Lecture-1
Course: Applied Data Science

Linear Algebra

By

Dr. Sibarama Panigrahi

Assistant Professor, Department of Computer Sc. & Engineering
National Institute of Technology, Rourkela, Odisha, 769008, India
Mobile No.: +91-7377302566
Email: panigrahis[at]nitrkl[dot]ac[dot]in
panigrahi[dot]sibarama[at]gmail[dot]com

Outlines…

- Course Objectives
- Course Outcomes
- Course Contents
- References
- Linear Algebra

Course Objectives

- To extract valuable information for use in strategic decision making, product development, trend analysis, and forecasting.
- Apply quantitative modeling and data analysis techniques to the solution of real world business problems, communicate findings, and effectively present results using data visualization techniques.
- Employ cutting edge tools and technologies to analyze Big Data.

Course Outcomes

- Develop in depth understanding of the key technologies in data science and business analytics: data mining, machine learning, visualization techniques, predictive modeling, and statistics.
- Practice problem analysis and decision-making.
- Gain practical, hands-on experience with statistics programming languages and big data tools through coursework and applied research experiences.
- Students will apply data science concepts and methods to solve problems in real-world contexts and will communicate these solutions effectively.

Course Contents

- **Module 1:**
  - Mathematical Foundation: Linear Algebra and Vector Calculus, Probability and Statistics.
  - Introduction and Motivation to Data Science, Data: Definition, Types and Facets of Data, Data Quality Data Preprocessing: Aggregation, Sampling, Dimensionality reduction, Feature subset selection, Feature creation, Discretization and Binarization, Variable transformation, Measures of Similarity and Dissimilarity, Data science process.
  - Data Warehousing: Data Preprocessing, Warehouse Architecture, ETL, OLAP, Data Lakes, Big Data Pipeline.

- **Module 2:**
  - Descriptive Statistics: Introduction, Data Preparation, Exploratory data analysis: Summarizing Data and Plotting, Scatter plot, Pair plot, Histogram, Probability Density Function (PDF), Univariate analysis using PDF, Cumulative distribution function (CDF), Percentiles and Quantiles, Inter-Quartile Range (IQR), MAD (Median Absolute Deviation), Box-plot with whiskers, Violin plots, Univariate, Bivariate, and Multivariate analysis, contour plot, Outlier Treatment, Measuring Asymmetry: Skewness and Pearson's Median Skewness Coefficient, Kernel Density Estimation: Sample and Estimated Mean, Variance and Standard Scores, Covariance, and Pearson's and Spearman's Rank Correlation.
  - Predictive Modeling: Regression, Decision Tree, Support Vector Machine (SVM), Ensemble Models: Bagging, Boosting.
  - Deep Learning: Introduction to Neural Network, Exploding and Vanishing Gradient Problem, Dropout, Regularization, Weight Initialization, Batch Normalization, Optimizers. Convolutional Neural Network (CNN): Convolution, Padding, Strides, Pooling, Convolution over RGB Images, Transfer Learning, Fusion: Early Fusion (Feature Level Fusion), Late Fusion (Decision Level Fusion) and Intermediate Fusion. Recurrent Neural Network (RNN), Long Short Term Neural Network (LSTM), Gated Recurrent Unit (GRU).

Course Contents

- **Module 3:**
  - Time Series Data Analysis: Introduction to Time Series, Univariate Time Series Forecasting using Statistical, Deep Learning and Hybrid Models: Autoregressive Moving Average (ARMA) models, Autoregressive Integrated Moving Average (ARIMA) models, Seasonal ARIMA (SARIMA), Exponential Smoothing, LSTM, GRU, Hybrid statistical and deep learning models. Multivariate Time Series Forecasting, Point Forecasting, Interval Forecasting, Spatial-Temporal Forecasting.
  - Text Mining and Analytics: Introduction, Data Cleaning, Text Mining Techniques: Stemming, Stop-Word Removal, Tokenization, Lemmatization, Uni-gram, Bi-gram, n-gram, tf-idf, Word2Vec, Bag of Words, Case Study.
- **Module 4:**
  - Descriptive Modelling: k-means, Hierarchical, DBScan clustering.
  - Recommender System: Introduction, Content-Based Filtering, Collaborative Filtering, Hybrid Recommenders, Modelling User Preferences, Evaluating Recommenders.
  - Data Visualization and Condensation: Introduction to Data Visualization, Basic charts and dashboard, Dimensions and Measures, Visual analytics, Dashboard design & principles, Advanced design components/ principles: Enhancing the power of dashboards, Special chart types.

