← Go Back to Making Sense of Unstructured Data

Course Content

Distance and Scaling Measures

Unsupervised learning algorithms use different distance measures or similarity/dissimilarity measures between each pair of observations to group data into different clusters. Points which are close to each other on that distance metric are likely to be grouped into the same cluster, while points far apart on that distance metric are likely to be members of different clusters. All this is done by computing a distance matrix that has distances between every pair of observations. There are different ways of calculating these distances, some of which are mentioned below:

Euclidean distance: Euclidean distance is calculated as the square root of the sum of the squared differences between two vectors.

Say there are two points P (x_1, y_1) and Q (x_2, y_2). The Euclidean distance between these two points would be calculated as:

$$d = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Manhattan distance: Also called city block distance, it calculates the distance between two points by drawing an orthogonal, zig-zag grid between them.

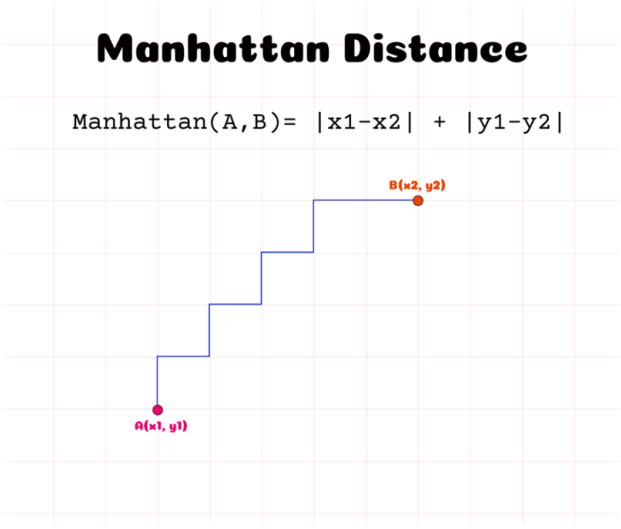

Image Source

Scaling:

Input variables in a dataset may have different units e.g. kilometers, hours, kilograms, etc. i.e. different scales. This difference in scales increases the difficulty in modeling, in turn resulting in an unstable model. Unless you

This difference in scales increases the difficulty in modeling, in turn resulting in an unstable model. Unless you scale the data, the importance of 1 km would be the same as 1 kg, which would be the same as 1 hr, the same as 1 cm, etc.

In other words, while 1000 gms and 1 kg mean the same thing, a quantity of 1000 ml from another feature is at a different scale as a number relative to 1000 g, than it is with respect to 1 kg. Unless both the weight (i.e. gms) and the volume (i.e. ml) are on the same scale, machine learning algorithms might give one of them more weightage than the other simply due to the number in that quantity. That in turn, diminishes the effect of the variable that has the smaller number (lower scale).

Thus, scaling the variables is an important step in machine learning models. By scaling the variables, we can ensure that the machine learning algorithm gives every variable a similar weightage in terms of its likelihood of contributing to the decision making of the algorithm, and that no single variable with a high numerical quantity as its value unnecessarily influences the algorithm's predictive power.

Normalization: One of the ways of scaling the data is so that all the values lie between 0 and 1. This is called Normalization. A value can be normalized as follows:

$$y = (x - min) / (max - min)$$

where,

- **y**: normalized version of the variable
- **x**: variable of interest
- **min**: minimum value of the variable x in this dataset
- **max**: maximum value of the variable x in this dataset

Standardization is another way of scaling the data, where the mean of the observations becomes 0 and the standard deviation is 1. A value can be standardized as follows:

$$y = (x - mean) / std\_dev$$

where,

- **y**: standardized version of the variable
- **x**: variable of interest
- **mean**: average (arithmetic mean) of the variable x in the dataset
- **std_dev**: standard deviation of the variable x in the dataset

‹ Previous

Next ›