UNIT II	**Linear Algebra & Statistical Modeling for Data Science** Linear Algebra for Data science, Solving Linear Equations, Linear Algebra - Distance, hyperplanes and half spaces, Eigen values, Eigenvectors, Statistical Modeling, Random Variables and Probability Mass/Density Functions, Sample Statistics, descriptive statistics,notion of probability, distributions,Mean, variance, Covariance, Hypotheses Testing, Type 1 and Type 2 errors. Testing for parameters of a normal distribution and for percentages based on a single sample and based on two samples. Introduction to the chisquared test. The concept of p-value. Mean-square estimation and Kalman filtering.	**06 - Hrs**

Recap

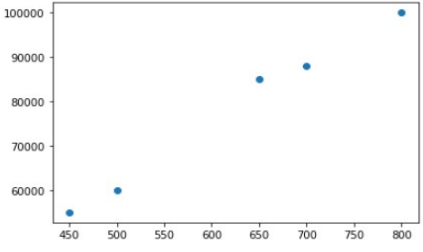
- What is Linear Algebra
- Building Blocks of Linear Algebra
- What is Scalar?
- Why Linear Algebra in DS?
- Application of Linear Algebra in DS
- Linear Algebra Example

Introduction – Linear Equations

- Consider a scenario where you need to predict the **price of a house** given its area in square feet. The prediction can be made based on historical data of the houses sold in that region as given belo

	Area (in Sq. ft.)	Price (US Dollar)
0	500	60000
1	800	100000
2	650	85000
3	450	55000
4	700	88000

- Predicting the price of a house can be achieved by finding the relationship between **Area** and **Price** from the given data.
- The following is a scatter plot of Area and Price:



Introduction – Linear Equations

- This plot indicates that there is a linear relationship between the variables Area and Price. And the linear relationship is represented by an expression as shown below:

  $Y = \beta 0 + \beta 1x1 + \epsilon$ ------- (1)

  where,

  Y is the target variable **'Price'**

  x1 is the variable **'Area'**

  β0, β1 are coefficients

  ϵ is the random error

- When there are more independent variables in the dataset, the target variable can be in general represented as: $Y = \beta 0 + \beta 1x1 + \ldots + \beta rxr + \epsilon$
- To find the optimal values of the coefficients β0 and β1, you can proceed as follows:

  1. Find the sum of squares of the errors as given below:

$$SS = \sum_{i=1}^{n} (Y_i - A - Bx_i)^2$$

- In the above expression, A and B are estimators of the coefficients β0 and β1.

Solving a System of Linear Equations

- Solving a System of Linear Equations implies finding the value of the variables such that each of the equations is satisfied.
- Consider a System of Linear Equations as shown below:

$$x_1 + 3x_2 - x_3 = 4$$
$$2x_1 + 5x_2 + 4x_3 = 19$$
$$2x_1 + 3x_2 - x_3 = 7$$

- This System of Linear Equations can be represented in the form Ax = b as:

$$\begin{bmatrix} 1 & 3 & -1 \\ 2 & 5 & 4 \\ 2 & 3 & -1 \end{bmatrix} * \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = \begin{bmatrix} 4 \\ 19 \\ 7 \end{bmatrix}$$

- Where,

$$A = \begin{bmatrix} 1 & 3 & -1 \\ 2 & 5 & 4 \\ 2 & 3 & -1 \end{bmatrix}, b = \begin{bmatrix} 4 \\ 19 \\ 7 \end{bmatrix}, x = \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix}$$

- In general, the matrix A is of dimension (m, n), x is a column vector of dimension n and b is a column vector of dimension m. A is called the Coefficient matrix and the matrix [A b] is called the Augmented matrix.

Solving a System of Linear Equations

$$[A\ b] = \begin{bmatrix} 1 & 3 & -1 & 4 \\ 2 & 5 & 4 & 19 \\ 2 & 3 & -1 & 7 \end{bmatrix}$$

- In general, the matrix A is of dimension (m, n), x is a column vector of dimension n and b is a column vector of dimension m. A is called the Coefficient matrix and the matrix [A b] is called the Augmented matrix.
- Note: For a given coefficient matrix A of dimension (m, n), if m = n and A is a Full Rank matrix, then, there exists precisely one solution for the System of Linear Equations.

