Data Normalization and Standardization: A Technical Report

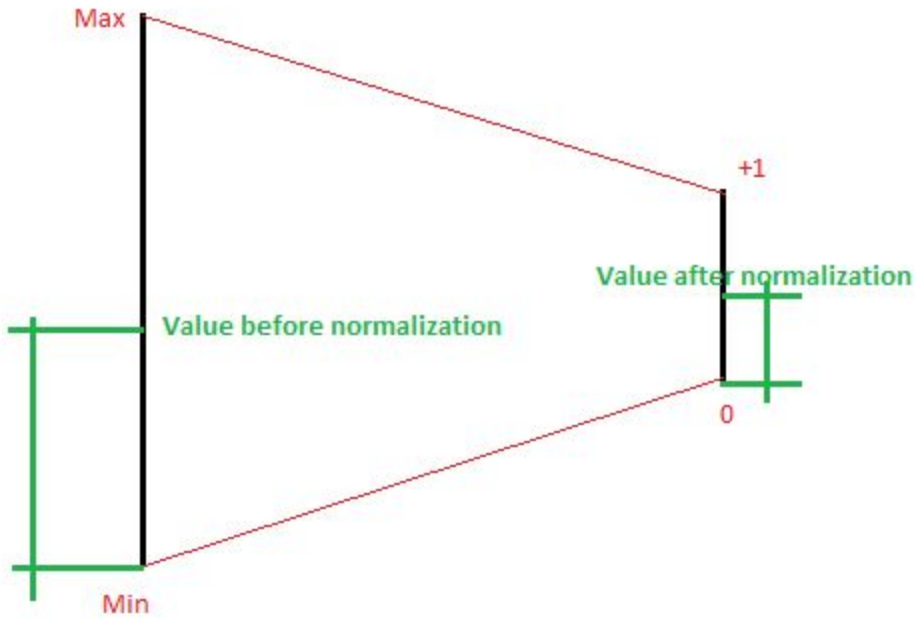
Peshawa Jamal Muhammad Ali*, and Rezhna Hassan Faraj
The Machine Learning Lab. at Koya University
Koya, Erbil, Iraq.
peshawa.jammal@koyauniversity.org

Cite the technical report:
Peshawa J. Muhammad Ali, Rezhna H. Faraj; "*Data Normalization and Standardization: A Technical Report*", Machine Learning Technical Reports, 2014, 1(1), pp 1-6.
https://docs.google.com/document/d/1x0A1nUz1WWtMCZb5oVzF0SVMY7a_58KQulqQVT8LaVA/edit#

Information about the publisher:
Machine Learning Technical Reports is a periodical technical report published by the Machine Learning Lab. at Koya University
Koya University, Building of Faculty of Engineering, KOY45
Koya 44023, Erbil, F.R. of Iraq
Contact: +9647707578801/+9647501138655, email: feng.dswe@koyauniversity.org

Abstract
This paper aims to clarify how and why data are normalized or standardized, these two processes are used in the data preprocessing stage in which the data is prepared to be processed later by one of the data mining and machine learning techniques like support vector machine, neural network, etc. The two methods try to scale the data set. These two processes are helpful in some cases and necessary in some other cases, most of the data mining and machine learning tools include these two preprocessing techniques like in Weka or in Matlab. This paper will simply define and present the use of these two data preprocessing techniques.

Normalization
It's the process of casting the data to the specific range, like between 0 and 1 or between -1 and +1. Normalization is required when there are big differences in the ranges of different features. This scaling method is useful when the data set does not contain outliers. The theoretical background of normalization can be easily understood from Figure (1). If it is required to cast the data to the range 0,1 then:



From Trigonometry:

$$\frac{valueAfterNormalization - 0}{1 - 0} = \frac{valueBeforeNormalization - min}{max - min}$$

$$\frac{valueAfterNormalization}{1} = \frac{valueBeforeNormalization - min}{max - min}$$

