Overview

When we work with some collection of data, it is often comprised of a number of different types of features. Let's consider an example involving loan data. Let the data set $\underset{n\times(p+1)}{\mathbf{X}}$ represent historical loan data with as $p+1$ column vectors in the format of $\begin{bmatrix}\mathbf{1}_n & \mathbf{x}_1 & \mathbf{x}_2 & \cdots & \mathbf{x}_p\end{bmatrix}$.

Suppose that $\mathbf{y}$ represented the loan amount that was awarded. Which features do you think would be necessary for us to determine how much of a loan an individual can receive? Let's define our dataset with the following features:

1. $\mathbf{x}_1$ – annual income
2. $\mathbf{x}_2$ – number of years in current position
3. $\mathbf{x}_3$ – number of vehicles

Which of the above features would be larger compared to the others? If one feature is substantially larger than other features, then consider the following linear regression equation:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

If the values for x_1 are substantially larger than x_2 and x_3, but each of these features are important in determining y, then it may take longer for the weights to converge using techniques like gradient descent (e.g. linear regression, support vector machines, neural networks, etc.). This is due to the fact the magnitude of one covariate can easily eclipse that of another covariate. How do we rectify this problem? Simple: we need to normalize the data.

Data Normalization

With data normalization, we seek to reduce the range of each of the covariates. Consider the data from this example. We want to ensure that we represent $\mathbf{x}_1$, $\mathbf{x}_2$, and $\mathbf{x}_3$ on the same scale. Consider examinations that you have taken across classes. Suppose that one student received 343 points on some examination in one class, and 9 points in another class. How did this student perform? In short, you have absolutely no basis for comparison. If you were told the maximum number of points from the first class was 350, then that student achieved a score of $343/350 = 0.98$, which is outstanding. For the other examination score, the individual would have scored $9/20 = 0.45$, which is failing. From this example, you were already applying a form of data normalization where you scaled the data from both tests between the range of $[0, 1]$ to make the results easier to interpret.

There are many ways that you can scale data. Consider some of the following options where x is the input value and x' is the newly scaled value:

1. Log scaling: $x' = \log(x)$
2. Min-max scaling: $x' = \dfrac{x - \min(\mathbf{x})}{\max(\mathbf{x}) - \min(\mathbf{x})}$
3. z-score: $x' = \dfrac{x - \mu}{\sigma}$

For z-scores, the closer a x is to the mean of its respective covariate, the z-score value will be 0. The further the value the data point is away from its respective mean, then the larger the departure the z-score will be away from 0 noting that it can positive or negative.