Dimensionality Reduction

What is Dimensionality Reduction?

- The number of input features, variables, or columns present in a given dataset is known as dimensionality, and the process to reduce these features is called dimensionality reduction.

- A dataset contains a huge number of input features in various cases, which makes the predictive modeling task more complicated.

- Because it is very difficult to visualize or make predictions for the training dataset with a high number of features, for such cases, dimensionality reduction techniques are required to use.

- Dimensionality reduction technique can be defined as, *"**It is a way of converting the higher dimensions dataset into lesser dimensions dataset ensuring that it provides similar information.**"* These techniques are widely used in machine learning for obtaining a better fit predictive model while solving the classification and regression problems.

It is commonly used in the fields that deal with high-dimensional data, such as **speech recognition, signal processing, bioinformatics, etc. It can also be used for data visualization, noise reduction, cluster analysis**, etc.

The Curse of Dimensionality

- Handling the high-dimensional data is very difficult in practice, commonly known as the *curse of dimensionality*.

- If the dimensionality of the input dataset increases, any machine learning algorithm and model becomes more complex. As the number of features increases, the number of samples also gets increased proportionally, and the chance of overfitting also increases.

- If the machine learning model is trained on high-dimensional data, it becomes overfitted and results in poor performance.

- Hence, it is often required to reduce the number of features, which can be done with dimensionality reduction.

Benefits of applying Dimensionality Reduction

Some benefits of applying dimensionality reduction technique to the given dataset are given below:

- By reducing the dimensions of the features, the space required to store the dataset also gets reduced.
- Less Computation training time is required for reduced dimensions of features.
- Reduced dimensions of features of the dataset help in visualizing the data quickly.
- It removes the redundant features (if present) by taking care of multicollinearity.

Disadvantages of dimensionality Reduction

There are also some disadvantages of applying the dimensionality reduction, which are given below:

- Some data may be lost due to dimensionality reduction.

- In the PCA dimensionality reduction technique, sometimes the principal components required to consider are unknown.

Approaches of Dimension Reduction

There are two ways to apply the dimension reduction technique, which are given below:

1. Feature Selection

Three methods are used for the feature selection:

. Filters Methods

In this method, the dataset is filtered, and a subset that contains only the relevant features is taken. Some common techniques of filters method are:

- **Correlation**
- **Chi-Square Test**
- **Information Gain, etc.**

- **Wrappers Methods**

The wrapper method has the same goal as the filter method, but it takes a machine learning model for its evaluation.

In this method, some features are fed to the ML model, and evaluate the performance.

The performance decides whether to add those features or remove to increase the accuracy of the model.

This method is more accurate than the filtering method but complex to work.

Some common techniques of wrapper methods are:

- Forward Selection
- Backward Selection
- Bi-directional Elimination

- **Embedded Methods:**

Embedded methods check the different training iterations of the machine learning model and evaluate the importance of each feature. Some common techniques of Embedded methods are:

- **LASSO**
- **Elastic Net**
- **Ridge Regression, etc.**
- **Decision Tree**

2. Feature Extraction:

Feature extraction is the process of transforming the space containing many dimensions into space with fewer dimensions.

This approach is useful when we want to keep the whole information but use fewer resources while processing the information.

Some common feature extraction techniques are:

1. Principal Component Analysis
2. Linear Discriminant Analysis
3. Kernel PCA
4. Quadratic Discriminant Analysis