- In the previous example of the System of Linear Equations, dimension of A is (3, 3) and its rank is 3. Therefore, there exists precisely one solution for the Linear Equations.

Solving a System of Linear Equations

$$[A\ b] = \begin{bmatrix} 1 & 3 & -1 & 4 \\ 2 & 5 & 4 & 19 \\ 2 & 3 & -1 & 7 \end{bmatrix}$$

- In general, the matrix A is of dimension (m, n), x is a column vector of dimension n and b is a column vector of dimension m. A is called the Coefficient matrix and the matrix [A b] is called the Augmented matrix.
- Note: For a given coefficient matrix A of dimension (m, n), if m = n and A is a Full Rank matrix, then, there exists precisely one solution for the System of Linear Equations.

- In the previous example of the System of Linear Equations, dimension of A is (3, 3) and its rank is 3. Therefore, there exists precisely one solution for the Linear Equations.

hyperplane in linear algebra?

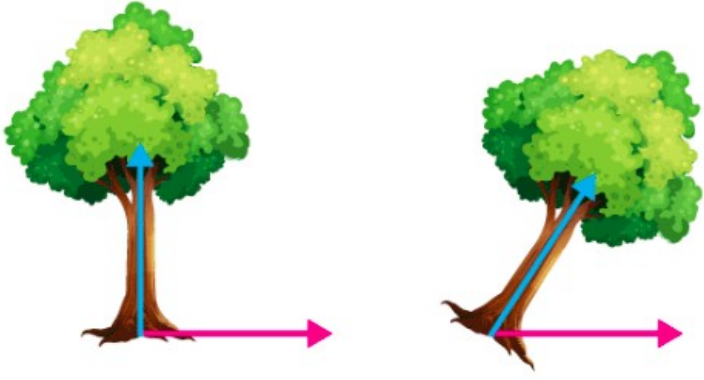
- A hyperplane is a higher-dimensional generalization of lines and planes. The equation of a hyperplane is $w \cdot x + b = 0$, where w is a vector normal to the hyperplane and b is an offset. ... If $y > 0$, then x is on one side of the hyperplane, and if $y < 0$, then x is on the other side of the hyperplane.
- **For example**, in two-dimensional space a hyperplane is a straight line, and in three-dimensional space, a hyperplane is a two-dimensional subspace. Imagine a knife cutting through a piece of cheese that is in cubical shape and dividing it into two parts.
- **How do you calculate hyperplane?**
- A hyperplane is a higher-dimensional generalization of lines and planes. The equation of a hyperplane is $w \cdot x + b = 0$, where w is a vector normal to the hyperplane and b is an offset.

Eigenvalue

- Eigenvalues are the special set of scalars associated with the system of linear equations. It is mostly used in matrix equations. ‘Eigen’ is a German word that means ‘proper’ or ‘characteristic’. Therefore, the term eigenvalue can be termed as characteristic value, characteristic root, proper values or latent roots as well.
- In simple words, the eigenvalue is a scalar that is used to transform the eigenvector.
- The basic equation is $Ax = \lambda x$,
- In Mathematics, an eigenvector corresponds to the real non zero eigenvalues which point in the direction stretched by the transformation whereas eigenvalue is considered as a factor by which it is stretched. In case, if the eigenvalue is negative, the direction of the transformation is negative.
- For every real matrix, there is an eigenvalue. Sometimes it might be complex. The existence of the eigenvalue for the complex matrices is equal to the fundamental theorem of algebra.

