Data Normalization and Standardization

When approaching data for modeling, some standard procedures should be used to prepare the data for modeling:

1. First the data should be filtered, and any outliers removed from the data (watch for a future post on how to scrub your raw data removing only legitimate outliers).
2. The data should be normalized or standardized to bring all of the variables into proportion with one another. For example, if one variable is 100 times larger than another (on average), then your model may be better behaved if you normalize/standardize the two variables to be approximately equivalent. Technically though, whether normalized/standardized, the coefficients associated with each variable will scale appropriately to adjust for the disparity in the variable sizes. However, if normalized/standardized, then the coefficients will reflect meaningful relative activity between each variable (i.e., a positive coefficient will mean that the variable acts positively towards the objective function, and vice versa, plus a large coefficient versus a small coefficient will reflect the degree to which that variable influences the objective function. Whereas the coefficients from un-normalized/un-standardized data will reflect the positive/negative contribution towards the objective function, but will bill much more difficult to interpret in terms of their relative impact on the objective function.
3. Non-numeric qualitative data should be converted to numeric quantitative data, and normalized/standardized. For example, if a survey question asked an interviewee to select where the economy will be for the next six months (i.e., deep recession, moderate recession, mild recession, neutral, mild recovery, moderate recovery, or strong recovery), these can be converted to numerical values of 1 through 7, and thus quantified for the model.

So when we speak of data normalization and data standardization, what is meant? To normalize data, traditionally this means to fit the data within unity (1), so all data values will take on a value of 0 to 1. Since some models collapse at the value of zero, sometimes an arbitrary range of say 0.1 to 0.9 is chosen instead, but for this post I will assume a unity-based normalization. The following equation is what should be used to implement a unity-based normalization:

$$X_{i,\ 0\ to\ 1} = \frac{X_i - X_{Min}}{X_{Max} - X_{Min}}$$

Where:

X_i = Each data point i

X_{Min} = The minima among all the data points

X_{Max} = The maxima among all the data points

$X_{i,\ 0\ to\ 1}$ = The data point i normalized between 0 and 1

If you desire to have a more centralized set of normalized data, with zero being the central point, then the following equation can be used instead to normalize your data:

$$X_{i,\,-1\text{ to }1} = \frac{X_i - \left(\frac{X_{Max} + X_{Min}}{2}\right)}{\left(\frac{X_{Max} - X_{Min}}{2}\right)}$$

Where:

