Data Science 1b

raymond mugandiwa

July 2024

1 Data formatting and validation

Data formatting and validation can improve the quality of data analysis by filling missing values and adjusting the distribution of the data to follow a normal distribution. An example of this can be a dataset that is left skewed, the data scientist could use squared root transform or log transformer to fix the distribution of the data. This would help in analysing it and applying the data to a model. For some statistical methods the data needs to be formatted before applied. An example of this is that a log and square root transform can never take in a negative number as input. So in certain situations data validation needs to take place to check whether the data is positive, negative or if it even exists. There are a variety of ways to format data these include normalisation, standardisation, square root transform, Log Transform, and square transform. Normalisation is the process of scaling your data so it is between 0 and 1 this is done in order to lessen the influence of bigger numbers. Standardization is the process of the data having a mean of 0 and a standard deviation of 1. Log, square root and square transform is just applying their respectful functions onto every data point in the dataset to transform it. Data formatting can even be done in simpler terms such as having a consistent date format or a consistent name format like first name and surname for all the values rather than a weird mismatch.