EigenValue Example

- In this shear mapping, the **blue arrow** changes direction, whereas the pink arrow does not. Here, the pink arrow is an eigenvector because it does not change direction. Also, the length of this arrow is not changed; its eigenvalue is **1.**



Properties of Eigenvalues

- Eigenvectors with Distinct Eigenvalues are Linearly Independent
- Singular Matrices have Zero Eigenvalues
- If A is a square matrix, then $\lambda = 0$ is not an eigenvalue of A
- For a scalar multiple of a matrix: If A is a square matrix and λ is an eigenvalue of A. Then, $a\lambda$ is an eigenvalue of aA.
- For Matrix powers: If A is square matrix and λ is an eigenvalue of A and $n \geq 0$ is an integer, then λ^n is an eigenvalue of A^n.
- For polynomials of matrix: If A is a square matrix, λ is an eigenvalue of A and p(x) is a polynomial in variable x, then $p(\lambda)$ is the eigenvalue of matrix p(A).
- Transpose matrix: If A is a square matrix, λ is an eigenvalue of A, then λ is an eigenvalue of A^t
- Inverse Matrix: If A is a square matrix, λ is an eigenvalue of A, then λ^{-1} is an eigenvalue of A^{-1}

Introduction – Statistics Model

- The health and welfare council wants to survey your city to understand the lifestyle of the residents and if need be, improve the facilities provided based on the conclusions drawn from the data collected.
- Some of the data being collected for this analysis is listed below:

  1. Personal information such as name, age, income of household and educational qualifications.

  2. Quality of water supply across the city.

  3. Medical facilities such as availability of hospitals, doctors etc.
- The collection, analysis, interpretation, presentation and organization of data is termed as **statistics.**

Introduction – Statistics Model

- Statistics is used to study a population/process (data).
- In the example cited earlier, the data is personal information of residents, water being supplied and information about the medical facilities.
- We can leverage the analysis of this data to answer a few questions which could be but not limited to,
  1. What is the average income of households in the city?
  2. Which is the most common disease in a given area?
  3. How much water does a household use in a month?
  4. How good is the water being supplied?
  5. Are there enough medical facilities being provided?
- When all data required for observation/analysis is collected and studied, the data is referred to as the population.
- On the other hand, when limited data is being collected/analyzed, this data is referred to as sample and is used as an indicative of the entire population.

Types of Statistics

- **Example:**
- **Population:** Data is being collected from all the residents of the city about their age, educational qualification and medical history.
- **Sample:** On the other hand to analyze the water supply, a few litres of water are being collected from different water supply lines that run in the city.
- Statistics can be broadly classified into descriptive and inferential statistics.

1. Descriptive statistics:

- Summarization of data to describe the main features of the sample.
- **For example,** in the survey cited earlier, the knowledge of descriptive statistics can be leveraged to answer some questions like:

1. What is the average income of households in the city?
2. Which is the most common disease in a given area?

Types of Statistics

2. Inferential statistics:

- When working with samples of the population the techniques and processes we use to draw conclusions come under inferential statistics
- For example, in the survey cited earlier, inferential statistics can be used to answer the following questions:

  1. How good is the water being supplied?

  2. Are there enough medical facilities being provided?

Statistical Modeling

- Statistical modeling is the process of applying statistical analysis to a dataset. A statistical model is a mathematical representation (or mathematical model) of observed data.
- When data analysts apply various statistical models to the data they are investigating, they are able to understand and interpret the information more strategically.
- Rather than sifting through the raw data, this practice allows them to identify relationships between variables, make predictions about future sets of data, and visualize that data so that non-analysts and stakeholders can consume and leverage it.
- When you analyze data, you are looking for patterns,"Say Hello". You are using a sample to make an inference about the whole.

Reasons to Learn Statistical Modeling

- While **data scientists** are most often tasked with building models and writing algorithms, analysts also interact with statistical models in their work on occasion.
- For this reason, **analysts** who are looking to excel should aim to obtain a solid understanding of what makes these models successful.
- As machine learning and artificial intelligence become more commonplace, more and more companies and organizations are leveraging **statistical modeling** in order to make predictions about the future data.
- Below are some of the benefits that come from having a thorough understanding of statistical modeling.

  1. You will be better equipped to choose the right model for your needs.
  2. You will be better able to prepare your data for analysis.
  3. You will become a better communicator.

Techniques in Statistical Modeling

- There are several statistical modeling techniques used during data exploration/analysis/ prediction etc,

  1. Linear Regression
  2. Classification
  3. Tree-Based Methods
  4. Supervised Learning
  5. Unsupervised Learning
  6. Neural Networks