# DEEP LEARNING (20A05703c)

**UNIT I**

Linear Algebra: Scalars, Vectors, Matrices and Tensors, Matrix operations, types of matrices, Norms, Eigen decomposition, Singular Value Decomposition, Principal Components Analysis. Probability and Information Theory: Random Variables, Probability Distributions, Marginal Probability, Conditional Probability, Expectation, Variance and Covariance, Bayes' Rule, Information Theory. Numerical Computation: Overflow and Underflow, Gradient-Based Optimization, Constrained Optimization, Linear Least Squares.

## LINEAR ALGEBRA

Linear algebra is a branch of mathematics that deals with linear equations and their representations in vector space using matrices. In other words, linear algebra is the study of linear functions and vectors. It is one of the most central topics of mathematics. Most modern geometrical concepts are based on linear algebra. A good understanding of linear algebra is essential for understanding and working with many machine learning algorithms, especially deep learning algorithms.

### Scalars

**Scalars** are single numbers and are an example of a 0th-order tensor. In mathematics, it is necessary to describe the set of values to which a scalar belongs. The notation $x \in R$ states that the scalar value x is an element of (or member of) the set of real-valued numbers, R.

There are various sets of numbers of interest within machine learning. N represents the *set of positive integers* (1,2, 3…). Z represents the integers, which include *positive, negative, and zero values*. Q represents the set of *rational* numbers that may be expressed as a fraction of two integers whereas P represents the *irrational* numbers. And R is a set that contains all the above number sets i.e., N, Z, Q, and P.

### Vectors

A vector is often represented as a 1-dimensional array of numbers, referred to as components, and is displayed either in column form or row form. It is an example of a 1st-order tensor.

Vectors are often represented using a lowercase character such as “v”; for example:

v = (v1, v2, v3)

Where v1, v2, and v3 are scalar values, often real values.

Vectors are also shown using a vertical representation or a column; for example:

$$v = \begin{pmatrix} v1 \\ v2 \\ v3 \end{pmatrix}$$

## Matrices

Matrices are rectangular arrays consisting of numbers and can be seen as 2*nd*-order tensors. If m and n are positive integers, that is m, n ∈ ℕ then the m×n matrix contains m*n numbers of elements, with m number of rows and n number of columns.

The pictorial representation of an m×n matrix is shown below:

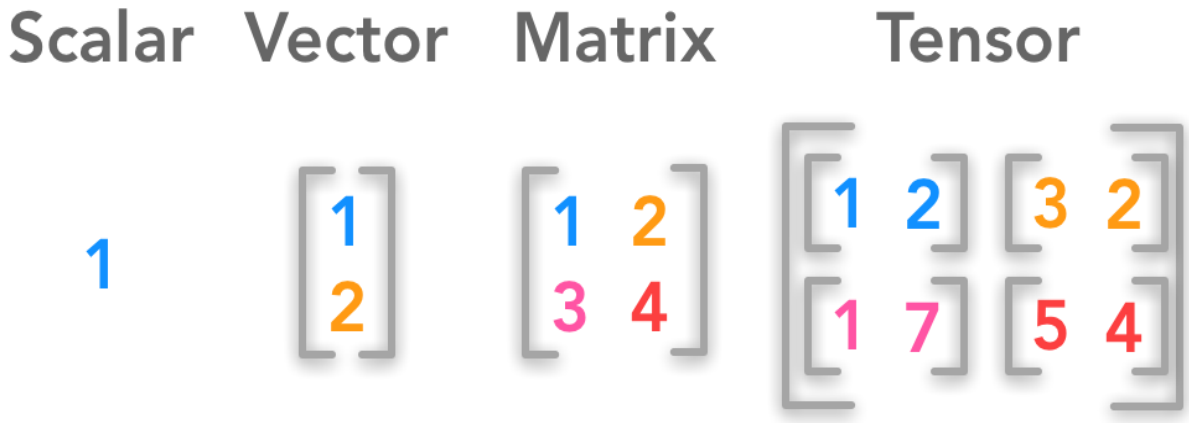
$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1n} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2n} \\ a_{31} & a_{32} & a_{33} & \cdots & a_{3n} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & a_{m3} & \cdots & a_{mn} \end{bmatrix}$$

Sometimes, instead of describing the full matrix components, we use the following abbreviation of a matrix:

$$A=[a_{ij}]_{m\times n}$$

**Tensors**

A tensor is a generalization of vectors and matrices and is easily understood as a multidimensional array. The inputs, outputs, and transformations within neural networks are all represented using tensors, and as a result, neural network programming utilizes tensors heavily.




A matrix is like a simple BOX. A matrix can be rectangular(nXm) or square(nXn). So tensor is an n-dimensional array satisfying a particular transformation law. Unlike a matrix, it shows an object placed in a specific coordinate system**.**

**Matrix operations:**

**Operations on Matrices**

Addition, subtraction and multiplication are the basic operations on the matrix. To add or subtract matrices, these must be of identical order, and for multiplication, the number of columns in the first matrix equals the number of rows in the second matrix.

- Addition of Matrices

- Subtraction of Matrices
- Scalar Multiplication of Matrices
- Multiplication of Matrices

## i. Addition of Matrices

If $A[a_{ij}]_{mxn}$ and $B[b_{ij}]_{mxn}$ are two matrices of the same order, then their sum A + B is a matrix, and each element of that matrix is the sum of the corresponding elements, i.e. $A + B = [a_{ij} + b_{ij}]_{mxn}$

Consider the two matrices, A and B, of order 2 x 2. Then, the sum is given by:

$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} + \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} a_1 + a_2 & b_1 + b_2 \\ c_1 + c_2 & d_1 + d_2 \end{bmatrix}$$

## ii. Subtraction of Matrices

If A and B are two matrices of the same order, then we define

$A - B = A + (-B)$.

Consider the two matrices, A and B, of order 2 x 2. Then, the difference is given by:

$$\begin{bmatrix} a_1 & b_1 \\ c_1 & d_1 \end{bmatrix} - \begin{bmatrix} a_2 & b_2 \\ c_2 & d_2 \end{bmatrix} = \begin{bmatrix} a_1 - a_2 & b_1 - b_2 \\ c_1 - c_2 & d_1 - d_2 \end{bmatrix}$$

We can subtract the matrices by subtracting each element of one matrix from the corresponding element of the second matrix, i.e. $A - B = [a_{ij} - b_{ij}]_{mxn}$.

## iii. Scalar Multiplication of Matrices

If $A = [a_{ij}]_{m \times n}$ is a matrix and k any number, then the matrix which is obtained by multiplying the elements of A by k is called the scalar multiplication of A by k, and it is denoted by k A, thus if $A = [a_{ij}]_{m \times n}$,

Then

$$kA_{m \times n} = A_{m \times n}k = [ka_{i \times j}]$$

## iv. Multiplication of Matrices

If A and B be any two matrices, then their product AB will be defined only when the number of columns in A is equal to the number of rows in B.

If

$A = [a_{ij}]_{m \times n}. and\, B = [b_{ij}]_{n \times p} then\ their\ product\ AB = C = [c_{ij}]_{m \times p}$
will be a matrix of order m×p where
$(AB)_{ij} = C_{ij} = \sum_{r=1}^{n} a_{ir} b_{rj}$

## Types of matrices:

There are many types of matrices depending on the elements in the matrix, order, and certain sets of conditions. The different types of matrices are mentioned below:

- Singleton Matrix
- Null Matrix
- Row Matrix
- Column Matrix
- Horizontal Matrix
- Vertical Matrix
- Rectangular Matrix
- Square Matrix
- Diagonal Matrix
- Scalar Matrix
- Stochastic Matrix
- Identity Matrix
- Triangular Matrix
- Upper Triangular Matrix
- Lower Triangular Matrix
- Singular Matrix
- Non Singular Matrix
- Symmetric Matrix
- Skew Symmetric Matrix
- Orthogonal Matrix
- Idempotent Matrix
- Nilpotent Matrix
- Periodic Matrix
- Involutory Matrix
- Hermitian Matrix
- Skew Hermitian Matrix
- Boolean Matrix

Let’s learn the above types of matrices in detail

**Singleton Matrix**

A matrix that has only one element is called a singleton matrix. In this type of matrix number of columns and the number of rows is equal to 1. A singleton matrix is represented as [a]$1 \times 1$.