$$valueAfterNormalization = \frac{valueBeforeNormalization - min}{max - min}$$

$$or \quad x' = \frac{x - min}{max - min}$$

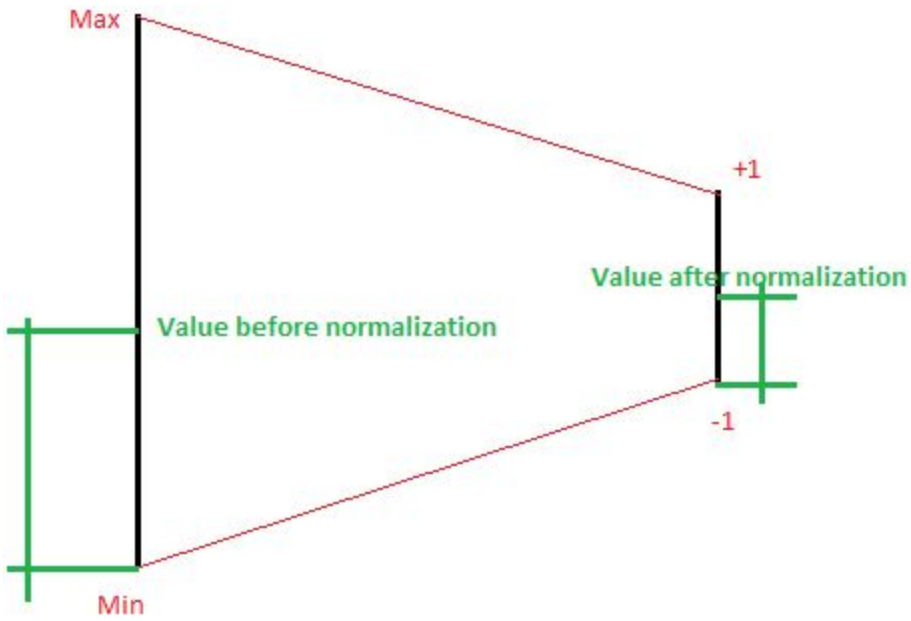
Denormalization

This process should be done if normalization applied. For example, to denormalize the a data from the range 0, 1 below equation can be used:

$$x = [x' * (max - min)] + min$$

where x' is the normalized data and x is denormalized data, min and max are the same values used previously in the normalization process.

To normalize the data to the range -1, +1 see Fig(2):



$$\frac{valueAfterNormalization - (-1)}{1 - (-1)} = \frac{valueBeforeNormalization - min}{max - min}$$

$$\frac{valueAfterNormalization + 1)}{2} = \frac{valueBeforeNormalization - min}{max - min}$$

$$valueAfterNormalization = 2 * (\frac{valueBeforeNormalization - min}{max - min}) - 1$$

$$or \qquad x' = 2 * (\frac{x - min}{max - min}) - 1$$

Denormalization from range -1, +1

$$x = [\ (\frac{x' + 1}{2})(max - min)\ ] + \ min$$

In WEKA, for the range -1,+1, the formula is organized as follow:

$$x' = 2 * (\frac{x - min}{max - min}) - 1$$

$$x' = (\frac{x - min}{\frac{max - min}{2}}) - 1 = [\frac{x - min - (\frac{max - min}{2})}{\frac{max - min}{2}}]$$

$$x' = [\frac{x - min - \frac{max}{2} + \frac{min}{2})}{\frac{max - min}{2}}] = [\frac{x - \frac{max}{2} - \frac{min}{2})}{\frac{max - min}{2}}]$$

$$x' \ = [\frac{x - (\frac{max}{2} + \frac{min}{2})}{\frac{max - min}{2}}]$$

$$x' = \frac{x - (\frac{max + min}{2})}{\frac{max - min}{2}}$$

Z-score standardization

Making a data set with mean=0, and standard deviation =1. This scaling method is useful when the data follows a normal distribution (Gaussian distribution), if the data does not follow normal distribution then this will make problems.

Example: -20, -6, 0, 40, 70,120

$$Mean = \frac{-20-6+0+40+70+120}{6} = 34$$

$$sd = \sqrt{\frac{(-20-34)^2+(-6-34)^2+(0-34)^2+(40-34)^2+(70-34)^2+(120-34)^2}{6}}$$

sd = 48.98979

z-score standardization

$$x'' = \frac{x-mean}{sd} = \frac{-20-34}{48.98979} = -1.1022$$

Other values are changed too,

Accordingly, values are changed to:

-1.10227

-0.8165

-0.69402

0.122474

0.734847

1.755468

Now, if you calculate the average and sd of these new values you will see that the mean is zero and sd=1.

Important note:

However, the point must be made that N/S are _not_ good where the raw measurement is desirable and where the N/S is irreversible, thus losing much of the information in the raw measurement, this is according to a note made by Kevin Hankins (kevin.hankins@dbhds.virginia.gov).

References

1. Yazen A. Khalil and Peshawa J. Muhammad Ali; "A proposed method for colorizing grayscale images", International Journal of Computer, Science and Engineering, 2013, 2(2), pp.104-109. http://www.iaset.us/view_acrhives.php?year=2013&id=14&jtype=2&page=2
2. Peshawa J. Muhammad Ali, Nigar M.S. Suramerry, Abdul-rahman M. Yunis, Ladeh S.Abdulrahman, "Gender prediction of journalists from writing style", Aro Journal, 2013, 1(1), pp.22-28. http://aro.koyauniversity.org/issues/volumeone/aro-10031
3. Peshawa J. Muhammad Ali; "Predicting the gender of the Kurdish writers in Facebook" Sulaimani Journal for Engineering Sciences, 2013, 1(1), pp.18-28. http://www.univsul.edu.iq/Wenekan_KS/12111313102014_Sulaimani%20Journal-ENG.%2020-30.pdf