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). A lot of emphasis is placed on the significance of transformation [6]. In this paper we present the statistical properties of untransformed data with a transformation approach.

1 Introduction

1.1 SCALING

Scaling is a necessary step that precedes modeling. This is essentially because real-world datasets always come in different scales. For instance, a housing dataset include features such as size of room (in square feet), price (in USD) or age of house (in years). Scaling the dataset helps reduce the spread or difference between this features.

1.1.1 Standardisation

Standardisation is useful for comparing variables expressed in different units. The standardised values are often between 0 and 1 and unitless.

standardised value

$$z = \frac{x - \bar{x}}{\sigma} \tag{1}$$

where $\bar{x}$ is the mean of feature vector x and σ is the standard deviation.

1.1.2 Normalization

Normalization is also a useful feature scaling process, rescaling the range of features to scale [0, 1] or [-1, 1].

Normalizaed value

$$z = \frac{x - min(x)}{max(x) - min(x)} \tag{2}$$

This is often referred to as $min - max$ normalization.

Mean normalization is a variant of normalization given by

$$z = \frac{x - mean(x)}{max(x) - min(x)} \tag{3}$$

Features can also be scaled to arbitrary numbers [a, b] using

$$z = a + \frac{x - min(x)}{max(a) - min(x)}(b - a) \tag{4}$$

1.2 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.

$$mean = median = mode \tag{5}$$

- 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.

$$mean > median > mode \tag{6}$$

- 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.

$$mean < median < mode \tag{7}$$

1.2.1 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

$$S_{FP} = g_1 = \sum_{i=1}^{N} \frac{(X_i - \bar{X})^3}{N\sigma^3} \tag{8}$$

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

$$S_{FP} = G_1 = \frac{\sqrt{N(N-1)}}{N-2} \sum_{i=1}^{N} \frac{(X_i - \bar{X})^3}{N\sigma^3} \tag{9}$$

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

1.2.2 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

$$S_k = \frac{Mean - Median}{\sigma} \tag{10}$$

- Median skewness

$$S_k = \frac{3 * (Mean - Median)}{\sigma} \tag{11}$$

σ : Standard deviation

1.2.3 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

$$S_B = \frac{(Q_3 - Median) - (Median - Q_1)}{(Q_3 - Median) + (Median - Q_1)} \tag{12}$$

$$= \frac{Q_3 - 2Median + Q_1}{Q_3 - Q_1} \tag{13}$$

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

1.2.4 Distance Skewness

$$d_{Skew} = 1 - \frac{\sum_{i=1}^{N} ||X_i - \bar{X}||}{\sum_{i=1}^{N} ||X_i + \bar{X} - 2\theta||} \tag{14}$$

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

1.2.5 Kelly's Measure of Skewness

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

Kelly's coefficient S_P, is given by

$$S_P = \frac{(P_{90} - Median) - (Median - P_{10})}{(P_{90} - Median) + (Median - P_{10})} \tag{15}$$

$$= \frac{P_{90} - 2Median + P_{10}}{P_{90} - P_{10}} \tag{16}$$

1.2.6 Test Statistics

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

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

$$Z_{g1} = \frac{G_1}{SES} \tag{17}$$

Where Standard Error of Skewness (SES) is given by

$$SES = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \tag{18}$$

- 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.

1.3 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.

1.3.1 MEASURE OF KURTOSIS

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

$$K = g_2 = \sum_{i=1}^{N} \frac{(X_i - \bar{X})^4}{Ns.d^4} \tag{19}$$

Excess kurtosis is measured using

$$K_{exc} = g_2 - 3 = \sum_{i=1}^{N} \frac{(X_i - \bar{X})^4}{Ns.d^4} - 3 \tag{20}$$