Example of Singleton Matrix

[5]

In the above example of the Singleton Matrix, there is only one element 5. Hence there is only one column and only one row.

**Null Matrix**

A matrix whose all elements are zero is called a Null Matrix. A null matrix is also called a Zero Matrix because its all elements are zero.

## Row Matrix

A matrix that contains only one row and any number of columns is known as a row matrix. A row matrix is represented as [a]1×n where 1 is the number of row and n is the number of columns present in a row matrix. An example of a row matrix is given below

### Example of Row Matrix

$$\begin{bmatrix} 1 & 3 & 7 \end{bmatrix}_{1\times3}$$

In the above example of a row matrix, the number of rows is 1, and the number of columns is 3. Hence the order of the matrix is 1 × 3.

## Column Matrix

A matrix that contains only one column and any number of rows is called a Column Matrix. A Column Matrix is represented as [a]n×1 where n is the number of rows and 1 is the number of columns. An example of a column matrix is given below:

### Example of Column Matrix

$$\begin{bmatrix} 1 \\ 15 \\ 4 \\ 5 \end{bmatrix}_{4\times1}$$

In the above example of a column matrix the number of rows is 4 and the number of columns is 1 thus making it a matrix of order 4 × 1.

### Horizontal Matrix

A matrix in which the number of rows is lower than the number of columns is called a Horizontal Matrix. columns

#### Example of Horizontal Matrix

$$\begin{bmatrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{bmatrix}_{2\times4}$$

In the above matrix, the number of rows is 2 while the number of columns is 4 thus making it a horizontal matrix.

### Vertical Matrix

The matrix in which the number of rows exceeds the number of columns is called a Vertical Matrix. A Vertical matrix is represented as [a]i×j where i > j. An example of a Vertical Matrix is mentioned below:

#### Example of Vertical Matrix

$$\begin{bmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{bmatrix}_{4\times2}$$

In the above matrix, the number of rows is 4 while the number of columns is 2 thus making it a Vertical matrix.



## Rectangular Matrix

A matrix that does not have an equal number of rows and columns is known as a Rectangular Matrix. A rectangular matrix can be represented as $[A]_{m\times n}$ where m ≠ n. An example of a rectangular matrix is mentioned below:

### Example of Rectangular Matrix

$$\begin{bmatrix} 1 & 3 & 7 & 15 \\ 3 & 4 & 6 & 11 \\ 5 & 2 & 9 & 8 \end{bmatrix}_{3\times4}$$

In the above example, we see that the number of rows is 3 while the number of columns is 4 i.e. both are unequal thus making it a rectangular matrix. We can say that both horizontal and vertical matrices are examples of rectangular matrices.

### Square Matrix

A matrix that has an equal number of rows and an equal number of columns is called a Square Matrix. Generally, the representation used for the square matrix is $[A]_{n\times n}$. An example of Square Matrix is mentioned below:

**Example of Square Matrix**

$$\begin{bmatrix} 8 & 3 & 2 \\ 6 & 4 & 6 \\ 5 & 7 & 9 \end{bmatrix}_{3\times 3}$$

In the above example of Square Matrix, both the number of rows and columns are 3, thus making them seem like a square structure.

### Diagonal Matrix

A matrix that has all elements as 0 except diagonal elements is known as a diagonal matrix. A Diagonal Matrix is only possible in the case of a Square Matrix. An example of a Diagonal Matrix is mentioned below:

**Example of Diagonal Matrix**

$$\begin{bmatrix} 8 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 9 \end{bmatrix}_{3\times 3}$$

In the above example, the diagonal elements are 8, 4, and 9 and the rest elements are zero.

## Scalar Matrix

A diagonal matrix whose all diagonal elements are non-zero and the same is called a Scalar Matrix. Scalar Matrix is a kind of diagonal matrix where all diagonal elements are the same. Identity Matrix is a special case of Scalar Matrix.

### Example of Scalar Matrix

$$\begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \\ 0 & 0 & 4 \end{bmatrix}_{3\times 3}$$

In the above example, the given matrix is a diagonal matrix whose all diagonal elements are 4, and hence, this is an example of a Scalar Matrix.

## Identity Matrix

A diagonal matrix where all the diagonal elements are 1 and all non-diagonal elements are 0 is called an Identity Matrix. The Identity Matrix is called Unit Matrix. The identity matrix or unit matrix always has an equal number of rows and columns.

### Example of Identity Matrix

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}_{3\times 3}$$

In the above diagonal matrix of order 3 × 3, all the diagonal elements are 1, and non-diagonal elements are zero. Hence this diagonal matrix is an Identity Matrix.

## Triangular Matrix

A square matrix in which the non-zero elements form a triangular below and above the diagonal is called a Triangular Matrix. Based on the triangle formed below or above the diagonal, the triangular matrix is classified as:

- Upper Triangular Matrix
- Lower Triangular Matrix

Let's learn them in detail.

## Upper Triangular Matrix

A square matrix in which all the elements below the diagonal are zero and the elements from the diagonal and above are non-zero elements is called an Upper Triangular Matrix. In an Upper Triangular Matrix, the non-zero elements form a triangular-like shape.

### Example of Upper Triangular Matrix

$$\begin{bmatrix} 8 & 5 & 6 \\ 0 & 4 & 7 \\ 0 & 0 & 9 \end{bmatrix}_{3\times3}$$

In the above example of the Upper Triangular Matrix, all the elements below the diagonal are zero.

## Lower Triangular Matrix

A square matrix in which all the elements above the diagonal are zero and the elements from the diagonal and below are non-zero elements is called a Lower Triangular Matrix. In a Lower Triangular Matrix, the non-zero elements form a triangular-like shape from the diagonal and below.

### Example of Lower Triangular Matrix

$$\begin{bmatrix} 8 & 0 & 0 \\ 6 & 4 & 0 \\ 5 & 7 & 9 \end{bmatrix}_{3\times3}$$

In the above example of the lower triangular matrix, all the elements above the diagonal are zero.

## Singular Matrix

A singular matrix is referred to as a square matrix whose determinant is zero and is not invertible. If det A = 0, a square matrix "A" is said to be singular; otherwise, it is said to be non-singular.

### Example of Singular Matrix

$$A = \begin{bmatrix} 3 & 6 & 9 \\ 6 & 12 & 18 \\ 2 & 4 & 6 \end{bmatrix}$$

$\Rightarrow |A| = 3(12\times6 - 18\times4) - 6(6\times6 - 18\times2) + 9(6\times4 - 12\times2)$

$\Rightarrow |A| = 3(72 - 72) - 6(36 - 36) + 9(24 - 24)$

$\Rightarrow |A| = 3\times0 - 6\times0 + 9\times0 = 0$

## Non Singular Matrix

A Non-Singular matrix is defined as a square matrix whose determinant is not equal to zero and is invertible.

### Example of a Non-Singular Matrix

$|A| = \begin{bmatrix} 1 & 5 \\ 9 & 8 \end{bmatrix}$

$\Rightarrow |A| = 8 \times 1 - 9 \times 5 = 8 - 45 = -37$

## Symmetric Matrix

A square matrix "A" of any order is defined as a symmetric matrix if the transpose of the matrix is equal to the original matrix itself, i.e., $A^T = A$.

### Example of Symmetric Matrix

$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

## Skew Symmetric Matrix

A square matrix "A" of any order is defined as a skew-symmetric matrix if the transpose of the matrix is equal to the negative of the original matrix itself, i.e., $A^T = -A$.

