



# INTRODUCTION TO DATA ANALYTICS

***Class #7***

**Data Pre-processing – Data Transformation**

**Dr. Sreeja S R**

*Assistant Professor*

**Indian Institute of Information Technology  
IIIT Sri City**

## QUOTE OF THE DAY..

Nothing great was ever achieved without enthusiasm.

- RALPH WALDO EMERSON, American philosopher

# DATA TRANSFORMATION

# DATA TRANSFORMATION

Data transformation are needed to bring uniformity to the data. In addition, it can be used to scale the data to a preferred range. Following are few methods as data transformation.

- Change of origin
- Change of scale
- Change of origin and scale
- Decimal scaling
- Min-Max normalization
- Standard normalization

# DATA TRANSFORMATION

## 1. Change of Origin:

- Arbitrarily choose a constant. If sample values are integers, an integer constant is preferred.
- Shift data point by subtracting (or adding) the chosen constant from each sample observation.
- This technique is useful when data values are large, and variability is not so large.

### Lemma 6.1:

- a) The range of original data is preserved by a change of origin transformation.
- b) If  $\bar{x}$  is the old mean and  $c$  is the chosen constant, then the new mean of the transformed data is  $\bar{x}' = c + \bar{x}$

# DATA TRANSFORMATION

## Example:

Apply change of origin method to the following data and calculate the old and new means.

$$x = \{115, 128, 110, 104, 133\}$$

Let us arbitrarily chose the constant 120 and subtract this from each value in  $x$  to get the transformed data  $x'$  as

$$x' = \{-5, 8, -10, -16, 13\}$$

The new mean  $\bar{x}' = \frac{\sum x'_i}{5} = -2$

The mean of the original data is

$$\bar{x} = c + \bar{x}' = 120 - 2 = 118$$

# DATA TRANSFORMATION

## 2. Change of scale:

- This method is used to shorten the range of large numbers or lengthen the range of very small numbers.
- Chose (arbitrarily) a non-zero constant  $c$ . If  $c$  is less than the minimum of the observation, each value will be transformed to a value greater than 1. On the other hand, if it is greater than the maximum of the observation, then each value will be transformed to a value less than 1.
- If the value is between min and max of the sample, then the transformed values lie on the real line (positive real line if all  $x_i$ 's are positive)
- If all values are small fractions, we may multiply a large constant to scale them up and vice-versa.

# DATA TRANSFORMATION

## 3. Change of origin and scale:

- This is the most frequently used method to standardize data values. Depending upon the constants used to change the origin and scale, a variety of transformation intervals can be obtained.

**Example:** A sample in the range  $(a, b)$  can be transformed to a new interval  $(c, d)$  by the following transformation.

Let  $x$  is the original and  $y$  is the transformed value. Then

$$y = c + \frac{(d - c)}{(b - a)} \times (x - a)$$

[Prove that all values in the range  $(a, b)$  are mapped onto the range  $(c, d)$ ].



# DATA TRANSFORMATION

## 4. Min-Max Normalization:

- Min-Max normalization performs a linear transformation on the original data.
- Suppose,  $min_A$  and  $max_A$  are the minimum and maximum values of an attribute  $A$ . Min-Max normalization maps a value,  $v$  of  $A$  to  $v'$  in the range  $min'_A$  and  $max'_A$  using the following transformation:

$$v' = \frac{(v - min_A)}{(max_A - min_A)} \times (max'_A - min'_A) + min'_A$$

If  $[min'_A, max'_A] = [0,1]$ , then it is a special case of Min-max normalization.

# DATA TRANSFORMATION

## 5. Standard Normalization:

- This transformation is so called because it is extensively used in statistics in standardizing normal scores.
- Here, the origin is changed using the mean of the sample, and the scale is changed using the standard deviation of the sample.

$$v' = \frac{(v - \bar{A})}{\sigma_A}$$

where  $\bar{A}$  and  $\sigma_A$  are the mean and std for the attribute A.

- This method is also alternatively termed as **z-score normalization (zero-mean normalization)** and the transformed values are called z-scores.

# DATA TRANSFORMATION

## 5. Standard Normalization:

**Example:** Given  $X = \{32, 80, 56, 75, 69, 26, 44, 50\}$ . Apply the standard normalization.

Here, the mean  $\bar{A} = 54$ ,  $\sigma_A^2 = 390$ . Thus the z-scores are

$$v' = \frac{(v - \bar{A})}{\sigma_A} = \{-1.1140133 \quad 1.3165612 \quad 0.1012739 \quad 1.0633763 \\ 0.7595545 \quad -1.4178351 \quad -0.5063697 \quad -0.2025479\}$$

**Note:** It is very interesting to note that z-scores will always lie in the interval  $[-3, +3]$

## 6. Decimal Scaling:

This data transformation method is same as the “change of scale” method by either scale up or scale down.

$$v' = \frac{v}{10^j}$$

Here  $j$  represents movement of decimal points. Decided based on the maximum value in the data.

# DATA TRANSFORMATION

## Non-linear transformation

All the transformation methods discussed above are called linear transformation methods. Linear data transformation methods may be insufficient in some data missing applications. The popular non-linear transformations are:

- Square-root transformation
- Logarithmic and exponential transformation
- Hyperbolic transformation
- Trigonometric transformation
- Power transformation
- Polynomial transformation, etc.

These transformations are used either to stabilize the variance or bring the data into one of the well-known distributional form.