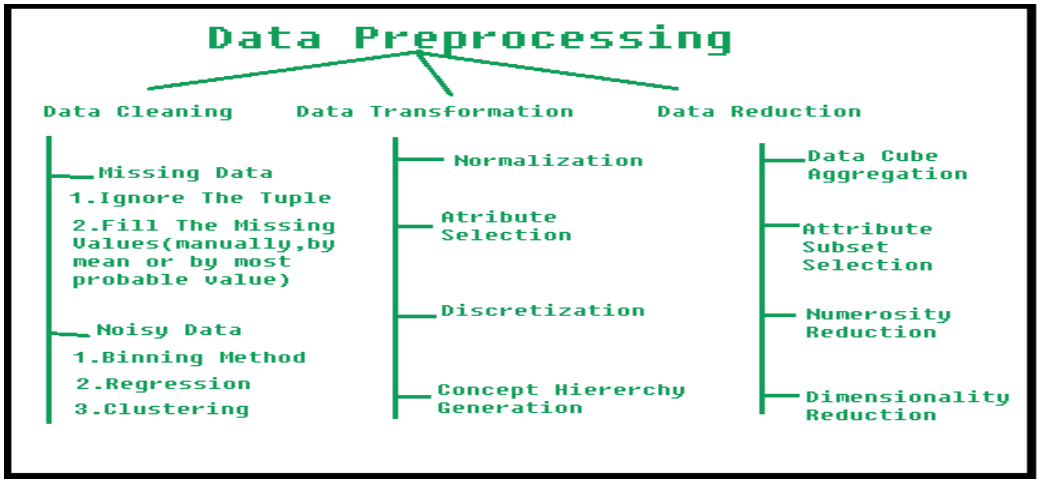
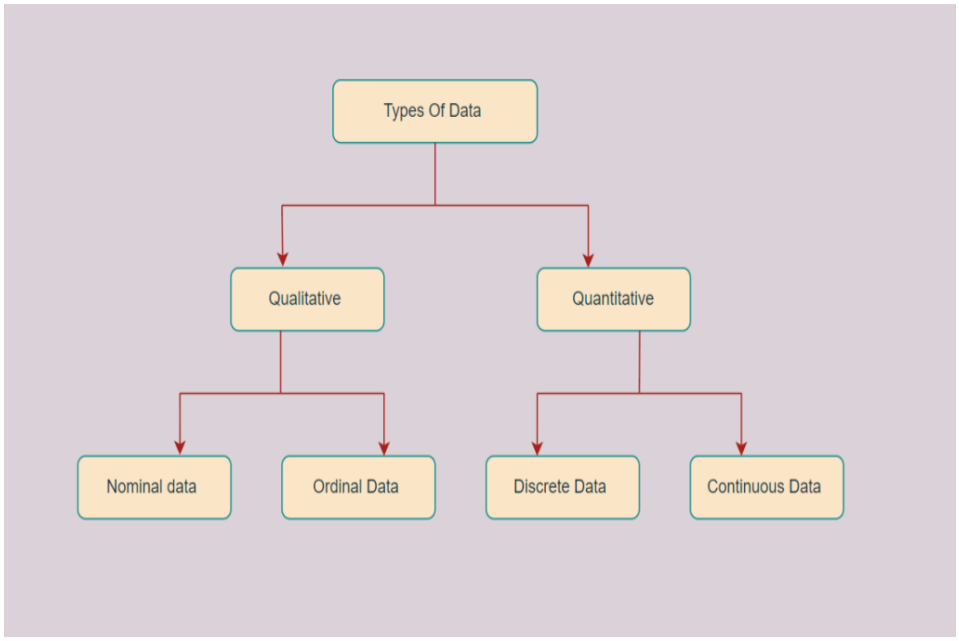
Module -2

Data Preprocessing

Data Preprocessing can be defined as a process of converting raw data into a format that is understandable and usable for further analysis. It is an important step in the Data Preparation stage. It ensures that the outcome of the analysis is **accurate**, **complete**, and **consistent**.



Types of data



A) Qualitative or Categorical Data

This type of data is descriptive and cannot be measured in numbers. It is often divided into two categories:

1. **Nominal Data**
    - **Definition**: Data that is used for labeling variables without any quantitative value.
    - **Example**: Gender, Nationality, Color.
2. **Ordinal Data**
    - **Definition**: Data that involves order or ranking.
    - **Example**: Ratings (good, better, best), Education level (high school, college, university).

B) Quantitative Data

Quantitative data is numerical and can be measured. This category also splits into two types:

3. **Discrete Data**
    - **Definition**: Data that can only take certain values and cannot be made more precise.
    - **Example**: Number of students in a class, Number of cars in a parking lot.
4. **Continuous Data**
    - **Definition**: Data that can take any value within a given range.
    - **Example**: Height, Weight, Temperature.

Sampling Theory

Sampling is a process of selecting subset of observations/records from a population to make inference about various population parameters such as mean, proportion, standard deviation, etc .

It is an important step in inferential statistics since an incorrect sample may lead to wrong inference about the population

POPULATION PARAMETERS:

Measures such as mean and standard deviation calculated using the entire population are called population parameters .

The population parameters mean and standard deviation are usually denoted using symbols μ and σ, respectively

SAMPLE STATISTIC

When population parameters are estimated from sample they are called sample statistic or statistic

The sample statistic is denoted using symbols (for mean) and S (or s for standard deviation)

The process of identifying a subset from a population of elements is called sampling process or simply sampling.

Steps used in any Sampling process:

1.Identification of target population that is important for a given problem under study.

2.Decide the sampling frame.

3.Determine the sample size.

4.Sampling method.

Random Sampling

Shewhart (1931) defines random sample as a ‘sample drawn under conditions such that the law of large number applies’

• Random sampling is usually carried out without replacement, that is, an observation which is selected in the sample is removed from the population for further consideration

• Random samples can also be created with replacement, that is, an observation which is selected for inclusion in the sample can again be considered since it is replaced (not removed) in the population.

Stratified Sampling

• The population can be divided into mutually exclusive groups using some factor (for example, age, gender, marital status, income, geographical regions, etc.).

The groups, thus, formed are called stratum

• It is important that the groups are mutually exclusive and exhaustive of the population.