### Example of Skew Symmetric Matrix

$\begin{bmatrix} 0 & 3 & 5 \\ -3 & 0 & -2 \\ -5 & 2 & 0 \end{bmatrix}$

## Orthogonal Matrix

A square matrix whose transpose is equal to its inverse is called Orthogonal Matrix. In an Orthogonal Matrix if $A^T = A^{-1}$ then $AA^T = I$ where I is the Identity Matrix.

### Example of Orthogonal Matrix

$A = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix}$

and $A^T = \begin{bmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{bmatrix}$

$\Rightarrow A \times A^T = \begin{bmatrix} \cos^2(\theta) + \sin^2(\theta) & \cos(\theta)\sin(\theta) - \cos(\theta)\sin(\theta) \\ \sin(\theta)\cos(\theta) - \cos(\theta)\sin(\theta) & \cos^2(\theta) + \sin^2(\theta) \end{bmatrix}$

$\Rightarrow A \times A^T = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = I_{(2\times 2)}$

## Idempotent Matrix

An idempotent matrix is a special type of square matrix that remains unchanged when multiplied by itself, i.e., $A^2 = A$.

### Example of Idempotent Matrix

$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$

Hence, $A \cdot A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = A$

## Nilpotent Matrix

A Nilpotent is a square matrix that when raised to some positive power results in zero matrix. The least power let's say 'p' for which the matrix yields zero matrix, then it is called the Nilpotent Matrix of power 'p'.

Example of Nilpotent Matrix

$A = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$

$\Rightarrow A^2 = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix}$

$\Rightarrow A^2 = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

and $A^3 = A \cdot A^2$

$\Rightarrow A^3 = \begin{bmatrix} 0 & 1 & 2 \\ 0 & 0 & 3 \\ 0 & 0 & 0 \end{bmatrix} \cdot \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$

Hence, A is a Nilpotent Matrix of index 3.

### Periodic Matrix

A periodic matrix is a square matrix which exhibits periodicity, i.e. when raised to some power let's say p+1 then $A^{p+1} = A$. If p = 1 then $A^2 = A$ it means A is an Idempotent Matrix. Thus we can say that Idempotent Matrix is a case of Periodic Matrix.

**Example of Periodic Matrix**

$$A = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$$

The above square matrix is a Periodic Matrix of Period 2, where p = 1.

### Involuntary Matrix

An involuntary matrix is a special type of square matrix whose inverse is the original matrix itself, i.e., $P = P^{-1}$, or, in other words, its square is equal to an identity matrix i.e. $P^2 = I$.

**Example of Involuntary Matrix**

$$A = \begin{bmatrix} 2 & 1 \\ -3 & -2 \end{bmatrix}$$

### Hermitian Matrix

A square matrix whose transpose is equal to its conjugate matrix is called Hermitian Matrix.

**Example of Hermitian Matrix**

$$A = \begin{bmatrix} 3+2i & 1-i \\ 1+i & 4 \end{bmatrix}$$

### Skew Hermitian Matrix

A square matrix whose transpose is equal to the negative of its conjugate matrix is called Skew Hermitian Matrix.

**Example of Skew Hermitian Matrix**

$$A = \begin{bmatrix} 0 & 2i & -3i \\ -2i & 0 & 4 \\ 3i & -4 & 0 \end{bmatrix}$$

### Boolean Matrix

The matrix which represents the binary relationship and takes 0 and 1 as its element is called Boolean Matrix.

**Example of Boolean Matrix**

$$\begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \end{bmatrix}$$

### Stochastic Matrix

A Square Matrix which represents the probability data i.e. a non-negative element such that the summation of each row is 1.

**Example of Stochastic Matrix**

$$\begin{bmatrix} 0.2 & 0.5 & 0.3 \\ 0.1 & 0.3 & 0.6 \\ 0.4 & 0.2 & 0.4 \end{bmatrix}$$

## Trace of a Matrix

The sum of diagonal elements of a matrix is known as the trace of a matrix. The Trace of a matrix A can be represented as tr(A). The Trace of a matrix can be calculated for a square matrix only.

**Example:**

$$A = \begin{bmatrix} 15 & 12 & 9 \\ 4 & 6 & 11 \\ 5 & 9 & 0 \end{bmatrix}_{3\times3}$$

***tr(A)** = 15 + 6 + 0 = 21*

# Norms:

Norm is a function that returns length/size of any vector (except zero vector).

Lets assume a vector **x** such that

$$\vec{x} = [x_1, x_2, \ldots\ldots, x_n]$$

For any function *f* to be a norm, it has to satisfy three conditions

**Condition 1**

If norm of x is greater than 0 then x is not equal to 0 (Zero Vector) and if norm is equal to 0 then x is a zero vector.

$$\text{if } f(x) > 0 \text{ then } x \neq 0$$
$$\text{if } f(x) = 0 \text{ then } x = 0$$

---

**Condition 2**

For any scalar quantity, say K

$$f(Kx) = Kf(x)$$

**Condition 3**

Assuming that we have another vector y

$$f(x+y) \leq f(x) + f(y)$$

If these three conditions are satisfied then the function *f* is a norm.

---

**Commonly used Norms**

The most commonly used norms are clubbed under *p-norms* or (*$l_p$-norms*) family, where p is any number greater than or equal to 1.

The p-norm of vector x will be denoted as

$$||x||_p$$

To calculate the p-norm of vector x we have the formula

$$||x||_p = (x_1^p + x_2^p + x_3^p + \ldots\ldots + x_n^p)^{1/p}$$

Every element of vector x is raised to the power p. Then their sum is raised to the power (1/p)

---

It could be re-written in simplified form as

$$||x||_p = (\sum_{i=1}^{n} x_i^p)^{1/p}$$

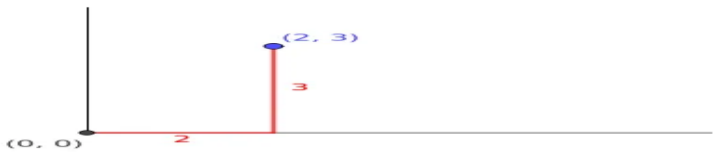
**Manhattan Distance (1-norm)**

1-norm is also called Manhattan distance because it measures distance between two points in a city given that you can only travel along orthogonal city blocks.

Suppose, for a vector **a** we have to calculate 1-norm.

$$\vec{a} = [2, 3]$$

1-norm could be represented on a figure as




Line in red represents 1-norm of vector a

To calculate 1-norm using formula, we could just replace p by 1

$$||a||_1 = (2 + 3)^1$$
$$||a||_1 = 5$$

**Euclidean Norm (2-norm)**

Top hi

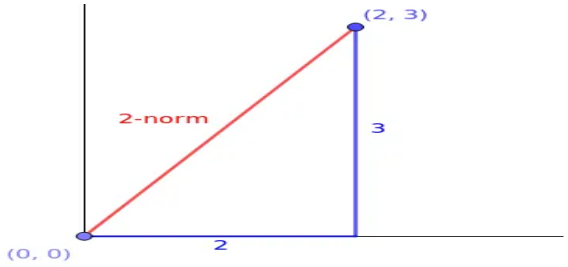
The most used norm within p-norm family is the Euclidean Norm or 2-norm. We have used it earlier to calculate the magnitude of vector.

Euclidean Norm returns the shortest distance between two points.

So, 2-norm of vector **a** will be

$$||a||_2 = (2^2 + 3^2)^{1/2}$$
$$||a||_2 = (4 + 9)^{1/2}$$
$$||a||_2 = \sqrt{13}$$

and to show 2-norm of vector **a** on figure




Line in red represents 2-norm. It is the shortest distance between origin and point represented by vector **a**

**Infinity-norm**

The infinity-norm returns maximum absolute value in the given vector.