Reference Books

Essential Reading

- Pang-Ning Tan, Michael Steinbach, Vipin Kumar, *Introduction to Data Mining*, Pearson.
- Laura Igual and Santi Seguí, *Introduction to Data Science*, Springer

Supplementary Reading

- Davy Cielin, Arno Meysman, Mohamed Ali, *Introducing Data Science*, Manning
- Andreas, *Practical Data Science*, Apress

Linear Algebra

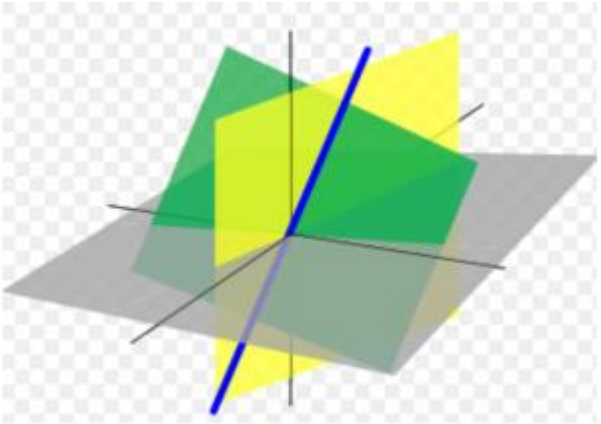
- Linear algebra is the branch of mathematics concerning linear equations such as

  $a_1x_1+ a_2x_2\ldots..+a_nx_n=b$

  - By default, Vectors are denoted as Column Vectors.
  - In vector notation we say $\boldsymbol{a}^{\mathrm{T}}\boldsymbol{x}=b$

  $$[a_1, a_2, \ldots a_n]\begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}=b$$

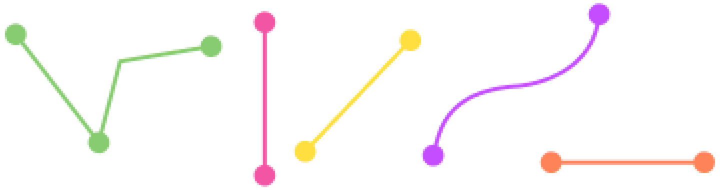
  - Called a linear transformation of ***x***



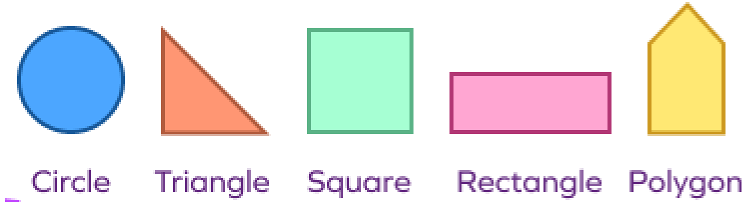
- **Linear algebra is fundamental to geometry, for defining objects such as lines, planes, rotations.**

Why Linear Algebra?

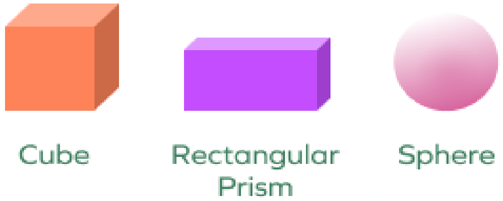
- Linear Algebra provides us the mathematical tool to understand in lower dimensions (2-D/3-D) and generalize for higher dimensions (n-D).
- 0-Dimensional : . (dot)
- 1-Dimensional