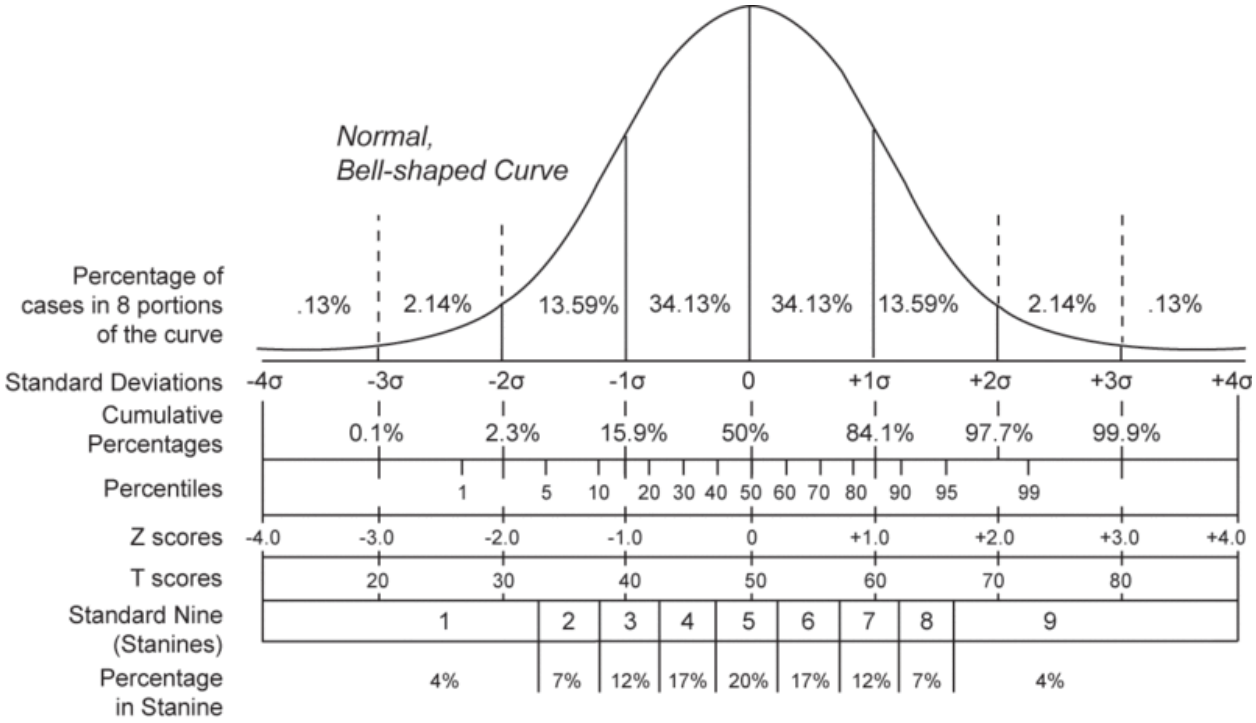
X_i = Each data point i

X_{Min} = The minima among all the data points

X_{Max} = The maxima among all the data points

$X_{i,\,-1\text{ to }1}$ = The data point i normalized between 0 and 1

Finally, to standardize your data, you will want the data to reflect how many standard deviations from the average that that data lies, with the following normal distribution curve representing the probability of each standard deviation for a normal distribution (this graphic is borrowed from Wikipedia). The Z-Score is what will be calculated to standardize the data, and it reflects how many standard deviations from the average that the data point falls.



To determine the Z-Score of each data point, the following equation should be used:

$$X_{i,\,1\sigma} = \frac{X_i - \bar{X}_S}{\sigma_{X,\,S}}$$

Where:

X_i = Each data point i

$\bar{X}_S$ = The average of all the sample data points

$\sigma_{X,\,S}$ = The sample standard deviation of all sample data points

$X_{i,\,1\sigma}$ = The data point i standardized to 1σ, also known as Z-Score

The question now arises when should each of these techniques be used, and why. Please refer to the following spreadsheet:

Data Norm & Stand.xlsx

A thousand randomly generated dice rolls were simulated in Excel using the following formula *=INT(RAND()*6+1)*. This data is located in Column C of the Analysis worksheet. A second die was then simulated (Column D), and the two die were then averaged to obtain a normal distribution (Column E). Finally, column A was multiplied by 100 to create a variable that is scaled 100 times higher than another (Column F). The data for the single 6-face die (Column C), the averaged 6-face dice (Column E), and the 100× scaled 6-face die (Column F) were normalized between 0 to 1 and -1 to 1, and standardized to 1σ. The results were then plotted using histograms, which can be found in the Graphs tab.

What can be seen from the histograms is that it didn't matter whether the raw data was linear (single 6-face die) or in a distribution to begin with (average of two 6-face dice), the same distributions were obtained no matter whether normalization or standardization were applied. The true value of normalizing or standardizing the data though could be seen with the 100× scaled single 6-face die which produced identical normalized an standardized output as its un-scaled single 6-face die (i.e., the histograms are identical between these two).

The conclusion from this experiment is that data should be normalized or standardized to remove their scale from your modeling, but both techniques produce identical results to this desired outcome. However, standardizing is the preferred method because it produces meaningful information about each data point, and where it falls within its normal distribution, plus provides a crude indicator of outliers (i.e., anything above or below a Z-Score of ±4.