Infinity-norm of vector **a** will be

$$||a||_\infty = 3$$

---

Suppose we have to find infinity-norm of another vector, say **b**

$$\vec{b} = [4, 3, -6]$$

then

$$||b||_\infty = 6$$

4 is the largest number in vector **b** but infinity-norm returns maximum **absolute** value

---

## Eigen Value Decomposition

One of the most widely used kinds of matrix decomposition is called Eigen decomposition, in which we decompose a matrix into a set of eigenvectors and eigenvalues. However, we often want to decompose matrices into their eigenvalues and eigenvectors. Doing so can help us to analyze certain properties of the matrix, much as decomposing an integer into its prime factors can help us understand the behavior of that integer. Eigen decomposition can also be used to calculate the principal components of a matrix in the Principal Component Analysis method or PCA which can be used to reduce the dimensionality of data in machine learning.

Condition: $A = \Omega U U^{-1}$

## Eigen Value Decomposition of a matrix

This method converts the given matrix into a combination of Eigen values and Eigen vectors.

Condition :- $A \cdot U = \Omega \cdot U$

$A = \Omega \cdot U \cdot U^{-1}$

Where,

A = The given matrix

U = Combination matrix of eigen vectors with descending order of eigen values.

$\Omega$ = A diagonal matrix which has eigen values as it's diagonal elements.

### Terminology in this technique

Eigen value :- These are special set of scalar values that is associated with the set of linear equations most probably in the matrix equations.

They are also termed as characteristic roots.

Eigen vectors :- This is a vector that is associated with a set of linear equations.

This is also termed as characteristic vector.

### Example problem

* Apply Eigen value Decomposition on $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$.

Solution :- $A = \begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix}$

Finding eigen values

$$\begin{bmatrix} (2-\lambda) & 1 \\ 1 & (2-\lambda) \end{bmatrix} = 0 \quad \text{---} (1)$$

$(2-\lambda)(2-\lambda) - 1 = 0$

$\lambda^2 - 4\lambda + 3 = 0$

$\lambda = 1, 3.$

Substitute $\lambda = 1, 3$ in (1), to find eigen vectors,

$\lambda = 1$

$$\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$x_1 + x_2 = 0$$

$$x_1 = -x_2$$

$$V_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$$

$\lambda = 3$

$$\begin{bmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$\begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$-x_1 + x_2 = 0$$

$$x_1 = x_2.$$

$$V_2 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

$$\Omega = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}, \quad U = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}.$$

According to condition,

$$A = \Omega \cdot U \cdot U^{-1}$$

to find $U^{-1}$,

$$U = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \Longrightarrow U^{-1} = \frac{1}{-2}\begin{bmatrix} -1 & -1 \\ -1 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix}$$

Finally,

$$\begin{bmatrix} 2 & 1 \\ 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 3 \end{bmatrix}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}\begin{bmatrix} 0.5 & 0.5 \\ 0.5 & -0.5 \end{bmatrix}.$$

## Singular Value Decomposition

The Singular Value Decomposition (SVD) of a matrix is a factorization of that matrix into three matrices.

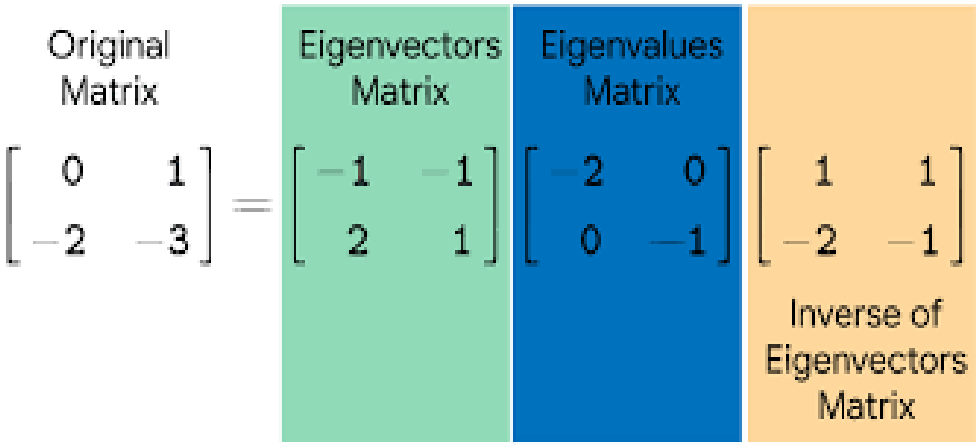
The SVD of matrix A of order mxn can be given by:

$A = U\Sigma V^T$

Where:

- U: Eigenvector matrix of $A.A^T$
- $V^T$: another Eigenvector matrix of $A^T$. A
- $\Sigma$: Eigen value matrix containing the square root of the Eigen values.

SVD is a widely used technique to decompose a matrix into several component matrices, exposing many of the useful and interesting properties of the original matrix.



## Singular Value Decomposition of a matrix

This method can state that a given matrix 'A' is factorized into three sub-matrices.

**Condition :-** $A = U \Sigma V^T$

Where,

A = Given matrix

U = Eigen vector's matrix

→ to obtain this matrix,

I, find $A \cdot A^T$

II, find eigen values & vectors by using following:

$A \cdot A^T - \lambda I = 0$

$\Sigma$ = Eigen value's matrix containing square root of eigen values.

$V^T$ = Another eigen vector's matrix

→ to obtain this matrix,

I, find $A^T \cdot A$

II, find eigen values & vectors by using following

$A^T \cdot A - \lambda I = 0$ to get 'V'

III, perform transpose on matrix 'V'

### Example problem

*Apply Singular Value Decomposition on $A = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$.

**Solution :-** $A = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix}$. $A^T = \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix}$

### finding matrix 'U'

$$A \cdot A^T = \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix} = \begin{bmatrix} 16 & 12 \\ 12 & 34 \end{bmatrix}.$$

$A \cdot A^T - \lambda I = 0$

$$\begin{bmatrix} 16-\lambda & 12 \\ 12 & 34-\lambda \end{bmatrix} = 0 \Rightarrow (16-\lambda)(34-\lambda) - (12)(12) = 0$$

$$\lambda^2 - 50\lambda + 400 = 0$$

$$\lambda = 40, 10.$$

### Eigen Vectors

When $\lambda = 10$,

$$\begin{bmatrix} 6 & 12 \\ 12 & 24 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0.$$

$$6x_1 + 12x_2 = 0$$

$$x_1 + 2x_2 = 0$$

$$x_1 = -2x_2$$

$$V_1 = \begin{bmatrix} -2 \\ 1 \end{bmatrix}.$$

When $\lambda = 40$,

$$\begin{bmatrix} -24 & 12 \\ 12 & -6 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$-24x_1 + 12x_2 = 0$$

$$-2x_1 + x_2 = 0$$

$$2x_1 = x_2$$

$$V_2 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}.$$

now, $U = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix}$ (descending order).

$$\Sigma = \begin{bmatrix} \sqrt{10} & 0 \\ 0 & \sqrt{40} \end{bmatrix}.$$

finding matrix 'VT'

$\because A \cdot A^T \neq A^T \cdot A$

$$A^T \cdot A = \begin{bmatrix} 4 & 3 \\ 0 & -5 \end{bmatrix} \begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} = \begin{bmatrix} 25 & -15 \\ -15 & 25 \end{bmatrix}.$$

$$A^T \cdot A - \lambda I = 0$$

$$\begin{bmatrix} 25-\lambda & -15 \\ -15 & 25-\lambda \end{bmatrix} = 0 \Longrightarrow$$

$$(25-\lambda)(25-\lambda) - (-15)(-15) = 0$$

$$\lambda^2 - 50\lambda + 400 = 0$$

$$\lambda = 10, 40.$$

Eigen vectors

When $\lambda = 10$,

$$\begin{bmatrix} 15 & -15 \\ -15 & 15 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$15x_1 - 15x_2 = 0$$

$$x_1 = x_2$$

$$V_1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

When $\lambda = 40$,

$$\begin{bmatrix} -15 & -15 \\ -15 & -15 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 0$$

$$-15x_1 - 15x_2 = 0$$

$$-x_1 = x_2$$

$$V_2 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

$$V = \begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix} \Longrightarrow V^T = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

