Dimensionality Reduction

Dimensionality reduction is a technique used to reduce the number of features in a dataset while retaining as much of the important information as possible.

In other words, it is a process of transforming high-dimensional data into a lower-dimensional space that still preserves the essence of the original data.

Dimensionality reduction can help to mitigate these problems by reducing the complexity of the model and improving its generalization performance.

There are two main approaches to dimensionality reduction: feature selection and feature extraction.

Feature Selection:

Feature selection involves selecting a subset of the original features that are most relevant to the problem at hand.

The goal is to reduce the dimensionality of the dataset while retaining the most important features.

It usually involves three methods:

1. Filter methods
2. Wrapper methods
3. Embedded methods

Filter methods rank the features based on their relevance to the target variable

wrapper methods use the model performance as the criteria for selecting features

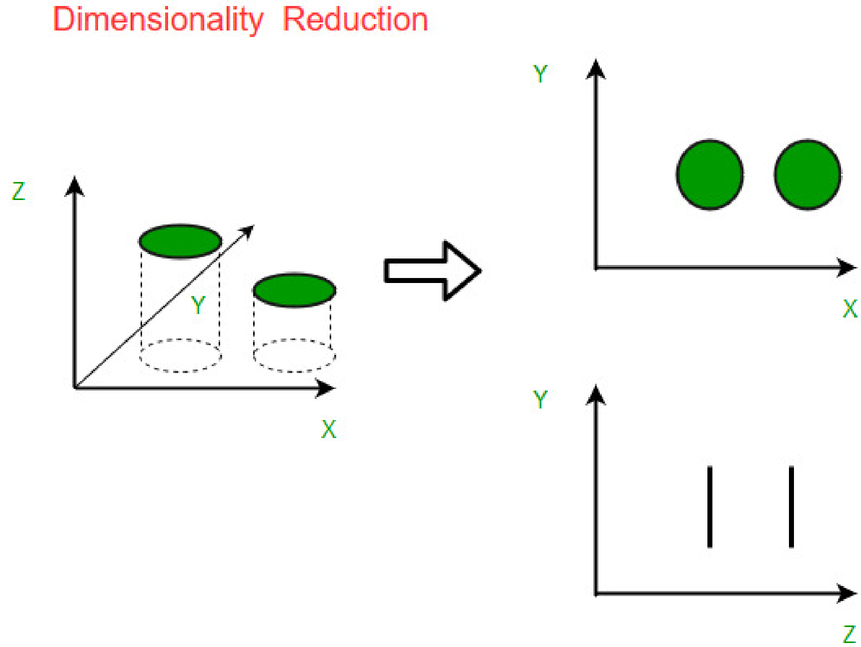
embedded methods combine feature selection with the model training process.

Feature Extraction:

Feature extraction involves creating new features by combining or transforming the original features. The goal is to create a set of features that captures the essence of the original data in a lower-dimensional space. There are several methods for feature extraction, including

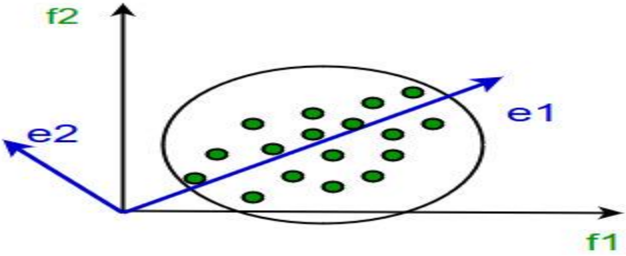
principal component analysis (PCA), linear discriminant analysis (LDA), and t-distributed stochastic neighbor embedding (t-SNE).

A 3-D classification problem can be hard to visualize, whereas a 2-D one can be mapped to a simple 2-dimensional space, and a 1-D problem to a simple line. The below figure illustrates this concept, where a 3-D feature space is split into two 2-D feature spaces, and later, if found to be correlated, the number of features can be reduced even further.



Principal Component Analysis

This method was introduced by Karl Pearson. It works on the condition that while the data in a higher dimensional space is mapped to data in a lower dimension space, the variance of the data in the lower dimensional space should be maximum.



It involves the following steps:

- Construct the covariance matrix of the data.

- Compute the eigenvectors of this matrix.
- Eigenvectors corresponding to the largest eigenvalues are used to reconstruct a large fraction of variance of the original data.

Advantages of Dimensionality Reduction

- It helps in data compression, and hence reduced storage space.
- It reduces computation time.
- It also helps remove redundant features, if any.

Disadvantages of Dimensionality Reduction

- It may lead to some amount of data loss.
- PCA tends to find linear correlations between variables, which is sometimes undesirable.
- PCA fails in cases where mean and covariance are not enough to define datasets.

Important points:

- Dimensionality reduction is the process of reducing the number of features in a dataset while retaining as much information as possible.

  This can be done to reduce the complexity of a model, improve the performance of a learning algorithm, or make it easier to visualize the data.

Independent Component Analysis

ICA is very similar to PCA, but the only assumption made by PCA that is alsomade by ICA is that there is a linear combination of the attributes.

]. By reducing the assumptions, ICA is able to find components with less redundancy than PCA, but at the cost of a higher processing time.

Another difference between PCA and ICA is that ICA does not rank the com- ponents. This is not a bad feature, since the principal components ranking order found by PCA is not always the best set of components.

Multidimensional Scaling

MDS involves a linear projection of a data set. However, while the previous techniques used the values of the attributes of the objects in the original data set, MDS uses the distances between pairs of objects.