INTERVIEW DAY → Scaling, Normalization & Standardization

Q) What is Feature Scaling / Scaling?

Ans - Feature Scaling is a technique to standardize the independent features present in the data in fixed range.

- It is mainly used to handle highly varying magnitudes or values or units.
- If the feature scaling is not done, then a machine learning algorithim tends to weigh greater values, higher and consider smaller values are lower, regardless of the unit of the value.

Q) Give an example where Feature Sealing should be used?

Ans - If an algorithim is not using feature scaling method then it can consider the value of 3000 meter to be greater than 5 km but that's actually not true and in this case, the algorithim will give wrong prediction. So we use Feature Scaling to bring all values to same magnitude.

Q) What is Normalization? Or what is Min-Max Scaler?

Ans - Normalization is a scaling technique in which values are shifted and rescale, so that they end up ranging between 0 and 1.

- It is known as Min-Max scaling.

Formula, $\boxed{x' = \frac{x - x_{min}}{x_{max} - x_{min}}}$

x_{max} → Maximum value of feature

x_{min} → Minimum value of feature

- When X is minimum, numerator will be 0 because $X = X_{min}$ = minimum value so $X - X_{min}$ equal $= 0$.

When X is maximum, numerator is equal to denominator, $X = X_{max}$.

$= \frac{x - X_{min}}{X_{max} - X_{max}}$ becomes $\frac{X_{max} - X_{min}}{X_{max} - X_{min}} = 1$. Hence range is set from 0 to 1.

Q) What is Standardization?

Ans - Standardization is another scaling techniques when the value are centered around mean with a unit standard deviation. This means that mean of the attributes become zero and the resultant distribution has a unit standard deviation.

Formula, $\boxed{x' = \frac{x - \mu}{\sigma}}$

- μ is the mean of the feature value.
- σ is standard deviation of the feature value.
- The value is not restricted to any range therefore not affected by outlier as it does not have strict range.

Q) When to use what? Scaling / Normalization / Standardization.

Ans - Scaling is used when we use distance based measures -

i) Scaling is critical while performing PCA.

ii) KNN uses Euclidean distance, measure is sensitive to magnitude and hence should be scaled for all feature to weigh in equally.

- Normalization is good to use when distribution of data does not follow normal distribution.
  This can be useful in algo that does not assume any distribution like KNN.
- Standardization can be useful when data follows Normal Distribution.

Q) Which algorithim do not need scaling?

Ans - Random forest / decision tree, as these algo relies always on some rules.