- 2-Dimensional

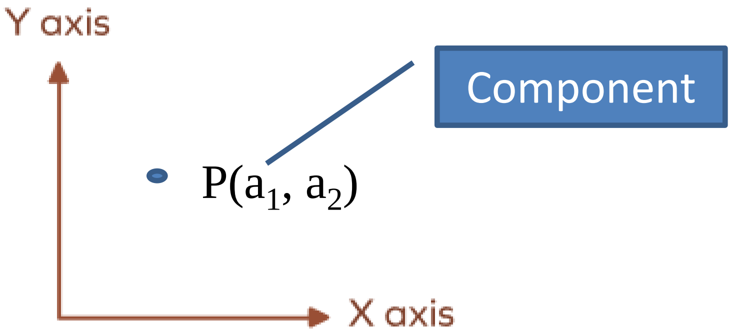


- 3-Dimensional

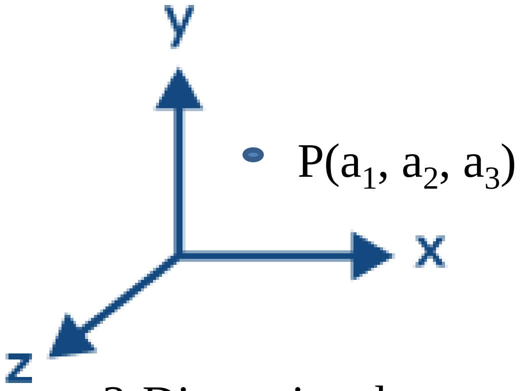


- n-Dimensional: Hypersphere, Hyperplane, Hypercube,…

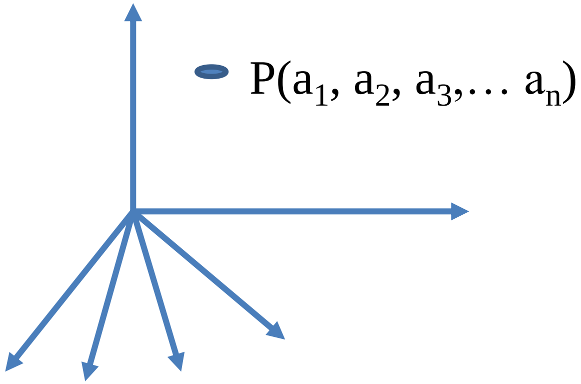
Point (Vector)



2-Dimensional Space

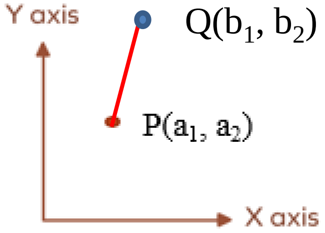


3-Dimensional Space



n-Dimensional Space

Distance between two Points

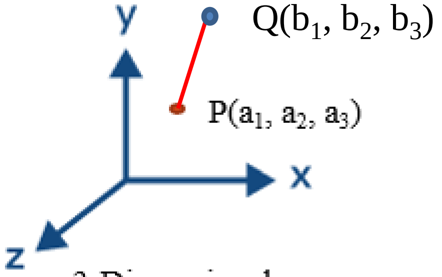


Q(b_1, b_2)

P(a_1, a_2)

2-Dimensional Space

$$D_{PQ} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2}$$

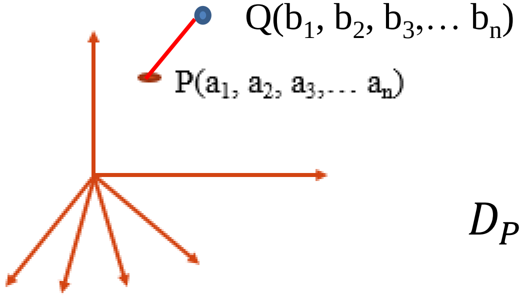


Q(b_1, b_2, b_3)

P(a_1, a_2, a_3)

3-Dimensional Space

$$D_{PQ} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + (a_3 - b_3)^2}$$



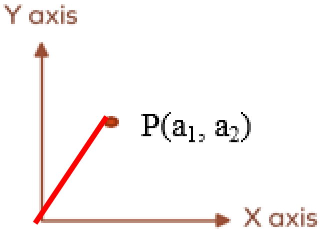
Q(b_1, b_2, b_3,… b_n)