Stratified Sampling -Examples

Amount of time spent by male and female users in sending messages in a day. Here the strata are male and female users.

b) Efficacy of a drug among different age groups. Age group can be classified into categories such as less than 40, between 41 and 60, and over 60 years of age.

c) Performance of children in school and the parents’ marital status. Here, marital status can be (a) Married, (b) Divorced. In this case we assume that the parent’s marital status may influence children’s academic performance.

d) Television rating points for a program across different geographical regions of a country. For India, geographical regions could be different states of the country.

Steps in creating stratified sample

a) Identify the factor that can be used for creating strata (for example: factor = Age; Strata 1: age less than 40; Strata 2: age between 41 and 60; and Strata 3: Age more than 60).

b) Calculate the proportion of each stratum in the population (say p1 , p2 , and p3 for three strata identified in step 1).

c) Calculate the sample size (say N). The sample size for strata 1, 2, and 3 identified in step 2 are p1 × N, p2 × N, and p3 × N, respectively. d) Use random sampling procedure to generate random samples in each strata. e) Combine samples from each stratum to create the final sample.

Cluster Sampling

In cluster sampling, the population is divided into mutually exclusive clusters

Cluster Sampling – Steps

Identify the clusters (example: different models of smart phones sold by a manufacturer, customers from different geographical locations).

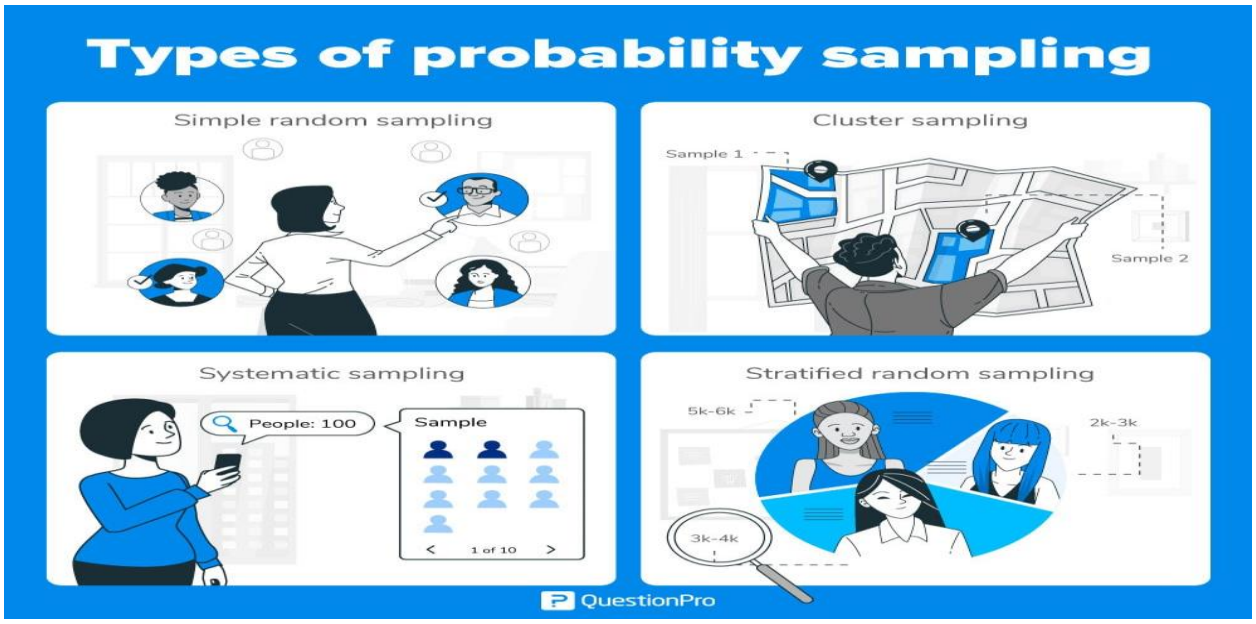
• Using random sampling select the clusters.

• Select all units in the clusters selected in step 2 and form the sample. If the size is too large, a random sampling within the clusters identified in step 2 may be used for final sample.

Types of sampling: sampling methods

Sampling in market action research is of two types – probability sampling and non-probability sampling. Let's take a closer look at these two methods of sampling.

1. **Probability sampling:** Probability sampling is a sampling technique where a researcher sets a selection of a few criteria and chooses members of a population randomly. All the members have an equal opportunity to be a part of the sample with this selection parameter.
2. **Non-probability sampling:** In non-probability sampling, the researcher chooses members for research at random. This sampling method is not a fixed or predefined selection process. This makes it difficult for all elements of a population to have equal opportunities to be included in a sample.



Correlation:

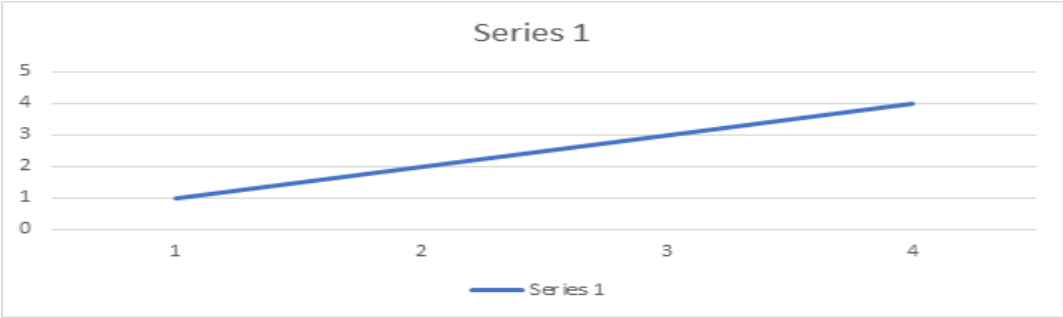
Correlation explains how one or more variables are related to each other. These variables can be input data features which have been used to forecast our target variable.

Correlation, statistical technique which determines how one variables moves/changes in relation with the other variable. It gives us the idea about the degree of the relationship of the two variables. It's a bi-variate analysis measure which describes the association between different variables. In most of the business it's useful to express one subject in terms of its relationship with others.

For example: No of testing vs no of positive cases in Corona.

