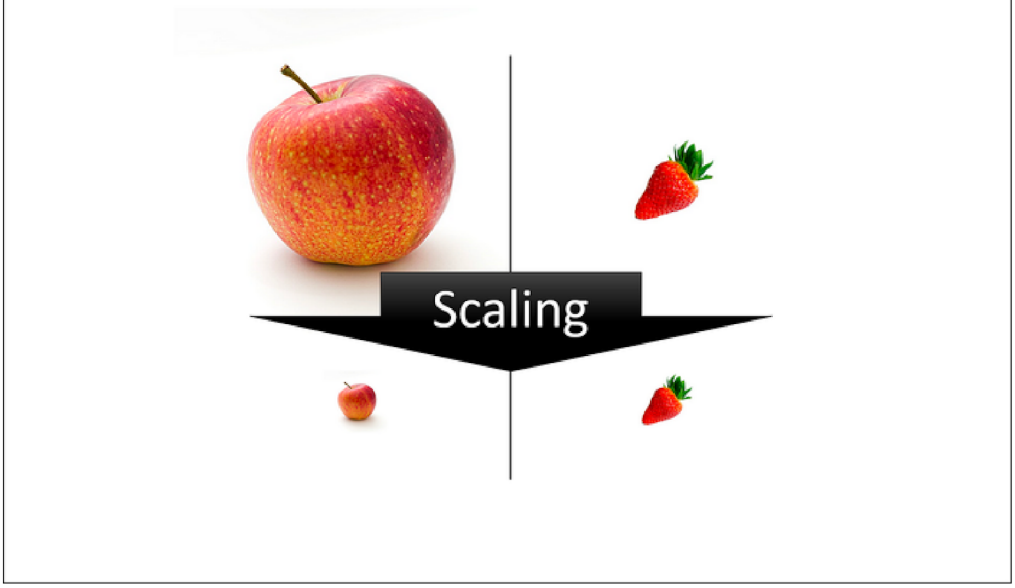
Day 5: Feature Scaling



Credit: Pixabay

Author: Alisha Metkari

What is Feature Scaling

- Feature scaling is the process of transforming data to the similar range.
- When dealing with data, there can be features having range from 0.1 to 0.9 and features having range from 1000 to 10000 or more than this. As you can see the scale is not the same.
- Using such features in ML model makes the model biased towards the features with large values.
- All features should contribute equally to the model and there should not be dominance of any feature.
- Feature scaling is a technique that applies the mathematical formula and transforms the features to the similar range.
- Tree based models are robust to the scale of features. Thus feature scaling is required to the models that are based on gradient descent or distance.
- Let’s see some common feature scaling methods.

Author: Alisha Metkari

Normalisation

- Normalisation also known as Min-Max scaling. It uses minimum and maximum values to transform the data.
- It transforms the data into the specified range, by default the range is 0 to 1.
- It does not reduce the effect of outliers but scales down all the values to the fixed range.
- Scikit-learn library has this implementation in the preprocessing module, MinMaxScaler.
- The formula is given below.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Author: Alisha Metkari

Standardisation

- Standardisation is the method in which all values are centred around the mean with unit standard deviation.
- It uses the standard deviation and mean of the data for transformation.
- The scaled values are not restricted to a particular range .
- It is used when the data has gaussian distribution.
- It does not reduce the effect of outliers.
- Scikit-learn library has this implementation in the preprocessing module, StandardScaler.
- The formula is given below.

$$x' = \frac{x - \bar{x}}{\sigma}$$

Author: Alisha Metkari

Robust Scaling

- Robust Scaling uses median and IQR(Inter-Quartile Range) to transform the data. It removes the median from data and scales according to the IQR.
- The statistics used here are insensitive to outliers hence it is robust to outliers and removes the effect of outliers.
- Scikit-learn library has this implementation in the preprocessing module, RobustScaler.
- The formula is given below.

$$X_{\text{scale}} = \frac{x_i - x_{\text{med}}}{x_{75} - x_{25}}$$

Author: Alisha Metkari

Advantages of Feature Scaling

- Improves the performance of the model.
- Speeds up the computational power, making the model learn fast and reach to the global minima.
- Removes the dominance of features with large values and make the features comparable.

Author: Alisha Metkari