P(a_1, a_2, a_3,… a_n)

n -Dimensional Space

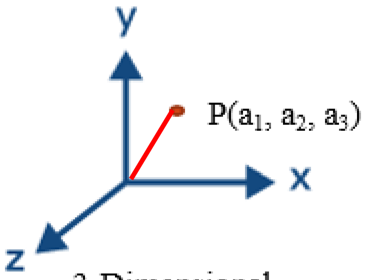
$$D_{PQ} = \sqrt{(a_1 - b_1)^2 + (a_2 - b_2)^2 + \cdots + (a_n - b_n)^2} = \sqrt{\sum_{i=1}^{n} (a_i - b_i)^2}$$

Distance of a Point from Origin



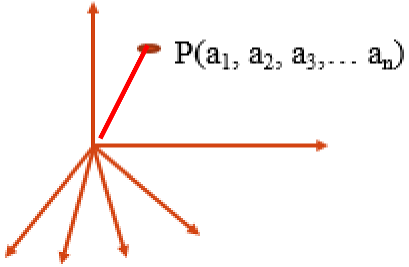
2-Dimensional Space

$$D = |P| = \sqrt{(a_1-0)^2+(a_2-0)^2} = \sqrt{a_1^2+a_2^2}$$



3-Dimensional Space

$$D = |P| = \sqrt{a_1^2+a_2^2+a_3^2}$$



n -Dimensional Space

$$D = |P| = \sqrt{a_1^2+a_2^2+a_3^2+\cdots+a_n^2} = \sqrt{\sum_{i=1}^{n} a_i^2}$$

Vector Operations

- $a = [a_1, a_2, \dots, a_n]$
- $b = [b_1, b_2, \dots, b_n]$
- Addition: $a + b = [a_1 + b_1, a_2 + b_2, \dots, a_n + b_n]$
- Subtraction: $a - b = [a_1 - b_1, a_2 - b_2, \dots, a_n - b_n]$
- Multiplication:
  - Dot Product:$a.b = [a_1 b_1 + a_2 b_2 + \cdots, a_n b_n]$

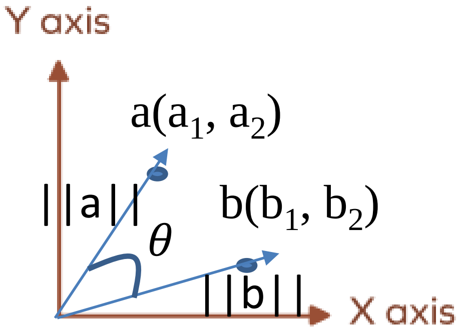
$$\mathrm{a.b} = [a_1, a_2, \dots a_n]\begin{bmatrix} b_1 \\ : \\ b_n \end{bmatrix} = a^T b$$

  - Cross Product (Not much used in Data Science)

Vector Operations

- Dot Product: (Visualize in 2-D)
  - $a.b = \|a\|\|b\|\cos(\theta)$ **[Geometry Definition]**
  - $Where\ \|a\| = \sqrt{a_1^2 + a_2^2}$ =*distance of a from Origin*
  - $a.b = a_1 b_1 + a_2 b_2$ **[Algebra Definition]**
  - The angle between two vectors=$\theta = \cos^{-1}\left(\frac{a.b}{\|a\|\|b\|}\right)$

$$\theta = \cos^{-1}\left(\frac{a_1 b_1 + a_2 b_2}{\|a\|\|b\|}\right)$$



When $\theta = 90^o$ ➔ $\cos(90) = 0$ ➔ *a.b=0*

Vector Operations

- Dot Product: (In n-D)
  - $a.b = \|a\|\|b\|\cos(\theta)$
  - $Where\ \|a\| = \sqrt{\sum_{i=1}^{n} a_i^2}$ =*distance of a from Origin*
  - $a.b = a_1b_1 + a_2b_2 + \cdots + a_nb_n = \sum_{i=1}^{n} a_i\, b_i$
  - The angle between two vectors=$\theta = \cos^{-1}\left(\frac{a.b}{\|a\|\|b\|}\right)$