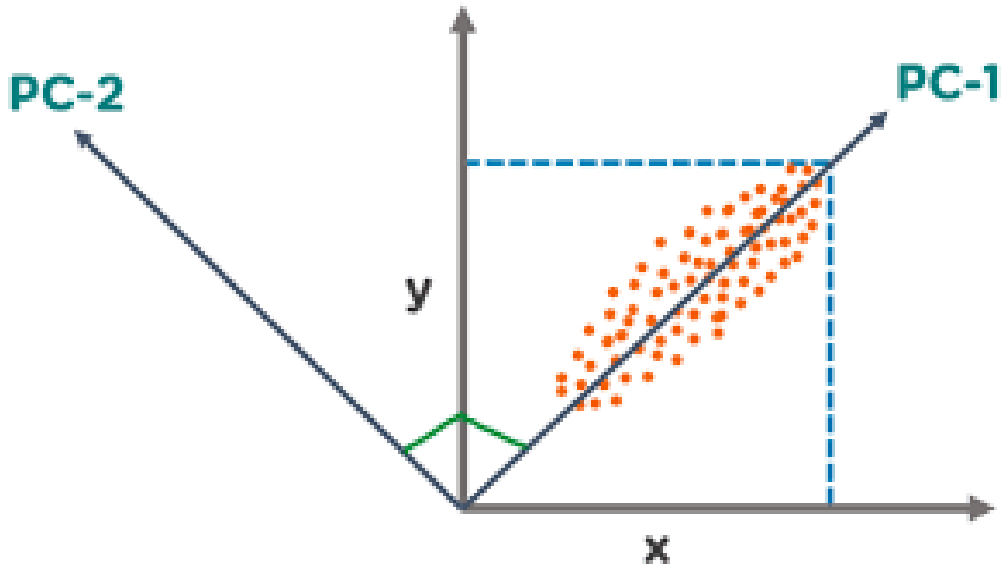
According to condition, $A = U \Sigma V^T$.

$$\begin{bmatrix} 4 & 0 \\ 3 & -5 \end{bmatrix} = \begin{bmatrix} 1 & -2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} \sqrt{10} & 0 \\ 0 & \sqrt{40} \end{bmatrix} \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}.$$

$\therefore$

# Principal Component Analysis

The Principal Component Analysis is a popular unsupervised learning technique for reducing the dimensionality of data. It increases interpretability yet, at the same time, it minimizes information loss. It helps to find the most significant features in a dataset and makes the data easy for plotting in 2D and 3D.




**Dimension Reduction Techniques-**

Principal Component Analysis (PCA)

**Principal Component Analysis-**

PCAworksbysuccessivelyidentifyingtheaxisofgreatestvarianceinadataset(theprincipalcomponents).Itd oesthisasfollows:

1. Identifyingthecenterpointofthedataset.
2. Calculatingthecovariancematrixofthedata.
3. Calculatingtheeigenvectorsofthecovariancematrix.
4. Orthonormalizingtheeigenvectors.
5. Calculatingtheproportionofvariancerepresentedbyeacheigenvector.

- **Covariance** is effectively variance applied to multiple dimensions; it is thevariancebetweentwoormorevariables.Whileasinglevaluecancapturethevariance in one dimension or variable, it is necessary to use a *2 x 2* matrix tocapture the covariance between two variables, a *3 x 3* matrix to capture thecovariance between three variables, and so on. So the first step in PCA is tocalculatethiscovariancematrix.

- An **Eigenvector** is a vector that is specific to a dataset and lineartransformation. Specifically, it is the vector that does not change in directionbefore and after the transformation is performed. To get a better feeling forhowthisworks,imaginethatyou'reholdingarubberband,straight,betweenboth hands. Let's say you stretch the band out until it is taut between yourhands. The eigenvector is the vector that did not change direction betweenbefore the stretch and during it; in this case, it's the vector running directlythrough thecenteroftheband fromonehandtothe other.

- **Orthogonalization** is the process of finding two vectors that areorthogonal(atrightangles)tooneanother.Inann-dimensionaldataspace,the process of orthogonalization takes a set of vectors and yields a set oforthogonalvectors.

- **Orthonormalization**isanorthogonalizationprocessthatalsonormalizestheproduct.

- **Eigenvalue**(roughlycorrespondingtothelengthoftheeigenvector)isusedto calculate the proportion of variance represented by each eigenvector.This is done by dividing the eigenvalue for each eigenvector by the sum ofeigenvaluesforalleigenvectors.

Principal Component Analysis is a well-known dimension reduction technique.It transforms the variables into a new set of variables called as principal components.These principal components are linear combination of original variables and are orthogonal.The first principal component accounts for most of the possible variation of original data.The second principal component does its best to capture the variance in the data.There can be only two principal components for a two-dimensional data set.

**PCA Algorithm-**

The steps involved in PCA Algorithm are as follows-

**Step-01:** Get data.

**Step-02:** Compute the mean vector (μ).

**Step-03:** Subtract mean from the given data.

**Step-04:** Calculate the covariance matrix.

**Step-05:** Calculate the eigen vectors and eigen values of the covariance matrix.

**Step-06:** Choosing components and forming a feature vector.

**Step-07:** Deriving the new data set.

**Problem-01:**

Given data = { 2, 3, 4, 5, 6, 7 ; 1, 5, 3, 6, 7, 8 }.

Compute the principal component using PCA Algorithm.

**OR**

Consider the two dimensional patterns (2, 1), (3, 5), (4, 3), (5, 6), (6, 7), (7, 8).

Compute the principal component using PCA Algorithm.

**OR**

Compute the principal component of following data-

CLASS 1

X = 2 , 3 , 4

Y = 1 , 5 , 3

CLASS 2

X = 5 , 6 , 7

Y = 6 , 7 , 8

**Solution-**

We use the above discussed PCA Algorithm-

**Step-01:**

Get data.

The given feature vectors are-

- $x_1 = (2, 1)$
- $x_2 = (3, 5)$
- $x_3 = (4, 3)$
- $x_4 = (5, 6)$
- $x_5 = (6, 7)$
- $x_6 = (7, 8)$

$$\begin{bmatrix} 2 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 5 \end{bmatrix} \begin{bmatrix} 4 \\ 3 \end{bmatrix} \begin{bmatrix} 5 \\ 6 \end{bmatrix} \begin{bmatrix} 6 \\ 7 \end{bmatrix} \begin{bmatrix} 7 \\ 8 \end{bmatrix}$$

**Step-02:**

Calculate the mean vector (μ).

Mean vector (μ)

= ((2 + 3 + 4 + 5 + 6 + 7) / 6, (1 + 5 + 3 + 6 + 7 + 8) / 6)

= (4.5, 5)

Thus,

$$\text{Mean vector } (\mu) = \begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$$

**Step-03:**

Subtract mean vector (μ) from the given feature vectors.

- $x_1 - \mu = (2 - 4.5, 1 - 5) = (-2.5, -4)$
- $x_2 - \mu = (3 - 4.5, 5 - 5) = (-1.5, 0)$
- $x_3 - \mu = (4 - 4.5, 3 - 5) = (-0.5, -2)$
- $x_4 - \mu = (5 - 4.5, 6 - 5) = (0.5, 1)$
- $x_5 - \mu = (6 - 4.5, 7 - 5) = (1.5, 2)$
- $x_6 - \mu = (7 - 4.5, 8 - 5) = (2.5, 3)$

Feature vectors ($x_i$) after subtracting mean vector (μ) are-

$$\begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 2.5 \\ 3 \end{bmatrix}$$

**Step-04:**

Calculate the covariance matrix.

