Chapter 1

Standardizing Data

Data standardization is a preprocessing step in many machine learning algorithms. Standardization transforms the variables in the dataset to have a mean of zero and a standard deviation of one.

The standardization of a variable X is calculated as follows:

$$Z = \frac{X - \mu}{\sigma} \tag{1.1}$$

where:

- Z is the standardized variable.
- X is the original variable.
- μ is the mean of X.
- σ is the standard deviation of X.

1.1 Why Do We Standardize Data?

There are several reasons why standardization is essential:

1.1.1 Homogeneity of Variances

Some statistical techniques assume that all variables have the same variance. Standardizing the data ensures this assumption.

1.1.2 Interpreting Coefficients

In regression analysis, standardizing allows us to interpret the coefficients of the predictors as the change in the response variable associated with a one-standard-deviation increase in the predictor.

1.1.3 Algorithm Convergence

For many machine learning algorithms (like gradient descent), standardization can help the algorithm converge more quickly to the optimum.

1.1.4 Comparing Variables

Standardization puts different variables on the same scale, allowing for meaningful comparisons. For example, it would be challenging to compare a variable measured in kilograms with another measured in kilometers without standardization.

1.1.5 Preventing Numerical Instabilities

Standardizing can help prevent numerical instabilities in computations, particularly when dealing with high-dimensional data.

Remember, though standardization is useful and necessary in many situations, it's not always required. For instance, tree-based models are scale-invariant and don't require standardization.

This is a draft chapter from the Kontinua Project. Please see our website (https://kontinua.org/) for more details.

Appendix A

Answers to Exercises