Sample excess kurtosis used by most statistical packages is given by

$$G_2 = g_2 \frac{(n-1)[(n+1)+6]}{(n-2)(n-3)} \tag{21}$$

2 TRANSFORMATION

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

$$x_i = x^{1/2} \tag{22}$$

2.1.2 Cube Root Transformation

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

$$x_i = x^{1/3} \tag{23}$$

Note that where negative values are present in x, equation above does not hold hence,

$$x_i = sign(x) * \sqrt[3]{abs(x)} \tag{24}$$

2.1.3 Square Transformation

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

$$x_i = x^2 \tag{25}$$

2.1.4 Reciprocal Transformation

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

$$x_i = 1/x \tag{26}$$

2.1.5 Log Transformation

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

$$x_i = log(x) \tag{27}$$

2.1.6 Power Transformation

Given a feature vector $x \in R^d$ where $x_i > 0$, The power transformation [2] is given by

$$x_i(\lambda) = \begin{cases} \frac{x_i^{\lambda-1}}{\lambda(GM(x))^{\lambda-1}}, & \lambda \neq 0 \\ GM(x) ln x_i, & \lambda = 0 \end{cases} \tag{28}$$

Where $GM(x_i) = (\Pi_{i=1}^{N} x_i)^{1/n}$ is the geometric mean.

2.1.7 Box Cox Transformation

Given a feature vector $x \in R^d$, Box Cox transformation [1] estimates the value lambda from -5 to 5 that maximizes the normality of the data using the equation below.

$$x_i(\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ log(x), & \lambda = 0 \end{cases} \tag{29}$$

2.2 Proportion/Percentage data

Data that are proportions (between 0 and 1) or percents (between 0 and 100) often benefit from special transformations.

2.2.1 logistic Transformation

logistic (logit) transformation handles very small and large values symmetrically, pulling out the tails and pulling in the middle around 50%. It is calculated thus,

$$logit(p) = log(\frac{p}{1-p}) = log(p) - log(1-p) \tag{30}$$

p : proportion or percent

2.2.2 Angular Transformation

the angle whose sine is the square root of p. Expressed as,

$$p = \arcsin(p^{1/2}) \tag{31}$$

In practice this angle behaves similar to

$$p = p^{.4} - (1-p)^{.4} \tag{32}$$

p : proportion or percent

2.3 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 transforming or rebalancing the classes. This can be done using the following methods

2.3.1 Random Oversampling

Random oversampling increases the weight of the minority class by replicating the minority class examples. The effect of this on a model is increasing the likelihood of overfitting.

2.3.2 Random Undersampling

Random undersampling decreases the weight of the majority class to create a balanced dataset. This approach tends to reduce the overfitting effect.

2.3.3 Synthetic Minority Over-Sampling Technique (SMOTE)

The minority class is over-sampled by taking each minority class sample and introducing synthetic examples along the line segments joining any or all of the k minority class nearest neighbors . Depending upon the amount of over-sampling required, neighbors from the k nearest neighbors are randomly chosen. The **SMOTE** algorithm is described in the authors publication [9].

References

[1] Box, G.E.P. and D.R. Cox. An analysis of transformations. *Journal of the Royal Statistical Society B 26: 211-252*, 1964.

[2] Carroll, RJ; Ruppert, D. On prediction and the power transformation family. *Biometrika. 68: 609–615*, 1981.

[3] Emerson, J.D. Mathematical aspects of transformation. In Hoaglin, D.C., F. Mosteller and J.W. *Understanding Robust and Exploratory Data Analysis.* New York: John Wiley, 247-282. 1983.

[4] John, J.A. and N.R. Draper. Quantifying Radiographic Knee Osteoarthritis Severity using Deep Convolutional Neural Networks *An alternative family of transformations* Applied Statistics 29: 190-197. 1980.

[5] Johnson, N.L. Systems of frequency curves generated by methods of translation. *Biometrika 36: 149-176.* Biometrika 36: 149-176. 1980.

[6] Tukey, J.W. *On the comparative anatomy of transformations* Biometrika 36: 149-176. 1957.

[7] Tukey, J.W. *The practical relationship between the common transformations of percentages or fractions and of amounts* Pacific Grove, CA: Wadsworth & Brooks-Cole, 211-219. 1960.

[8] Yeo, I. and R.A. Johnson. *A new family of power transformations to improve normality or symmetry* Biometrika 87: 954-959. 2000.

[9] Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, W. Philip Kegelmeyer. SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research 16 321–357* 2002.