Feature Scaling – Normalization

↳ Goal of Normalization is to change values in numeric columns in the data set to use common scale, w/o distorting differences in ranges of values or losing information

You can columns with different units of measurement Kg, Cm, Inch, Meter, grams, °C, °F etc all together.

We Ignore these measurement and try to bring all values's under a common Scale.

Techniques for Normalization

1. Min Max Scaling (used 90% of time)
2. Mean Normalization
3. Max Abs scaling
4. Robust Scaling.

Min max scaling

For Transforming value we use value

$$X_i' = \frac{X_i - X_{min}}{X_{max} - X_{min}}$$

X_i → value need to be transformed

X_{min} → min value in variable/column

X_{max} → max values of data

Transformed value

Now after all values are transformed, the Max value of these transformed value now is 1 and min value is 0.

Geometric Intution

So min max scaling squeeze the values under 1 unit of square/rectangle in 2D data.

1 unit of cube for 3D data. 1 unit of hyper cube for more dimensional data.

Remember the distribution remains same but after transformation the values gets squeezed under 1 unit of measurement.

Mean Normalization

Its kind of like standardization where data is centred towards mean

$$x_i' = \frac{x_i - x_{mean}}{x_{max} - x_{min}}$$

Gives value in the range of -1 & 1

Extremely rare to use.

People prefer standardization over this cause sciKit learn has no support of this

Max Absolute Scaling

used for sparse data ⇒ where we have lot's of zero's.

$$X_i^{o'} = \frac{X_i^o}{|X_{max}|} \rightarrow \text{Abs value}$$

Robust Scaling

$$X_i^{o'} = \frac{X_i^o - X_{median}}{IQR \quad (75^{th} \text{percentile} - 25^{th})}$$

RobustScalorClass

This thing is robust to "Outliers" because "median" doesn't gets effected by outliers like in case of mean.

The Robust Scalor can be utilized when we have lots of outliers in our data.

Normalization VS Standardization

90% we use this ⇐

1) Is Feature Scaling Required?
⇒ we need to understand what algo require feature scaling.

2) Most of time we get better results

with standardization.

3) Most Normalization (Min Max Scaler) is used where we before hand know the min & max values of data. Mostly used in CNN. (0-255) values.

Impact of Outlier on MinMax Scaling?
As value gets squeezed under 0 & 1,
the outlier value also now gets reduced under 1 & between 0.
Now this can effect the algorithm.

2)
Best to avoid this we can use Robust Scaling if we have lots of outliers in our data.