COIS 4400H Assignment 1 Q3

Q3. Normalise Data

In order to understand the process of having new values added to the dataset without re-normalizing all the old values while also having a min-max values set, we should first understand how the normalization happens.

We normalize data to organize the data so that it appears similar across all the records and fields. When data's different features vary drastically from each other, then finding trends and useful information from mining this data using algorithms is difficult. Therefore, we try to normalize the data, more like min-max normalize the data so that it is ea`sy to extract useful information from the data. This helps us in analysis with multiple variables measured on different scales.

There is a simple formula to ensure that we normalize data and also ensure it is between 0 and 1.

$$\textit{Normalized value } z_i = [(x_i - min(x))(max_{new} - min_{new}) / (max(x) - min(x))] + min_{new}$$

Z_i = i^{th} normalized value of dataset

X_i = i^{th} value in the dataset

Min(x) = minimum value of dataset

Max(x) = maximum value of dataset

Min_{new} = 0

Max_{new} = 1

This is the formula for normalizing a dataset.

Normalization approach :

Whenever a new value is added to the dataset, its normalized value can be easily created using this above formula :

$$\textit{Normalized value } z_i = [(x_i - min(x))(max_{new} - min_{new}) / (max(x) - min(x))] + min_{new}$$

Not only will it be between 0 and 1 but also, there would be no need to re-normalize all the values. Thus satisfying both the conditions :

- all values (old and new) have to lie in the range between 0 and 1
- no transformation or renormalization of the old values is allowed

However, an issue arises when the new value is less than the minimum or greater than the maximum, then there would be a need to renormalize all the old values.

Therefore, we need to find a normalization approach that will make sure that the exception case of the new value being less than minimum or greater than maximum does not rise an issue. This is possible if the extreme values are in two separate groups. Meaning let's say the entire dataset to be normalized (in order) has 50 values. We divide it into three equal groups. The first group from minimum to let's say one

third of the dataset and similarly the last group would include the maximum part. Each group would have values 0 to 1.

This way even if the new value to be added is the lowest value, we know we need to add it to B1. Similarly the maximum value would go into B3. We need a function to identify the range it lies in by comparing the new value to the not-normalized values of the old dataset. For efficiency, we would need to keep in mind the Big-O notation and probably use algorithm like Binary search.

The entire dataset may not be 0 to 1, but all the values in overall are still from 0 to 1. For representation it could be represented as a boxplot. There would be a few errors in being precise but better than having to re-normalize all data.

References :

https://stats.stackexchange.com/questions/70801/how-to-normalize-data-to-0-1-range

https://www.kristakingmath.com/blog/effect-on-mean-median-mode-by-changing-the-data