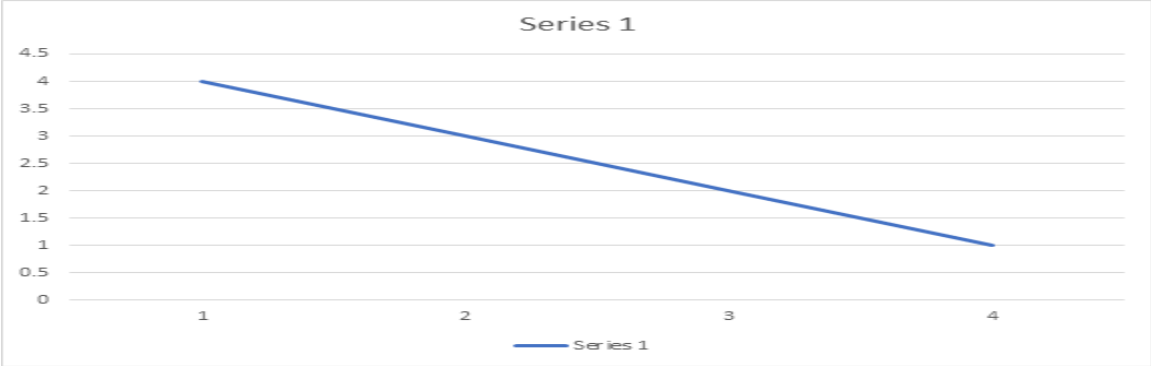
1. If two variables are closely correlated, then we can predict one variable from the other.
2. Correlation plays a vital role in locating the important variables on which other variables depend.
3. It's used as the foundation for various modeling techniques.
4. Proper correlation analysis leads to better understanding of data.
5. Correlation contribute towards the understanding of causal relationship (if any).

Positive Correlation: Two features (variables) can be positively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) also increases.



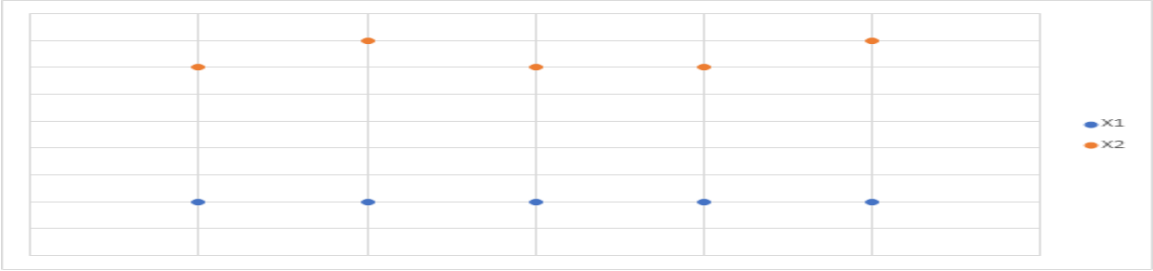
Positive Correlation

Negative Correlation: Two features (variables) can be negatively correlated with each other. It means that when the value of one variable increase then the value of the other variable(s) decreases.



Negative Correlation

No Correlation: Two features (variables) are not correlated with each other. It means that when the value of one variable increase or decrease then the value of the other variable(s) doesn't increase or decreases.



No Correlation

Feature selection:

The goal of feature selection in machine learning is to find the best set of features that allows one to build useful models of studied phenomena.

The techniques for feature selection in machine learning can be broadly classified into the following categories:

Supervised Techniques: These techniques can be used for labeled data, and are used to identify the relevant features for increasing the efficiency of supervised models like classification and regression.

Unsupervised Techniques: These techniques can be used for unlabeled data.

From a taxonomic point of view, these techniques are classified as under:

A. Filter methods

B. Wrapper methods

C. Embedded methods

D. Hybrid methods

Wrapper method

Forward Feature Selection

This is an iterative method wherein we start with the best performing variable against the target. Next, we select another variable that gives the best performance in combination with the first selected variable. This process continues until the preset criterion is achieved.

Backward Feature Elimination

This method works exactly opposite to the Forward Feature Selection method. Here, we start with all the features available and build a model. Next, we the variable from the model which gives the best evaluation measure value. This process is continued until the preset criterion is achieved.

Dimensionality Reduction:

https://www.gatevidyalay.com/tag/principal-component-analysis-numerical-example/

Feature selection:

Feature selection is a process that chooses a subset of features from the original features so that the feature space is optimally reduced according to a certain criterion.

Feature selection is a critical step in the feature construction process. In text categorization problems, some words simply do not appear very often. Perhaps the word “groovy” appears in exactly one training document, which is positive. Is it really worth keeping this word around as a feature ? It’s a dangerous endeavor because it’s hard to tell with just one training example if it is really correlated with the positive class or is it just noise. You could hope that your learning algorithm is smart enough to figure it out. Or you could just remove it.

There are three general classes of feature selection algorithms: **Filter methods, wrapper methods and embedded methods**.

The role of feature selection in machine learning is,

1. To reduce the dimensionality of feature space.

2. To speed up a learning algorithm.

3. To improve the predictive accuracy of a classification algorithm.

4. To improve the comprehensibility of the learning results.

Features Selection Algorithms are as follows:

1. **Instance based approaches:** There is no explicit procedure for feature subset generation. Many small data samples are sampled from the data. Features are weighted according to their roles in differentiating instances of different classes for a data sample. Features with higher weights can be selected.

2. Nondeterministic approaches: Genetic algorithms and simulated annealing are also used in feature selection.

3. Exhaustive complete approaches: Branch and Bound evaluates estimated accuracy and ABB checks an inconsistency measure that is monotonic. Both start with a full feature set until the preset bound cannot be maintained.

While building a machine learning model for real-life dataset, we come across a lot of features in the dataset and not all these features are important every time. Adding unnecessary features while training the model leads us to reduce the overall accuracy of the model, increase the complexity of the model and decrease the generalization capability of the model and makes the model biased. Even the saying "Sometimes less is better" goes as well for the machine learning model. Hence, **feature selection** is one of the important steps while building a machine learning model. Its goal is to find the best possible set of features for building a machine learning model.