Covariance matrix is given by-

$$\text{Covariance Matrix} = \frac{\sum (x_i - \mu)(x_i - \mu)^t}{n}$$

Now,

$$m_1 = (x_1 - \mu)(x_1 - \mu)^t = \begin{bmatrix} -2.5 \\ -4 \end{bmatrix} \begin{bmatrix} -2.5 & -4 \end{bmatrix} = \begin{bmatrix} 6.25 & 10 \\ 10 & 16 \end{bmatrix}$$

$$m_2 = (x_2 - \mu)(x_2 - \mu)^t = \begin{bmatrix} -1.5 \\ 0 \end{bmatrix} \begin{bmatrix} -1.5 & 0 \end{bmatrix} = \begin{bmatrix} 2.25 & 0 \\ 0 & 0 \end{bmatrix}$$

$$m_3 = (x_3 - \mu)(x_3 - \mu)^t = \begin{bmatrix} -0.5 \\ -2 \end{bmatrix} \begin{bmatrix} -0.5 & -2 \end{bmatrix} = \begin{bmatrix} 0.25 & 1 \\ 1 & 4 \end{bmatrix}$$

$$m_4 = (x_4 - \mu)(x_4 - \mu)^t = \begin{bmatrix} 0.5 \\ 1 \end{bmatrix} \begin{bmatrix} 0.5 & 1 \end{bmatrix} = \begin{bmatrix} 0.25 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$

$$m_5 = (x_5 - \mu)(x_5 - \mu)^t = \begin{bmatrix} 1.5 \\ 2 \end{bmatrix} \begin{bmatrix} 1.5 & 2 \end{bmatrix} = \begin{bmatrix} 2.25 & 3 \\ 3 & 4 \end{bmatrix}$$

$$m_6 = (x_6 - \mu)(x_6 - \mu)^t = \begin{bmatrix} 2.5 \\ 3 \end{bmatrix} \begin{bmatrix} 2.5 & 3 \end{bmatrix} = \begin{bmatrix} 6.25 & 7.5 \\ 7.5 & 9 \end{bmatrix}$$

Now,

Covariance matrix

$= (m_1 + m_2 + m_3 + m_4 + m_5 + m_6) / 6$

On adding the above matrices and dividing by 6, we get-

$$\text{Covariance Matrix} = \frac{1}{6}\begin{bmatrix} 17.5 & 22 \\ 22 & 34 \end{bmatrix}$$

$$\text{Covariance Matrix} = \begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix}$$

**Step-05:**

Calculate the eigen values and eigen vectors of the covariance matrix.

$\lambda$ is an eigen value for a matrix M if it is a solution of the characteristic equation $|M - \lambda I| = 0$.

So, we have-

$$\begin{vmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{vmatrix} - \begin{vmatrix} \lambda & 0 \\ 0 & \lambda \end{vmatrix} = 0$$

$$\begin{vmatrix} 2.92 - \lambda & 3.67 \\ 3.67 & 5.67 - \lambda \end{vmatrix} = 0$$

From here,

$(2.92 - \lambda)(5.67 - \lambda) - (3.67 \times 3.67) = 0$

$16.56 - 2.92\lambda - 5.67\lambda + \lambda^2 - 13.47 = 0$

$\lambda^2 - 8.59\lambda + 3.09 = 0$

Solving this quadratic equation, we get $\lambda = 8.22, 0.38$

Thus, two eigen values are $\lambda_1 = 8.22$ and $\lambda_2 = 0.38$.

Clearly, the second Eigen value is very small compared to the first eigen value.

So, the second Eigen vector can be left out.

Eigen vector corresponding to the greatest Eigen value is the principal component for the given data set.

Find the Eigen vector corresponding to eigen value $\lambda_1$.

Use the following equation to find the Eigen vector-

$MX = \lambda X$

where-

- M = Covariance Matrix
- X = Eigen vector
- λ = Eigen value

Substituting the values in the above equation, we get-

$$\begin{bmatrix} 2.92 & 3.67 \\ 3.67 & 5.67 \end{bmatrix} \begin{bmatrix} X1 \\ X2 \end{bmatrix} = 8.22 \begin{bmatrix} X1 \\ X2 \end{bmatrix}$$

Solving these, we get-

$2.92X_1 + 3.67X_2 = 8.22X_1$

$3.67X_1 + 5.67X_2 = 8.22X_2$

On simplification, we get-

$5.3X_1 = 3.67X_2$ ……(1)

$3.67X_1 = 2.55X_2$ ……(2)

From (1) and (2), $\mathbf{X_1 = 0.69X_2}$

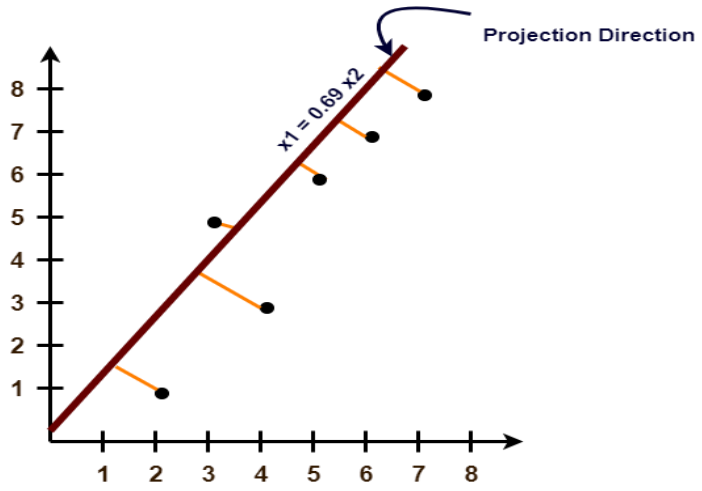
From (2), the eigen vector is-

$$\text{Eigen Vector :} \quad \begin{bmatrix} X1 \\ X2 \end{bmatrix} = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}$$

Thus, principal component for the given data set is-

$$\text{Principal Component :} \quad \begin{bmatrix} X1 \\ X2 \end{bmatrix} = \begin{bmatrix} 2.55 \\ 3.67 \end{bmatrix}$$

Lastly, we project the data points onto the new subspace as-




# PROBABILITY AND INFORMATION THEORY

## Random Variables

A random variable is a variable whose value is unknown or a function that assigns values to each of an experiment's outcomes. A random variable can be either discrete (having specific values) or continuous (any value in a continuous range).

(Or)

A random variable's likely values may express the possible outcomes of an experiment, which is about to be performed or the possible outcomes of a preceding experiment whose existing value is unknown. The domain of a random variable is a sample space, which is represented as the collection of possible outcomes of a random event. For instance, when a coin is tossed, only two possible outcomes are acknowledged such as heads or tails.

### Types of Random Variables

- Discrete Random Variable

  A discrete random variable can take only a finite number of distinct values such as 0, 1, 2, 3, 4, … and so on.

- Continuous Random Variable

If the random variable X takes any value in a given interval (a, b), it is said to be a continuous random variable in that interval.

**Mean of a random variable:** If X is the random variable and P is the respective probabilities, the mean of a random variable is defined by:

**Mean (μ) = ∑ XP**

where variable X consists of all possible values and P consist of respective probabilities.

**Variance of Random Variable:** The variance tells how much the spread of random variable X is around the mean value. The formula for the variance of a random variable is given;

$Var(X) = \sigma^2 = E(X^2) – [E(X)]^2$

where $E(X^2) = \sum X^2P$ and $E(X) = \sum XP$

Standard Deviation(X) = $\sigma$

## Probability Distributions

In Statistics, the **probability distribution** gives the possibility of each outcome of a random experiment or event. It provides the probabilities of different possible occurrences. A function that is used to define the distribution of a probability is called a Probability distribution function.

