma = 0.8$ and, after standardisation, we obtain scaled values of x_1 with $\mu = 0$ and $\sigma = 1$.
- Standardisation is useful in situations where we need to use data set on proximity based-algorithms such as, k-means and k-nearest neighbour etc.

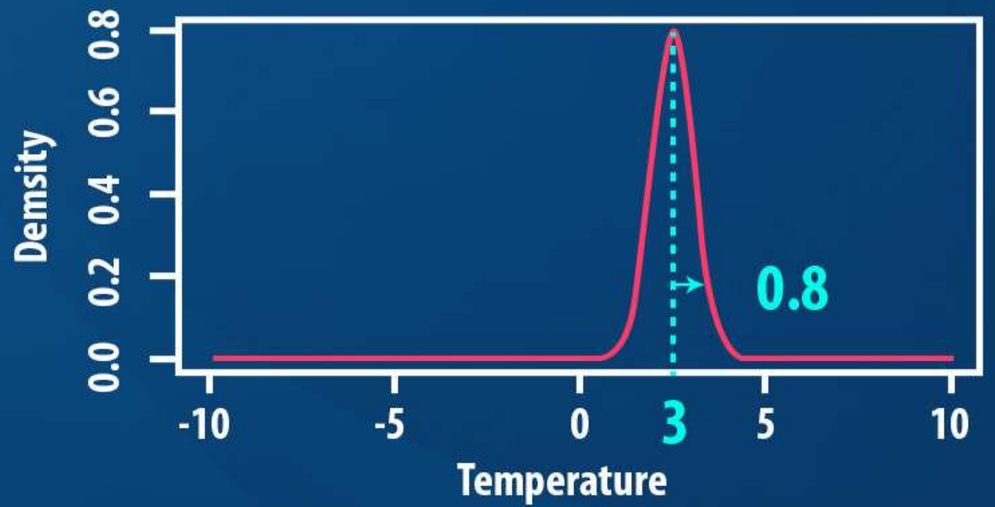
Figure 18: Process of Standardization

INTRODUCTION TO DATA TRANSFORMATION III

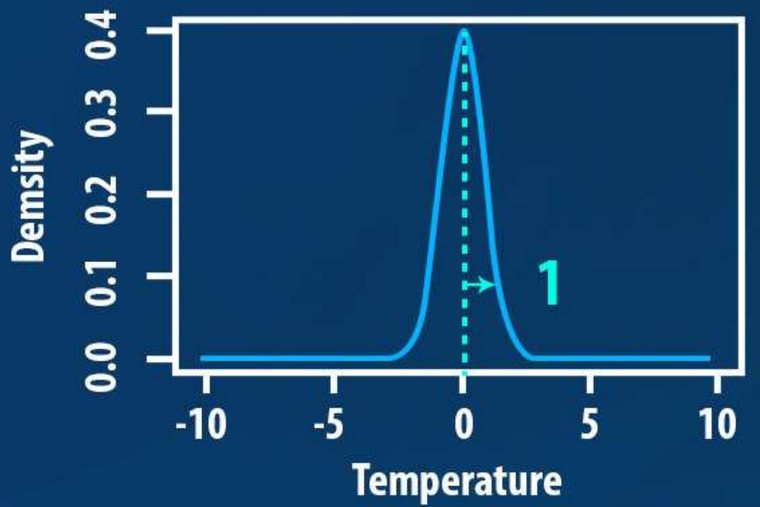
Quantitative variables can also be represented via a normal distribution plot.

In Figure 19a, feature variable "Temperature" is represented by normal distribution that captures μ and σ of the variable to represent this plot. The peak of the plot always indicates mean of the variable whereas, wideness of plot represents the standard deviation.

Figure 19b, represents the same variable "Temperature" but with standardisation technique applied on it. The change in the plot is observed with $\mu=0$ and $\sigma=1$.



(a) Example of normal distribution with $\mu=3$ and $\sigma=0.8$



(b) Example of standard normal distribution with $\mu=0$ and $\sigma=1$

Figure 19: Understanding normal and standard normal distribution plots

INTRODUCTION TO DATA TRANSFORMATION IV

Characteristics of Z-score

1. The Z-score can be positive or negative.
2. The distribution of converted data is standard normal distribution.
3. It signifies if a data point is below or above mean.
4. It scales down the feature by making it unitless.

Where to use Z-score?

1. It is useful in situations where we need to apply the data set on proximity based-algorithms such as, k-means, k-nearest neighbour etc.

INTRODUCTION TO DATA TRANSFORMATION V

2. Normalisation:

Normalisation is a data transformation technique that helps scaling down value of a feature.

However, unlike standardisation, normalisation does not work on mean and standard deviation of a given feature. Rather it takes into account minimum and maximum values taken by the attribute in order to scale down. This technique result in the range of [0, 1]. The normalised feature is obtained using Equation 5.

$$x^j = \frac{x_i - Min\ (x)}{Max(x) - Min(x)}$$

As shown in Figure 20, minimum and maximum value of x_1 is 11.06 and 58.67, respectively before Normalisation. However, after Normalisation, the value of x_1 scales down to minimum and maximum value 0 and 1, respectively.

Figure 20: Process of Normalization

INTRODUCTION TO DATA TRANSFORMATION VI

Characteristics of Min-Max normalisation

1. It scales down the feature to lie in range of [0-1].
2. Min-Max normalisation can result in data loss.
3. The Min-Max transformation will result only in positive values.
4. The distribution of converted data is a normal distribution.

Where to use Min-Max normalisation?

- It is useful in situations where we need data set to lie in a fixed range before feeding into the Machine learning algorithms.
- For example, image processing algorithms requires input pixels to fall in a definite range.