$$\theta = \cos^{-1}\left(\frac{\sum_{i=1}^{n} a_i b_i}{\|a\|\|b\|}\right)$$

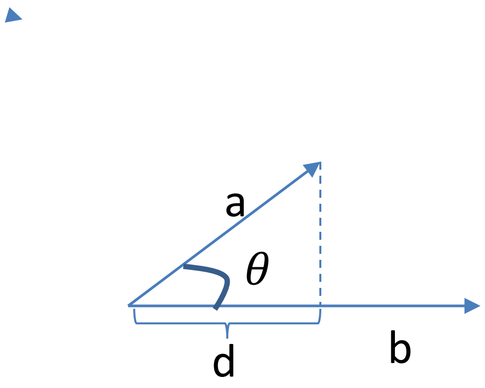
When $\theta = 90^o$ ➔ $\cos(90) = 0$ ➔ *a.b=0*

$$a.a = a_1a_1 + a_2a_2 + \cdots + a_na_n = \sum_{i=1}^{n} a_i^2 = \|a\|^2$$

Dot product between two same vectors =(distance from Origin)^2

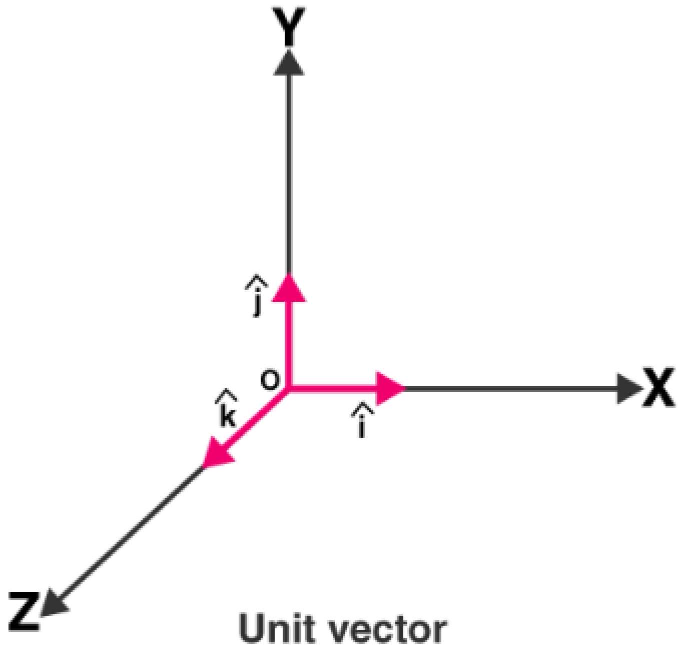
Projection

- $\cos(\theta) = \frac{p}{h} = \frac{d}{\|a\|}$
- $Projection\ of\ a\ on\ b\ i.e.\ d = \|a\|\cos(\theta)$
- $d = \frac{a.b}{\|b\|} = \frac{\|a\|\|b\|\cos(\theta)}{\|b\|} = \|a\|\cos(\theta)$



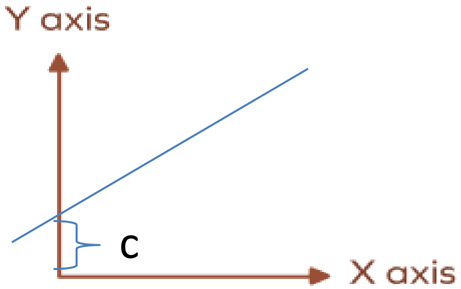
Unit Vector

- A vector is a quantity that has both magnitude, as well as direction.
- A vector that has a magnitude of 1 is a **unit vector**. It is also known as **Direction Vector**.
- Unit vector $\hat{a} = \frac{a}{\|a\|}$



Unit vector

Line (2-D), Plane (3-D) & Hyperplane(n-D)