Some popular techniques of feature selection in machine learning are:

- Filter methods
- Wrapper methods
- Embedded methods

Filter Methods

These methods are generally used while doing the pre-processing step. These methods select features from the dataset irrespective of the use of any machine learning algorithm. In terms of computation, they are very fast and inexpensive and are very good for removing duplicated, correlated, redundant features but these methods do not remove multicollinearity. Selection of feature is evaluated individually which can sometimes help when features are in isolation (don't

have a dependency on other features) but will lag when a combination of features can lead to increase in the overall performance of the model.

techniques used are:

- **Information Gain –** It is defined as the amount of information provided by the feature for identifying the target value and measures reduction in the entropy values. Information gain of each attribute is calculated considering the target values for feature selection.
- **Chi-square test —** Chi-square method (X2) is generally used to test the relationship between categorical variables. It compares the observed values from different attributes of the dataset to its expected value.

$$X^2 = \sum \frac{(Observed\ value - Expected\ value)^2}{Expected\ value}$$

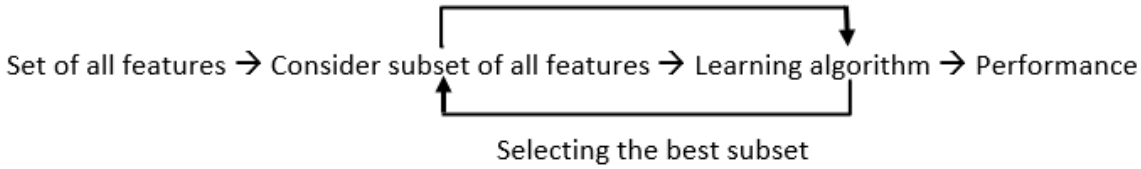
Chi-square Formula

- **Fisher's Score –** Fisher's Score selects each feature independently according to their scores under Fisher criterion leading to a suboptimal set of features. The larger the Fisher's score is, the better is the selected feature.
- **Correlation Coefficient –** Pearson's Correlation Coefficient is a measure of quantifying the association between the two continuous variables and the direction of the relationship with its values ranging from *-1 to 1*.
- **Variance Threshold –** It is an approach where all features are removed whose variance doesn't meet the specific threshold. By default, this method removes features having zero variance. The assumption made using this method is higher variance features are likely to contain more information.
- **Mean Absolute Difference (MAD) –** This method is similar to variance threshold method but the difference is there is no square in MAD. This method calculates the mean absolute difference from the mean value.

- **Dispersion Ratio –** Dispersion ratio is defined as the ratio of the Arithmetic mean (AM) to that of Geometric mean (GM) for a given feature. Its value ranges from $+1$ *to* ∞ *as* $AM \geq GM$ for a given feature. Higher dispersion ratio implies a more relevant feature.
- **Mutual Dependence –** This method measures if two variables are mutually dependent, and thus provides the amount of information obtained for one variable on observing the other variable. Depending on the presence/absence of a feature, it measures the amount of information that feature contributes to making the target prediction.
- **Relief –** This method measures the quality of attributes by randomly sampling an instance from the dataset and updating each feature and distinguishing between instances that are near to each other based on the difference between the selected instance and two nearest instances of same and opposite classes.

Wrapper methods:

Wrapper methods, also referred to as greedy algorithms train the algorithm by using a subset of features in an iterative manner. Based on the conclusions made from training in prior to the model, addition and removal of features takes place. Stopping criteria for selecting the best subset are usually pre-defined by the person training the model such as when the performance of the model decreases or a specific number of features has been achieved. The main advantage of wrapper methods over the filter methods is that they provide an optimal set of features for training the model, thus resulting in better accuracy than the filter methods but are computationally more expensive.



Wrapper Methods Implementation

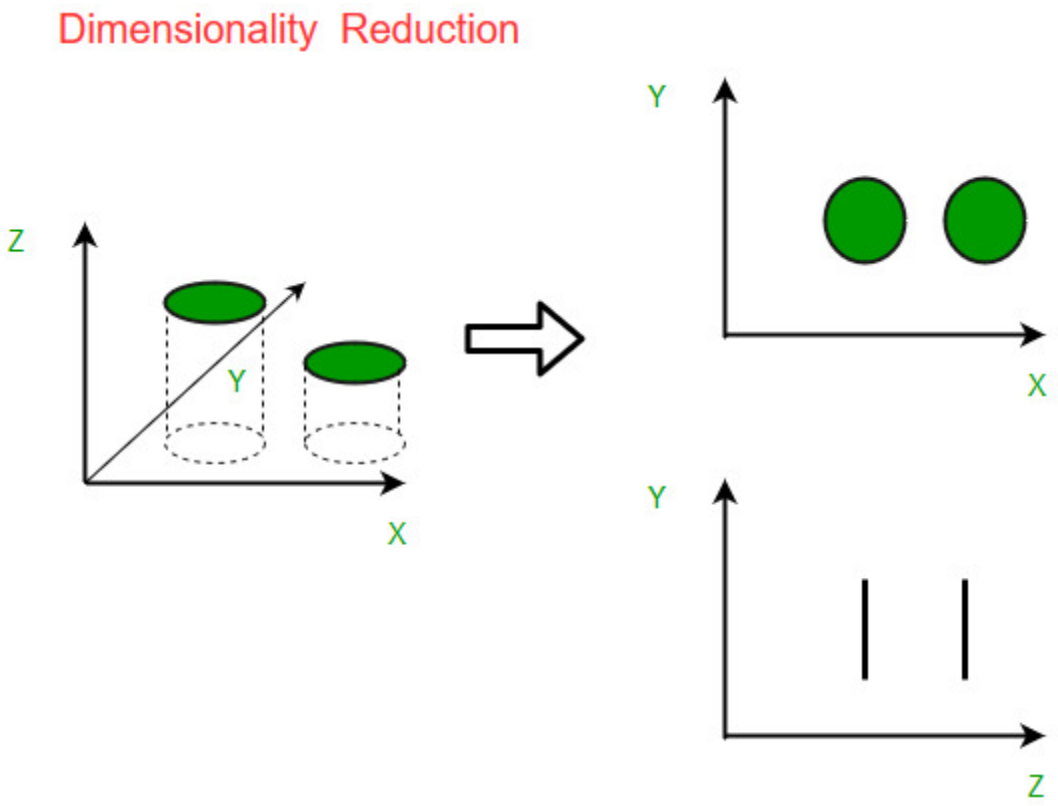
Some techniques used are:

- **Forward selection –** This method is an iterative approach where we initially start with an empty set of features and keep adding a feature which best improves our model after each iteration. The stopping criterion is till the addition of a new variable does not improve the performance of the model.
- **Backward elimination –** This method is also an iterative approach where we initially start with all features and after each iteration, we remove the least significant feature. The stopping criterion is till no improvement in the performance of the model is observed after the feature is removed.
- **Bi-directional elimination –** This method uses both forward selection and backward elimination technique simultaneously to reach one unique solution.
- **Exhaustive selection –** This technique is considered as the brute force approach for the evaluation of feature subsets. It creates all possible subsets and builds a learning algorithm for each subset and selects the subset whose model's performance is best.
- **Recursive elimination –** This greedy optimization method selects features by recursively considering the smaller and smaller set of features. The estimator is trained on an initial set of features and their importance is obtained using feature_importance_attribute. The least important features are then removed from the current set of features till we are left with the required number of features.
- 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.

