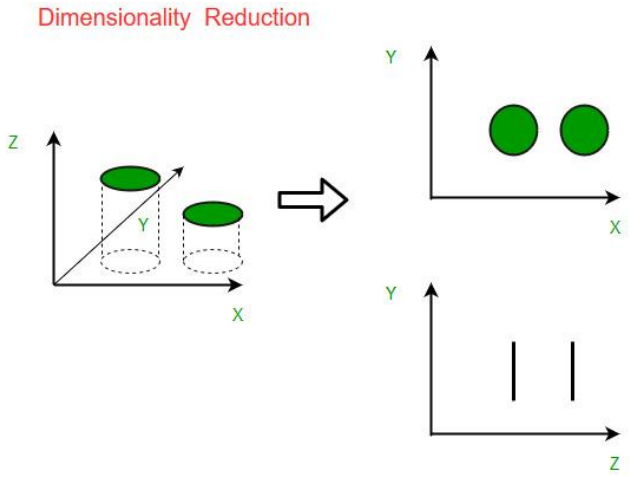
Dimensionality reduction

- In machine learning classification problems, there are often too many features on the basis of which the final classification is done
- Sometimes, most of these features are correlated, and hence redundant.
- Dimensionality reduction is the process of reducing the number of random variables under consideration, by obtaining a set of principal variables.
- This helps us in obtaining a better fit predictive model while solving the classification and regression problems.
- Commonly used for fields that deal with high-dimensional data, such as speech recognition, signal processing, bioinformatics, etc.

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.



Significance of DR

- 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.

DR Techniques

1. **Feature Selection**

- Feature selection is the process of selecting the subset of the relevant features and leaving out the irrelevant features present in a dataset to build a model of high accuracy.
- In other words, it is a way of selecting the optimal features from the input dataset
- Methods: Filter, Wrapper, Embedded

2. **Feature Extraction**

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

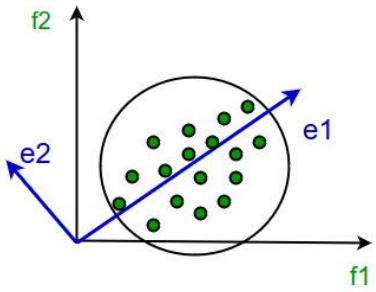
- This approach is useful when we want to keep the whole information but use fewer resources while processing the information.
- Methods : PCA, LDA

Methods used for DR

Principal Component Analysis (PCA)

- An unsupervised learning algo used for DR
- It is a statistical process that converts the observations of correlated features into a set of linearly uncorrelated features with the help of orthogonal transformation.
- These new transformed features are called the Principal Components.
- It is one of the popular tools that is used for exploratory data analysis and predictive modeling.
- Applications : image processing, movie recommendation system, etc

It works on a 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.



PCA Algorithm
The PCA algorithm is based on some mathematical concepts such as

- Variance and Covariance
- Eigenvalues and Eigen factors

Steps for PCA
https://www.javatpoint.com/principal-component-analysis

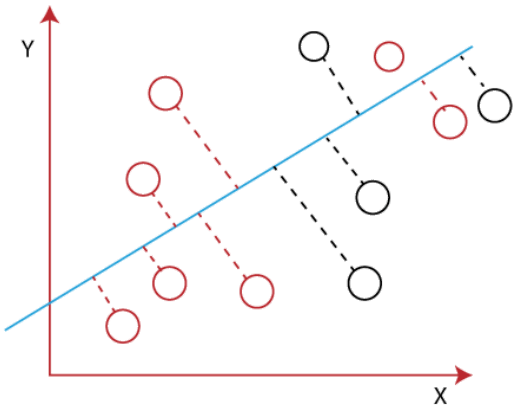
Linear Discriminant Analysis (LDA)

- It is a dimensionality reduction technique that is commonly used for supervised classification problems.
- It is used for modelling differences in groups i.e. separating two or more classes
- It is used to project the features in higher dimension space into a lower dimension space.

LDA working

- Let's consider an example where we have two classes in a 2-D plane having an X-Y axis, and we need to classify them efficiently.

- Here, LDA uses an X-Y axis to create a new axis by separating them using a straight line and projecting data onto a new axis.
- Hence, we can maximize the separation between these classes and reduce the 2-D plane into 1-D.



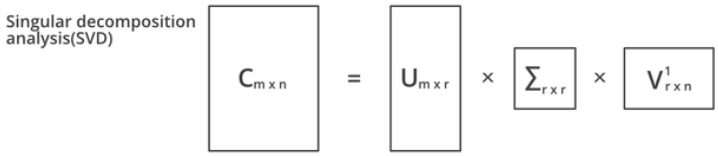
- To create a new axis, Linear Discriminant Analysis uses the following criteria:
    - It maximizes the distance between means of two classes.
    - It minimizes the variance within the individual class.

Steps for LDA

In book

Single valued decomposition

- The Singular Value Decomposition (SVD) of a matrix is a factorization of that matrix into three matrices.
- The SVD of mxn matrix A is given by the formula : A = USV^{T} where:
    - A represents mxn matrix
    - U represents mxn orthogonal matrix
    - S represents a nxn diagonal matrix
    - transpose of a nxn matrix



***Sums on all three in book**

Pca

https://youtu.be/MLaJbA82nzk