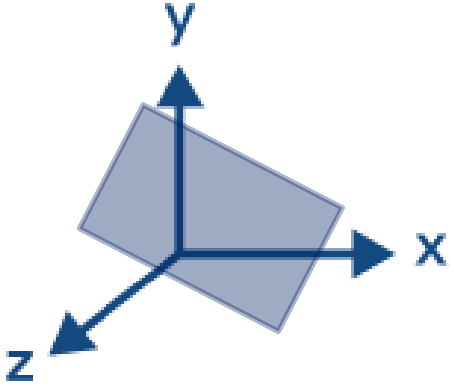


- 2-D:Line: $y = mx + c$

$$ax + by + c = 0$$

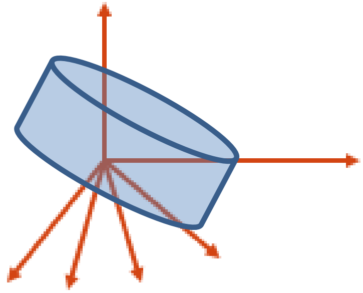
$$w_1x_1 + w_2x_2 + w_0 = 0$$

Note: *When* $w_0 = 0$, *the line passes through the origin.*



- 3-D: Plane:

$$w_1x_1 + w_2x_2 + w_3x_3 + w_0 = 0$$



n-Dimensional Space

- n-D:Hyperplane
- $w_1x_1 + \cdots + w_nx_n + w_0 = 0$

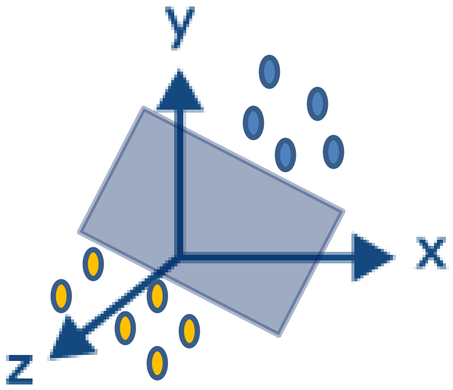
  $w.x + w_0 = 0$

When $w_0 = 0$, the hyperplane passes through origin.

Line (2-D), Plane (3-D) & Hyperplane(n-D)

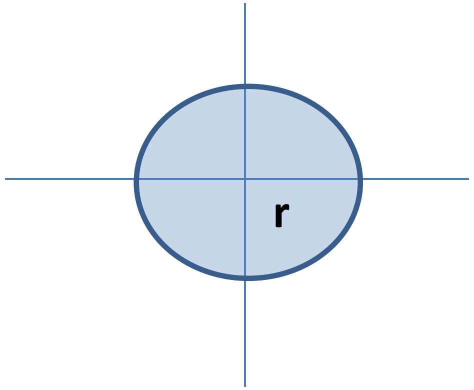
- A plane breaks a plane into two half-spaces one above the plane and one below the plane

- If the dot product of plane (w) and point (p) is positive then the point p is lying in the same half plane as that of w otherwise it is lying in the other half plane (i.e. opposite direction to w)



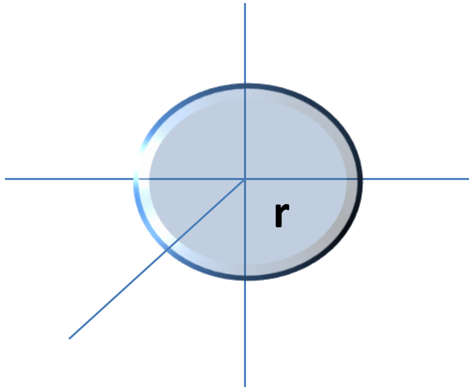
Circle (2-D), Sphere (3-D) & Hypersphere (n-D)

- 2-D (Circle)
- Equation of a circle centred at origin: $x^2 + y^2 = r^2$
- A point p(x_1,x_2) lies
  - Inside the circle if $x_1^2 + x_2^2 < r^2$
  - Outside the circle if $x_1^2 + x_2^2 > r^2$
  - On the circle if $x_1^2 + x_2^2 = r^2$