- In machine learning, high-dimensional data refers to data with a large number of features or variables. The curse of dimensionality is a common problem in machine learning, where the performance of the model deteriorates as the number of features increases. This is because the complexity of the model increases with the number of features, and it becomes more difficult to find a good solution. In addition, high-dimensional data can also lead to overfitting, where the model fits the training data too closely and does not generalize well to new data.

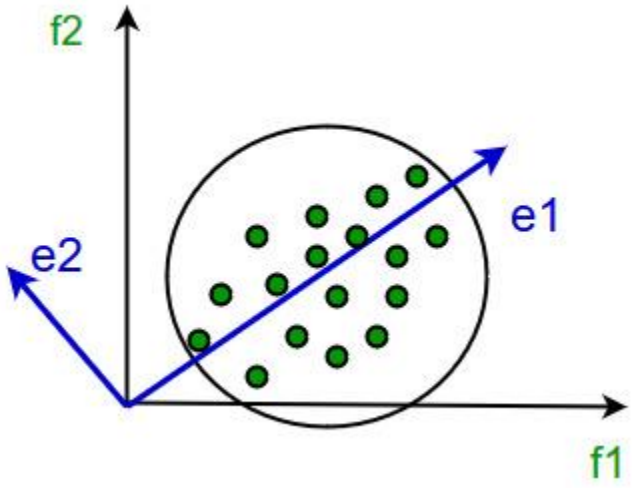
- Dimensionality reduction can help to mitigate these problems by reducing the complexity of the model and improving its generalization performance. There are two main approaches to dimensionality reduction: feature selection and feature extraction.
- Feature Selection:
  Feature selection involves selecting 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. There are several methods for feature selection, including filter methods, wrapper methods, and 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, and 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). PCA is a popular technique that projects the original features onto a lower-dimensional space while preserving as much of the variance as possible. **Why is Dimensionality Reduction important in Machine Learning and Predictive Modeling?**
- An intuitive example of dimensionality reduction can be discussed through a simple e-mail classification problem, where we need to classify whether the e-mail is spam or not. This can involve a large number of features, such as whether or not the e-mail has a generic title, the content of the e-mail, whether the e-mail uses a template, etc. However, some of these features may overlap. In another condition, a classification problem that relies on both humidity and rainfall can be collapsed into just one underlying feature, since both of the aforementioned are correlated to a high degree. Hence, we can reduce the number of features in such problems. 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.



- **Components of Dimensionality Reduction**
- There are two components of dimensionality reduction:
- **Feature selection:** In this, we try to find a subset of the original set of variables, or features, to get a smaller subset which can be used to model the problem. It usually involves three ways:
- Filter
- Wrapper
- Embedded
- **Feature extraction:** This reduces the data in a high dimensional space to a lower dimension space, i.e. a space with lesser no. of dimensions.
- **Methods of Dimensionality Reduction**
- The various methods used for dimensionality reduction include:
- Principal Component Analysis (PCA)
- Linear Discriminant Analysis (LDA)

- Generalized Discriminant Analysis (GDA)
- Dimensionality reduction may be both linear and non-linear, depending upon the method used. The prime linear method, called Principal Component Analysis, or PCA, is discussed below.
- **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.
- Hence, we are left with a lesser number of eigenvectors, and there might have been some data loss in the process. But, the most important variances should be retained by the remaining eigenvectors.
- **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.

- Improved Visualization: High dimensional data is difficult to visualize, and dimensionality reduction techniques can help in visualizing the data in 2D or 3D, which can help in better understanding and analysis.
- Overfitting Prevention: High dimensional data may lead to overfitting in machine learning models, which can lead to poor generalization performance. Dimensionality reduction can help in reducing the complexity of the data, and hence prevent overfitting.
- Feature Extraction: Dimensionality reduction can help in extracting important features from high dimensional data, which can be useful in feature selection for machine learning models.
- Data Preprocessing: Dimensionality reduction can be used as a preprocessing step before applying machine learning algorithms to reduce the dimensionality of the data and hence improve the performance of the model.
- Improved Performance: Dimensionality reduction can help in improving the performance of machine learning models by reducing the complexity of the data, and hence reducing the noise and irrelevant information in the data.
- **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.
- We may not know how many principal components to keep- in practice, some thumb rules are applied.
- Interpretability: The reduced dimensions may not be easily interpretable, and it may be difficult to understand the relationship between the original features and the reduced dimensions.
- Overfitting: In some cases, dimensionality reduction may lead to overfitting, especially when the number of components is chosen based on the training data.
- Sensitivity to outliers: Some dimensionality reduction techniques are sensitive to outliers, which can result in a biased representation of the data.

- Computational complexity: Some dimensionality reduction techniques, such as manifold learning, can be computationally intensive, especially when dealing with large 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.
- Techniques for dimensionality reduction include: principal component analysis (PCA), singular value decomposition (SVD), and linear discriminant analysis (LDA).
- Each technique projects the data onto a lower-dimensional space while preserving important information.
- Dimensionality reduction is performed during pre-processing stage before building a model to improve the performance
- It is important to note that dimensionality reduction can also discard useful information, so care must be taken when applying these techniques.

Eigenvalues and Eigenvectors are the scalar and vector quantities associated with matrices used for linear transformations. The vector that only changes by a scalar factor after applying a transformation is called an eigenvector, and the scalar value attached to the eigenvector is called the eigenvalue.

Eigenvalues Definition

Eigenvalues are the scalar values associated with the eigenvectors in linear transformation. The word 'Eigen' is of German Origin which means 'characteristic'.

Hence, these characteristic values indicate the factor by which eigenvectors are stretched in their direction. It doesn't involve the change in the direction of the vector

except when the eigenvalue is negative. When the eigenvalue is negative the direction is just reversed.