## Types of Probability Distribution

There are two types of probability distribution which are used for different purposes and various types of the data generation process.

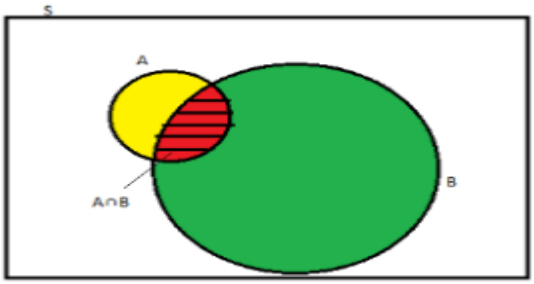
1. Normal or Cumulative Probability Distribution
2. Binomial or Discrete Probability Distribution

## Marginal Probability

Marginal probability is the probability of an event happening, such as (p(A)), and it can be mentioned as an unconditional probability. It does not depend on the occurrence of another event. For example, the likelihood that a card is drawn from a deck of cards is black (P(black) = 0.5), and the probability that a card is drawn is 7 (P (7) =1/13), both are independent events since the outcome of another event does not condition the result of one event.

## Conditional Probability

**Conditional probability** is known as the possibility of an event or outcome happening, based on the existence of a previous event or outcome. Imagine a student who takes leave from school twice a week, excluding Sunday. If it is known that he will be absent from school on Tuesday then what are the chances that he will also take a leave on Saturday in the same week? It is observed that in problems where the occurrence of one event affects the happening of the following event, these cases of probability are known as conditional probability. It is depicted by P(A|B).




As depicted by the above diagram, sample space is given by S, and there are two events A and B. In a situation where event B has already occurred, then our sample space S naturally gets reduced to B because now the chances of occurrence of an event will lie inside B.

As we have to figure out the chances of occurrence of event A, only a portion common to both A and B is enough to represent the probability of occurrence of A when B has already occurred. The common portion of the events is depicted by the intersection of both events A and B, i.e. $A \cap B$.

This explains the concept of conditional probability problems, i.e., the occurrence of any event when another event in relation to has already occurred.

## Conditional Expectation

The conditional expectation (or conditional expected value, or conditional mean) is the expected value of a random variable, computed with respect to a conditional probability distribution.

Let **X** and **Y** be two random variables. The conditional Expectation of **X** given **Y = y** is the weighted average of the values that **X** can take on, where each possible value is weighted by its respective conditional probability (conditional on the information that **Y = y**).

The expectation of a random variable **X** conditional on **Y = y** is denoted by:
$E[X|Y = y]$

## Variance and covariance

In probability, the variance of some random variable X is a measure of how much values in the distribution vary on average with respect to the mean.

$V(x) = \sigma^2 = E(X^2) - E(X)^2$

Where:

$E(X) = \mu = \sum xP(x)$.

$E(X^2) = \sum x^2P(x)$.

Variance is the difference between when we square the inputs to Expectation and when we square the Expectation itself.

**Covariance** is a measure of the relationship between two random variables and to what extent, they change together. Or we can say, in other words, it defines the changes between the two variables, such that change in one variable is equal to change in another variable.

### Types of Covariance

Covariance can have both positive and negative values. Based on this, it has two types:

1. Positive Covariance
2. Negative Covariance

### Positive Covariance

If the covariance for any two variables is positive, that means, both variables move in the same direction. Here, the variables show similar behavior. That means, if the values (greater or lesser) of one variable correspond to the values of another variable, then they are said to be in positive covariance.

### Negative Covariance

If the covariance for any two variables is negative, that means, both variables move in the opposite direction. It is the opposite case of positive covariance, where greater values of one variable correspond to lesser values of another variable and vice-versa.

The formula of the Covariance of two variables x and y is given below:

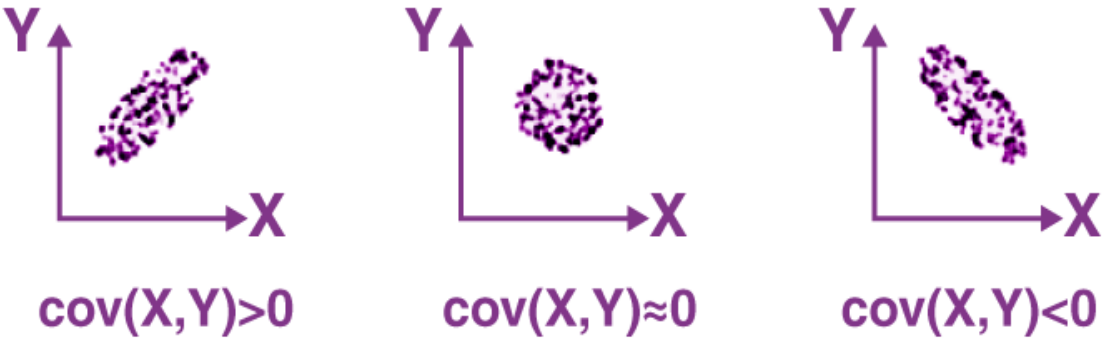
$$Cov(x,y) = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{N}$$

Where,

- $x_i$ = data value of x
- $y_i$ = data value of y
- $\bar{x}$ = mean of x
- $\bar{y}$ = mean of y
- N = number of data values.

## Covariance of X and Y

The below figure shows the covariance of X and Y.




If cov (X, Y) is greater than zero, then we can say that the covariance for any two variables is positive and both the variables move in the same direction.

If cov (X, Y) is less than zero, then we can say that the covariance for any two variables is negative and both the variables move in the opposite direction.

If cov (X, Y) is zero, then we can say that there is no relation between the two variables.

## Bayes' Rule

## Bayes Theorem Statement

Let $E_1, E_2, \ldots, E_n$ be a set of events associated with a sample space S, where all the events $E_1, E_2, \ldots, E_n$ have nonzero probability of occurrence and they form a partition of S. Let A be any event associated with S, then according to Bayes theorem,

$$P(E_i \mid A) = \frac{P(E_i)P(A \mid E_i)}{\sum_{k=1}^{n} P(E_k)P(A|E_k)}$$

for any k = 1, 2, 3, ...., n

## Bayes Theorem Proof

According to the conditional probability formula,

$$P(E_i \mid A) = \frac{P(E_i \cap A)}{P(A)} \ldots (1)$$

Using the multiplication rule of probability,

$$P(E_i \cap A) = P(E_i)P(A \mid E_i) \ldots (2)$$

Using total probability theorem,

$$P(A) = \sum_{k=1}^{n} P(E_k)P(A|E_k) \ldots (3)$$

Putting the values from equations (2) and (3) in equation 1, we get

$$P(E_i \mid A) = \frac{P(E_i)P(A \mid E_i)}{\sum_{k=1}^{n} P(E_k)P(A|E_k)}$$

**Information Theory:**

Information theory is based on probability theory and statistics and often concerns itself with measures of information of the distributions associated with random variables. Important quantities of information are entropy, a measure of information in a single random variable, and mutual information, a measure of information in common between two random variables. Information theory revolves around quantifying how much information is present in a signal. It was originally invented to study sending messages from discrete alphabets over a noisy channel, such as communication via radio transmission. The basic intuition behind information theory is that learning that an unlikely event has occurred is more informative than learning that a likely event has occurred. In the case of deep learning, the most common use case for information theory is to characterize

probability distributions and to quantify the similarity between two probability distributions.

A key measure in information theory is entropy. Entropy quantifies the amount of uncertainty involved in the value of a random variable or the outcome of a random process. For example, identifying the outcome of a fair coin flip (with two equally likely outcomes) provides less information (lower entropy, less uncertainty) than specifying the outcome from a roll of a die (with six equally likely outcomes).

Applications of fundamental topics of information theory include source coding/data compression (e.g. for ZIP files), and channel coding/error detection and correction (e.g. for DSL). The theory has also found applications in other areas, including statistical inference, cryptography, neurobiology, perception, linguistics, thermal physics, molecular dynamics, quantum computing, black holes, information retrieval, intelligence gathering, plagiarism detection, pattern recognition, anomaly detection and even art creation.

