Data Transformation for Machine Learning

Abstract

Data transformation is a crucial part of machine learning since an uncurated data would affect the performance of a machine learning algorithm (model). In this paper we present the statistical properties of untransformed data with a transformation approach.

1 Introduction

1.1 SKEW

Skewness is asymmetry in a probability distribution of a data, in which the histogram curve appears distorted or skewed either to the left or to the right of the mean. The value of a skew can either be positive or negative.

- Normal Skew
  A distribution is normally skewed if the mass of the distribution is concentrated at the center. This type of frequency distribution are said to have *zero* skewness since the measure of central tendency are all equal.

$$mean = median = mode \tag{1}$$

- Positive Skew
  A distribution is positively skewed if the mass of the distribution is concentrated on the left or right-skewed. It can also be referred to as a right-tailed distribution where the mean and median exceeds the mode.

$$mean > median > mode \tag{2}$$

- Negative Skew
  A distribution is negatively skewed if the mass of the distribution is concentrated on the right or left-skewed. It can also be referred to as a left-tailed distribution where the mean and median are less than the mode.

$$mean < median < mode \tag{3}$$

1.1.1 MEASURE OF SKEWNESS

The direction and extent of skewness can be measured in various ways. Skew is the **third moment** of a data and the Fisher-Pearson coefficient for univariant data X_1, X_2, ..., X_N is given by

$$S_{FP} = g_1 = \sum_{i=1}^{N} \frac{(X_i - \hat{X})^3}{N\sigma^3} \tag{4}$$

Adjusted Fisher-Pearson coefficient used by statistical packages is given by

$$S_{FP} = G_1 = \frac{\sqrt{N(N-1)}}{N-2} \sum_{i=1}^{N} \frac{(X_i - \hat{X})^3}{N\sigma^3} \tag{5}$$

N: number of samples, $\hat{X}$: Mean, σ: Standard deviation

1.1.2 Karl Pearson's Measure of Skewness

The Karl Pearson's measure of skewness is based upon the divergence of mean from mob in a skewed distribution. A relative measure, independent of the units of measurement, is defined as the Karl Pearson's Coefficient of Skewness S_k, given by

- Mode Skewness

$$S_k = \frac{Mean - Median}{\sigma} \tag{6}$$

- Median skewness

$$S_k = \frac{3 * (Mean - Median)}{\sigma} \tag{7}$$

σ: Standard deviation

1.1.3 Bowley's Measure of Skewness

Bowley's method of skewness is based on the values of median, lower and upper quartiles. This method is used in case of **open-end series**, where the importance of extreme values is ignored.

Bowley's coefficient S_B, is given by

$$S_B = \frac{(Q_3 - Median) - (Median - Q_1)}{(Q_3 - Median) + (Median - Q_1)} \tag{8}$$

$$= \frac{Q_3 - 2Median + Q_1}{Q_3 - Q_1} \tag{9}$$

Q_3: $Upper/third\ quartile$, Q_1: $Lower/first\ quartile$

1.1.4 Distance Skewness

$$d_{Skew} = 1 - \frac{\sum_{i=1}^{N} ||X_i - \hat{X}||}{\sum_{i=1}^{N} ||X_i + \hat{X} - 2\theta||} \tag{10}$$

$\hat{X}$: Mean, θ: location parameter

1.1.5 Kelly's Measure of Skewness

Kelly suggests a measure based on P_{10} (10th Percentile) and, P_{90} (90th Percentile) such that only 10% of the observations on each extreme are ignored.

Kelly's coefficient S_P, is given by

$$S_P = \frac{(P_{90} - Median) - (Median - P_{10})}{(P_{90} - Median) + (Median - P_{10})} \tag{11}$$

$$= \frac{P_{90} - 2Median + P_{10}}{P_{90} - P_{10}} \tag{12}$$

1.1.6 Test Statistics

measures how many standard errors separate the sample skewness from zero.

Test statistics (Z_{g1}) is given by:

$$Z_{g1} = \frac{G_1}{SES} \tag{13}$$

Where Standard Error of Skewness (SES) is given by

$$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{14}$$

- if Z_{g1} is less than -2, the data is skewed negatively.
- if Z_{g1} is between -2 and $+2$, the data might be symmetric or skewed to either direction.
- if Z_{g1} is greater than 2, the data is skewed positively.

1.2 KURTOSIS

Kurtosis is the measure of peakness or flatness of a data distribution. **High kurtosis** in a dataset is an indication of the presence of outliers, **Low kurtosis** is an indication of lack of outliers.

1.2.1 MEASURE OF KURTOSIS

Kurtosis is the **fourth moment** of a data and is defined by

$$K = g_2 = \sum_{i=1}^{N} \frac{(X_i - \hat{X})^4}{Ns.d^4} \tag{15}$$

Excess kurtosis is measured using

$$K_{exc} = g_2 - 3 = \sum_{i=1}^{N} \frac{(X_i - \hat{X})^4}{Ns.d^4} - 3 \tag{16}$$

Sample excess kurtosis used by most statistical packages is given by

$$G_2 = g_2 \frac{(n-1)[(n+1)+6]}{(n-2)(n-3)} \tag{17}$$

2 Tranformation

Efficiently maximizing the performance of a machine learning algorithm requires well transformed data, hence the need for transforming an inbalanced dataset.

2.1 Continous data

2.1.1 Square Root Transformation

Given a feature vector $X \in R^d$, the square root transformation is given by

$$X_{new} = \sqrt{X} \tag{18}$$

2.1.2 Reciprocal Transformation

Given a feature vector $x \in R^d$, the reciprocal transformation is given by

$$X_{new} = 1/X \tag{19}$$

2.1.3 Log Transformation

Given a feature vector $x \in R^d$, the log transformation is given by

$$X_{new} = log(X) \tag{20}$$

2.1.4 Box Cox Transformation

Given a feature vector $X \in R^d$, Box Cox transformation estimates the value lambda from -5 to 5 that maximizes the normality of the data using the equation below.

$$X_{new}(\lambda) = \begin{cases} \frac{X^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ log(X), & \lambda = 0 \end{cases} \tag{21}$$

2.2 Categorical data

Class imbalance is common in real-world datasets. For example, a dataset with examples of credit card fraud will often have exponentially more records of non-fraudulent activity than those of fraudulent cases. Training a model on such dataset is likely to result in a bias towards the majority class. This may impact our hypothesis if the initial goal is to favor the minority class.

One way of solving this problem is by transforming or rebalancing the classes. This can be done using the method listed below.

2.2.1 Random Oversampling

2.2.2 Random Undersampling

2.2.3 Synthetic Minority Over-Sampling Technique (SMOTE)