The equation for eigenvalue is given by

$$Av = \lambda v$$

Where,

- *A is the matrix,*
- *v is associated eigenvector, and*
- *λ is scalar eigenvalue.*

Types of Eigenvector

The eigenvectors calculated for the square matrix are of two types which are,

- Right Eigenvector
- Left Eigenvector

Right Eigenvector

The eigenvector which is multiplied by the given square matrix from the right-hand side is called the right eigenvector. It is calculated by using the following equation,

$$AVR = \lambda VR$$

Where,

- ***A*** *is given square matrix of order n×n,*
- ***λ*** *is one of the eigenvalues, and*
- ***VR*** *is the column vector matrix*

The value of VR is,

$VR=[v1v2v3..vn]\boldsymbol{VR}=v1v2v3..vn$

Left Eigenvector

The eigenvector which is multiplied by the given square matrix from the left-hand side is called the left eigenvector. It is calculated by using the following equation,

$\boldsymbol{VLA = VL\lambda}$

Where,

- $\boldsymbol{A}$ *is given square matrix of order* $n \times n$,
- $\boldsymbol{\lambda}$ *is one of the eigenvalues, and*
- ***VL*** *is the row vector matrix.*

The value of VL is,

$\boldsymbol{VL = [v1, v2, v3,\ldots, vn]}$

Eigenvectors of a Square Matrix

We can easily find the eigenvector of square matrices of order n × n. Now, let's find the following square matrices:

- Eigenvectors of a 2 × 2 matrix
- Eigenvectors of a 3 × 3 matrix.

Eigenvector of a 2 × 2 matrix

The Eigenvector of the 2 × 2 matrix can be calculated using the above mention steps. An example of the same is,

Example: Find the eigenvalues and the eigenvector for the matrix A = [1254][1524]

Solution:

If eigenvalues are represented using λ and the eigenvector is represented as v = [ab][ab]

Then the eigenvector is calculated by using the equation,

$|A- \lambda I| = 0$

$[1254]-\lambda[1001]=[0000][1524]-\lambda[1001]=[0000]$

$[1-\lambda 25 4-\lambda] \quad [1-\lambda 52 4-\lambda] \quad = 0$

$(1-\lambda)(4-\lambda) - 2.5 = 0$

$\Rightarrow 4 - \lambda - 4\lambda + \lambda 2 - 10 = 0$

$\Rightarrow \lambda 2 - 5\lambda - 6 = 0$

$\Rightarrow \lambda 2 - 6\lambda + \lambda - 6 = 0$

$\Rightarrow \lambda(\lambda-6) + 1(\lambda-6) = 0$

$\Rightarrow (\lambda-6)(\lambda+1) = 0$

$\lambda = 6$ *and* $\lambda = -1$

Thus, the eigenvalues are 6, and -1. Then the respective eigenvectors are,

For λ = 6

$(A-\lambda I)v = 0$

$\Rightarrow [1-6254-6].[ab] \quad [1-6524-6].[ab] \quad = 0$

$\Rightarrow [-525-2].[ab] \quad [-552-2].[ab] \quad = 0$

$\Rightarrow -5a + 2b = 0$

$\Rightarrow 5a - 2b = 0$

Simplifying the above equation we get,

$5a=2b$

The required eigenvector is,

$[ab]=[25][ab]=[25]$

For λ = -1

$(A-\lambda I)v = 0$

$\Rightarrow [1-(-1)254-(-1)].[ab] \quad [1-(-1)524-(-1)].[ab] \quad = 0$

$\Rightarrow [2255].[ab] \quad [2525].[ab] \quad = 0$

$\Rightarrow 2a + 2b = 0$

$\Rightarrow 5a + 5b = 0$

simplifying the above equation we get,

$a = -b$

The required eigenvector is,

$[ab]=[1-1][ab]=[1-1]$

Then the eigenvectors of the given 2 × 2 matrix are $[ab]=[25],[ab]=[1-1][ab]=[25]$ *,*$[ab]=[1-1]$

These are two possible eigen vectors but many of the corresponding multiples of these eigen vectors can also be considered as other possible eigen vectors.

Eigenvector of a 3 × 3 Matrix

The Eigenvector of the 3 × 3 matrix can be calculated using the above mention steps. An example of the same is,

Example: Find the eigenvalues and the eigenvector for the matrix A = [222222222] 222222222

Solution:

If eigenvalues are represented using λ and the eigenvector is represented as v = [abc] abc

Then the eigenvector is calculated by using the equation,

$|A- \lambda I| = 0$

$[222222222]-\lambda[100010001]=[000000000] 222222222-\lambda 100010001=000000000$

$[2-\lambda 2222-\lambda 2222-\lambda] \quad 2-\lambda 2222-\lambda 2222-\lambda \quad = 0$

Simplifying the above determinant we get

$\Rightarrow (2-\lambda)(\lambda 2) + 2\lambda 2 + 2\lambda 2 = 0$

$\Rightarrow (-\lambda 3) + 6\lambda 2 = 0$

$\Rightarrow \lambda 2(6 - \lambda) = 0$

$\Rightarrow \lambda = 0, \lambda = 6$

For λ = 0

$(A - \lambda I)\ v = 0$

$\Rightarrow [2–02222–02222−0].[abc] \quad 2–02222–02222−0.abc \quad = 0$

$\Rightarrow [222222222].[abc] \quad 222222222.abc \quad = 0$

Simplifying the above equation we get

$2a + 2b + 2c = 0$

$\Rightarrow 2(a+b+c) = 0$

$\Rightarrow a+b+c = 0$

Let b = k1 and c = k2

$a + k1 + k2 = 0$

$a = -(k1 + k2)$

Thus, the eigenvector is,

$[abc]=[−(k1+k2)k1k2]abc=−(k1+k2)k1k2$

taking k1 = 1 and k2 = 0

the eigenvector is,

$[abc]=[−110]abc=−110$

taking k1 = 0 and k2= 1

the eigenvector is,

$[abc]=[−101]abc=−101$

For λ = 6

$(A - \lambda I)\, v = 0$

$\Rightarrow [2–62222−62222−6].[abc] \quad 2–62222−62222−6.abc \quad = 0$

$\Rightarrow [−4222−4222−4].[abc] \quad −4222−4222−4.abc \quad = 0$

Simplifying the above equation we get,

$-4a + 2b + 2c = 0$

$\Rightarrow 2(-2a + b + c) = 0$

$\Rightarrow -2a = -(b + c)$

$\Rightarrow 2a = b + c$

Let b = k1 and c = k2 , and taking k1 = k2 = 1,

we get,

$[abc]=[111]abc=111$

Thus, the eigenvector is,

$[abc]=[111]abc=111$

Eigenspace

We define the eigenspace of a matrix as the set of all the eigenvectors of the matrix. All the vectors in the eigenspace are linearly independent of each other.