## Numerical Computation

### Overflow and Underflow

**Overflow** and **underflow** are both errors resulting from a shortage of space. On the most basic level, they manifest in data types like *integers* and *floating points*. **Overflow errors** come up when working with integers and floating points, and **underflow errors** are generally just associated with floating points.

### Overflow

Overflow indicates that we have done a calculation that resulted in a number larger than the largest number we can represent. Let's look at an example involving unsigned integers.

Let's assume we have an integer stored in 11 bytes. The greatest number we can store in one byte is 255, so let's take that. This is 11111111. Now, suppose we add 2 to it to get 00000010. The result is 257, which is 100000001. The result has 9 bits, whereas the integers we are working with consist of only 8.

What does a computer then do in this scenario? A computer will discard the *most-significant bit (MSB)* and keep the rest.

## Underflow

Underflow is a bit trickier to understand because it has to do with precision in floating points.

The floating-point convention comes up with techniques to represent fractional numbers. When we use these in calculations that result in a smaller number than our least value, we again exceeded our designated space. Without going into details of floating-point representation, we can see how this problem would manifest by considering a decimal example.

Suppose we are given designated boxes to write decimal numbers in. We have one box on the left of the decimal point and three boxes on the right. So, we can easily represent 0.004. Now, we want to perform a calculation, of 0.004×0.004. The answer to this is 0.000016, but we simply do not have these many places available to us. So, we discard the **least-significant** bits and store 0.000, which is quite obviously a wrong answer.

## Gradient-Based Optimization

In Neural Networks, we have the concept of Loss Functions, which tells us about the performance of our neural networks i.e., at the current instant, how good or poor the model is performing. Now, to train our network such that it performs better on unseen datasets, we need to take the help of the loss. Essentially, our objective is to minimize the loss, as a lower loss implies that our model is going to perform better. So, Optimization means minimizing (or maximizing) any mathematical expression.

Optimizers are algorithms or methods used to update the parameters of the network such as weights, biases, etc. to minimize the losses. Therefore, Optimizers are used to solve optimization problems by minimizing the function i.e., loss function in the case of neural networks.

Instances of Gradient-Based Optimizers
Different instances of Gradient descent-based Optimizers are as follows:

- Batch Gradient Descent or Gradient Descent (GD)
- Stochastic Gradient Descent (SGD)
- Mini batch Gradient Descent (MB-GD)

**Batch Gradient Descent** Batch Gradient Descent sums the error for each point in a training set, updating the model only after all training examples have been evaluated. This process is referred to as a training epoch.

### Stochastic Gradient Descent

Stochastic Gradient Descent runs a training epoch for each example within the dataset and it updates each training example's parameters one at a time.

### Mini – Batch Gradient Descent

Mini – Batch Gradient Descent combines concepts from both Batch and Stochastic Gradient Descent. It splits the training dataset into small batch sizes and performs updates on each of those batches. This approach strikes a balance between the computational efficiency of the Batch Gradient Descent and the speed of the Stochastic Gradient Descent.

## Constrained Optimization

In mathematical optimization, **constrained optimization** (in some contexts called **constraint optimization**) is the process of optimizing an objective function with respect to some variables in the presence of constraints on those variables. The objective function is either a cost function or energy function, which is to be minimized, or a reward function or utility function, which is to be maximized. Constraints can be either **hard constraints**, which set conditions for the variables that are required to be satisfied, or **soft constraints**, which have some variable values that are penalized in the objective function if, and based on the extent that, the conditions on the variables are not satisfied.

## Linear Least Squares

The least square method is the process of finding the best-fitting curve or line of best fit for a set of data points by reducing the sum of the squares of the offsets (residual part) of the points from the curve. It is quite obvious that the fitting of curves for a particular data set is not always unique. This is known as the best-fitting curve and is found by using the least-squares method.

The method of least squares actually defines the solution for the minimization of the sum of squares of deviations or the errors in the result of each equation. The least-squares method is often applied in data fitting.

In this topic, we can discuss Linear Regression and Least Squares method.

## Linear Regression Method

# Linear Regression and least squares method

When working with linear regression, our main goal is to find the best fit line that means the error between predicted values and actual values should be minimized. The best fit line will have the least error.

## Example problem on linear Regression

*perform linear regression on following data.

| x | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| y | 4 | 12 | 28 | 52 | 80. |

**Solution :-**

find line equation,

$m = \frac{y_2 - y_1}{x_2 - x_1}$ (Slope) (let take $x_1 = 1$ $y_1 = 4$ $x_5 = 5$ $y_5 = 80$)

$= \frac{80-4}{5-1} = \frac{76}{4} = 19$

for intercept c,

$(y - y_1) = m(x - x_1)$

$y - 4 = 19(x - 1)$

$y = 19x - 19 + 4$

$y = 19x + (-15)$.

Predicted values of y.

$y_1' = 19x - 15 = 19(1) - 15 = 4$

$y_2' = 19(2) - 15 = 23$

$y_3' = 19(3) - 15 = 42$

$y_4' = 19(4) - 15 = 61$

$y_5' = 19(5) - 15 = 80.$

accuracy $= y_i - y_i'$

$y_1 - y_1' = 4 - 4 = 0$

$y_2 - y_2' = 12 - 23 = -11$

$y_3 - y_3' = 28 - 42 = -14$

$y_4 - y_4' = 52 - 61 = -9$

$y_5 - y_5' = 80 - 80 = 0$

$\Sigma (\text{error})^2 = 0 + 121 + 196 + 81 + 0 = 398.$

# Least Squares Method

The least Square method is the process of finding the best fitting curve or line of best fit for a set of data points by reducing the sum of squares of offsets (residual part) of the points from the curve.

* Example problem on least squares

* perform least Square method on following data.

| x | 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|---|
| y | 4 | 12 | 28 | 52 | 80. |

Solution :-

mean of $x = \bar{x} = \frac{15}{5} = 3$

mean of $y = \bar{y} = \frac{176}{5} = 35.2.$

| x | y | $(x-\bar{x})$ | $(y-\bar{y})$ | $(x-\bar{x})^2$ | $(x-\bar{x})(y-\bar{y})$ |
|---|---|---|---|---|---|
| 1 | 4 | -2 | -31.2 | 4 | 62.4 |
| 2 | 12 | -1 | -23.2 | 1 | 23.2 |
| 3 | 28 | 0 | -7.2 | 0 | 0 |
| 4 | 52 | 1 | 16.8 | 1 | 16.8 |
| 5 | 80 | 2 | 44.8 | 4 | 89.6 |
| | | 0 | | 10 | 192 |

Slope $= m = \frac{\sum (x-\bar{x})(y-\bar{y})}{(x-\bar{x})^2} = \frac{192}{10} = 19.2.$

$c = \bar{y} - m(\bar{x})$

$= 35.2 - 19.2(3) = -22.4$

So, line equation is $y = 19.2x - 22.4$

predicted values of y

$y_1' = 19.2 - 22.4 = -3.2$

$y_2' = 19.2(2) - 22.4 = 16$

$y_3' = 19.2(3) - 22.4 = 35.2$

$y_4' = 19.2(4) - 22.4 = 54.4$

$y_5' = 19.2(5) - 22.4 = 73.6$

error (actual value − predicted value)

$y_1 - y_1' = 4 + 3.2 = 7.2$

$y_2 - y_2' = 12 - 16 = -4$

$y_3 - y_3' = 28 - 35.2 = -7.2$

$y_4 - y_4' = 52 - 54.4 = -2.4$

$y_5 - y_5' = 80 - 73.6 = 6.4.$

$\Sigma(\text{error})^2 = 51.84 + 16 + 51.84 + 5.76 + 40.96$

$= 166.4.$