To find the Eigenspace of the matrix we have to follow the following steps

Step 1: *Find all the eigenvalues of the given square matrix.*

Step 2: *For each eigenvalue find the corresponding eigenvector.*

Step 3: *Take the set of all the eigenvectors (say A). The resultant set so formed is called the Eigenspace of the following vector.*

From the above example of given 3 × 3 matrix A, the eigenspace so formed is{ [000],[−10−1],[−1−10] 000,−10−1,−1−10 }

Applications of Eigen Values in Engineering

Some of the common applications of eigen values are:

Linear Algebra

Diagonalization: Eigenvalues are used to diagonalize matrices, simplifying computations and solving linear systems more efficiently.

Matrix Exponentiation: Eigenvalues play a crucial role in computing the exponentiation of a matrix.

Quantum Mechanics

Schrödinger Equation: Eigenvalues of the Hamiltonian operator correspond to the energy levels of quantum systems, providing information about possible states.

Vibrations and Structural Analysis:

Mechanical Vibrations: Eigenvalues represent the natural frequencies of vibrational systems. In structural analysis, they help understand the stability and behavior of structures.

Statistics

Covariance Matrix: In multivariate statistics, eigenvalues are used in the analysis of covariance matrices, providing information about the spread and orientation of data.

Computer Graphics

Principal Component Analysis (PCA): Eigenvalues are used in PCA to find the principal components of a dataset, reducing dimensionality while retaining essential information.

Control Systems

System Stability: Eigenvalues of the system matrix are critical in determining the stability of a control system. Stability analysis helps ensure that the system response is bounded.

Diagonalize Matrix Using Eigenvalues and Eigenvectors

Eigenvalues and Eigenvectors are used to find diagonal matrices. A diagonal matrix is a matrix which can be written as,

A = XDX-1

Where,

- ***D*** *is the matrix which is formed by replacing the 1's in the identity matrix by eigenvalues, and*
- ***X*** *is the matrix formed by eigenvectors.*

We can understand the concept of a diagonal matrix by taking the following example.

Example: Diagonalize the matrix A = [222222222]222222222

Solution:

We have already solved for the eigenvalues and the eigenvectors of the A = [222222222]222222222

The eigenvalues of the A are $\lambda = 0$, $\lambda = 0$, *and* $\lambda = -8$

The eigenvectors of A are [000],[−10−1],[−1−10] 000,−10−1,−1−10

Thus,

D = [00000000−8]00000000−8

X = [0−1−100−10−10]000−10−1−1−10

We can easily find the inverse of X as,

X-1 = [0−1−100−10−10]000−10−1−1−10

Articles related to Eigenvalues:

- ***Elementary Operation on Matrices***
- ***Identity Matrix***
- ***Inverse of a Matrix***

Solved Examples on Eigenvectors

Example 1: Find the eigenvectors of the matrix A = [110011001]100110011

Solution:

The eigen values of the matrix is found using,

|A – λI| = 0

[1−λ1001−λ1001−λ] 1−λ0011−λ0011−λ = 0

(1 – λ)3 = 0

Thus, the eigen values are,

λ = 1, 1, 1

As the all the eigenvalues are equal we have three identical eigenvectors. We will find the eigenvectors for $\lambda = 1$, using $(A - \lambda I)v = O$

$[1-11001-11001-1].[abc]=[000]1-10011-10011-1.abc=000$

$[010001000].[abc]=[000]000100010.abc=000$

solving the above equation we get,

- *$a = K$*
- *$y = 0$*
- *$z = 0$*

Then the eigenvector is,

$[abc]=[k00]=k[100]abc=k00=k100$

Principal Component Analysis(PCA)

As the number of features or dimensions in a dataset increases, the amount of data required to obtain a statistically significant result increases exponentially. This can lead to issues such as overfitting, increased computation time, and reduced accuracy of machine learning models this is known as the curse of dimensionality problems that arise while working with high-dimensional data.

As the number of dimensions increases, the number of possible combinations of features increases exponentially, which makes it computationally difficult to obtain a representative sample of the data. It becomes expensive to perform tasks such as clustering or classification because the algorithms need to process a much larger feature space, which increases computation time and complexity. Additionally, some machine learning algorithms can be sensitive to the number of dimensions, requiring more data to achieve the same level of accuracy as lower-dimensional data.

What is Principal Component Analysis(PCA)?

Principal Component Analysis(PCA) technique was introduced by the mathematician **Karl Pearson** in 1901**.** 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.

- **Principal Component Analysis (PCA)** is a statistical procedure that uses an orthogonal transformation that converts a set of correlated variables to a set of uncorrelated variables.PCA is the most widely used tool in exploratory data analysis and in machine learning for predictive models. Moreover,

- Principal Component Analysis (PCA) is an unsupervised learning algorithm technique used to examine the interrelations among a set of variables. It is also known as a general factor analysis where regression determines a line of best fit.

- The main goal of Principal Component Analysis (PCA) is to reduce the dimensionality of a dataset while preserving the most important patterns or relationships between the variables without any prior knowledge of the target variables.

Principal Component Analysis (PCA) is used to reduce the dimensionality of a data set by finding a new set of variables, smaller than the original set of variables, retaining most of the sample's information, and useful for the regression and classification of data.

Step-By-Step Explanation of PCA (Principal Component Analysis)

Step 1: Standardization

First, we need to standardize our dataset to ensure that each variable has a mean of 0 and a standard deviation of 1.

$Z=X-\mu\sigma Z=\sigma X-\mu$

Here,

- $\mu\mu$ is the mean of independent features $\mu=\{\mu1,\mu2,\cdots,\mu m\}\mu=\{\mu1,\mu2,\cdots,\mu m\}$
- $\sigma\sigma$ is the standard deviation of independent features $\sigma=\{\sigma1,\sigma2,\cdots,\sigma m\}\sigma=\{\sigma1,\sigma2,\cdots,\sigma m\}$

Step2: Covariance Matrix Computation

Covariance measures the strength of joint variability between two or more variables, indicating how much they change in relation to each other. To find the covariance we can use the formula:

$cov(x1,x2)=\sum i=1n(x1i-x1^-)(x2i-x2^-)n-1cov(x1,x2)=n-1\sum i=1n(x1i-x1^-)(x2i-x2^-)$

The value of covariance can be positive, negative, or zeros.

- Positive: As the x1 increases x2 also increases.
- Negative: As the x1 increases x2 also decreases.
- Zeros: No direct relation

Step 3: Compute Eigenvalues and Eigenvectors of Covariance Matrix to Identify Principal Components

Let A be a square nXn matrix and X be a non-zero vector for which

$AX=\lambda XAX=\lambda X$

for some scalar values $\lambda\lambda$. then $\lambda\lambda$ is known as the eigenvalue of matrix A and X is known as the eigenvector of matrix A for the corresponding eigenvalue.

It can also be written as :

$AX-\lambda X=0(A-\lambda I)X=0AX-\lambda X(A-\lambda I)X=0=0$

where I am the identity matrix of the same shape as matrix A. And the above conditions will be true only if (A–λI)(*A–λI*) will be non-invertible (i.e. singular matrix). That means,

|A–λI|=0|*A–λI*|=0

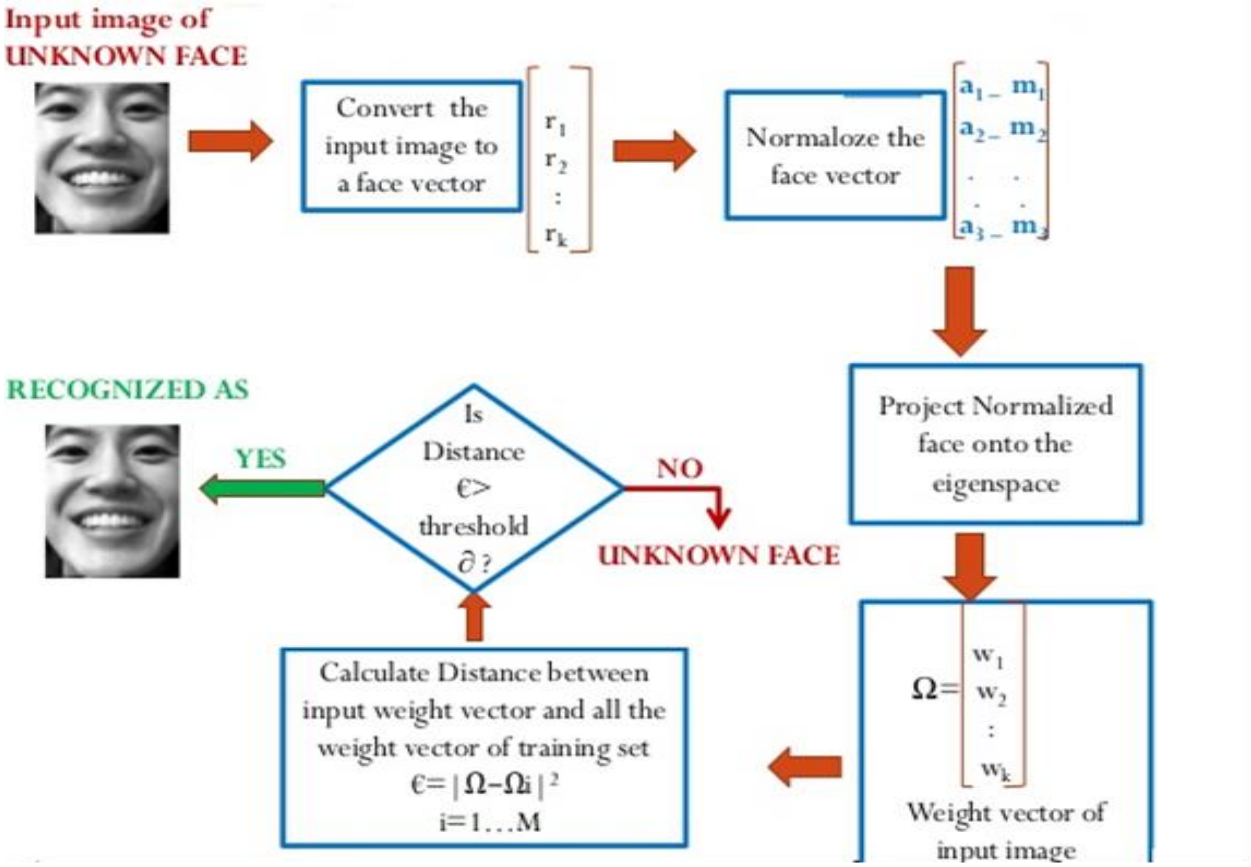
From the above equation, we can find the eigenvalues \lambda, and therefore corresponding eigenvector can be found using the equation AX=λX*AX=λX*.

Principal Component Analysis

1. Principal Component Analysis (PCA) is a technique for dimensionality reduction that identifies a set of orthogonal axes, called principal components, that capture the maximum variance in the data. The principal components are linear combinations of the original variables in the dataset and are ordered in decreasing order of importance. The total variance captured by all the principal components is equal to the total variance in the original dataset.

2. The first principal component captures the most variation in the data, but the second principal component captures the maximum variance that is orthogonal to the first principal component, and so on.

Process flow diagram of face recognition using PCA


Input image of UNKNOWN FACE
Convert the input image to a face vector
Normaloze the face vector
Project Normalized face onto the eigenspace
Weight vector of input image
Calculate Distance between input weight vector and all the weight vector of training set
i=1…M
Is Distance > threshold ∂ ?
YES
NO
UNKNOWN FACE
RECOGNIZED AS

Question Bank

1.Explain the types of Data in detail?

2.Explain Samploing theory?

3.Explain the types of Sampling Methods?

4.Explain Correlation and Types of Correlation ?

5.Explain Eigen values and Eigen vectors?

6.Difference between Eigen values and Eigen vectors?

7.Explain the methods of feature selection?

8.Explain feature selection Algorithms?

9.Explain dimensionality reduction techniques methods?

10. **Prioritize the Sampling theory in Data Science with suitable examples.**

11. **Estimate the Process of Feature Selection with suitable example.**

12. Describe the various steps in the Principal Component Analysis for the following dataset.

{ 2,3,4 ; 1,2,3}

13. **Identify the covariance matrix for the following dataset.{5,6,7 ; 6,7,8}**

14. Describe the importance of the Principal Component Analysis. Identify mean vector for the following dataset. { 2,3,4 ; 1,2,3}

15. **Identify the Eigen vectors for the following dataset.{2,3,4 ; 5,7,9}**

16.Explain methods of Principle component analysis

17.Explain the process flow of PCA with a neat diagram?

18.Explain the applications of Eigen Values in Engineering?

19. Find the eigenvectors of the matrix A = [110011001]100110011

20.write down the difference between Qualitative and quantitative data ?