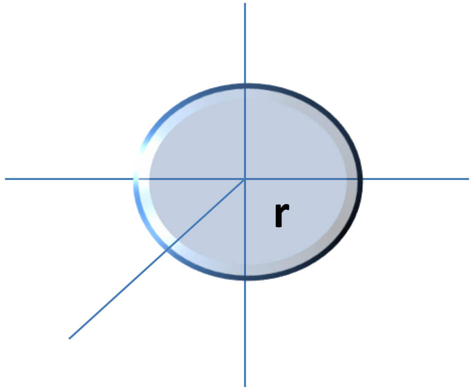
Circle (2-D), Sphere (3-D) & Hypersphere(n-D)

- 3-D (Sphere)
- Equation of a circle centred at origin: $x^2 + y^2 + z^2 = r^2$
- A point p(x$_1$,x$_2$, x$_3$) lies
  - Inside the circle if $x_1^2 + x_2^2 + x_3^2 < r^2$
  - Outside the circle if $x_1^2 + x_2^2 + x_3^2 > r^2$
  - On the circle if $x_1^2 + x_2^2 + x_3^2 = r^2$



Circle (2-D), Sphere (3-D) & Hypersphere(n-D)

- n-D (Hypersphere)
- Equation of a circle centred at origin: $\sum_{i=1}^{n} x_i^2 = r^2$
- A point p(x$_1$,x$_2$, x$_3$ … x$_n$) lies
  - Inside the circle if $\sum_{i=1}^{n} x_i^2 < r^2$
  - Outside the circle if $\sum_{i=1}^{n} x_i^2 > r^2$
  - On the circle if $\sum_{i=1}^{n} x_i^2 = r^2$



Ellipse, Ellipsoid, Hyperellipsoid

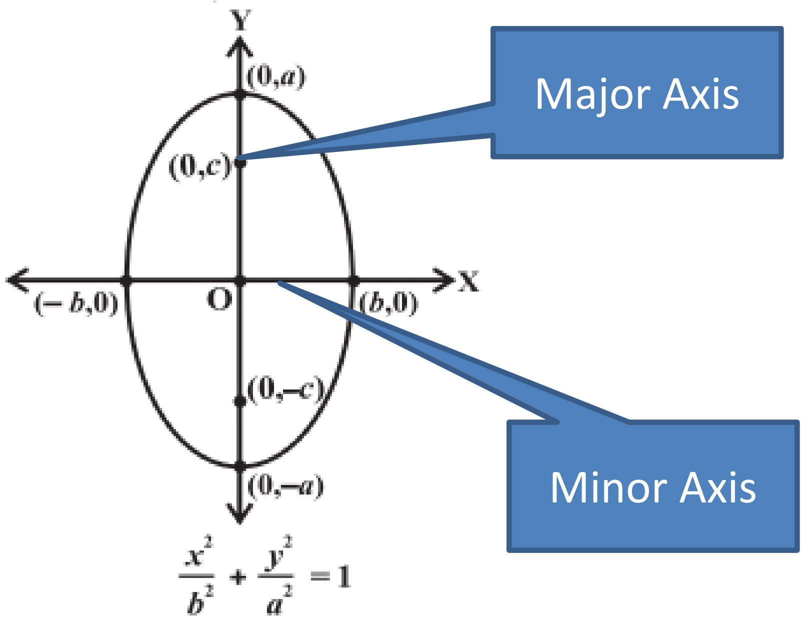
- *Ellipse*

  $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ point p(x,y) lies on the ellipse

  $\frac{x^2}{a^2} + \frac{y^2}{b^2} < 1$ point p(x,y) lies inside ellipse

  $\frac{x^2}{a^2} + \frac{y^2}{b^2} > 1$ point p(x,y) lies outside the ellipse

- *Ellipsoid(3-D)*
  - $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$
- *Hyperellipsoid(n-D)*
  - *Similarly*



$$\frac{x^2}{b^2} + \frac{y^2}{a^2} = 1$$

Major Axis

Minor Axis

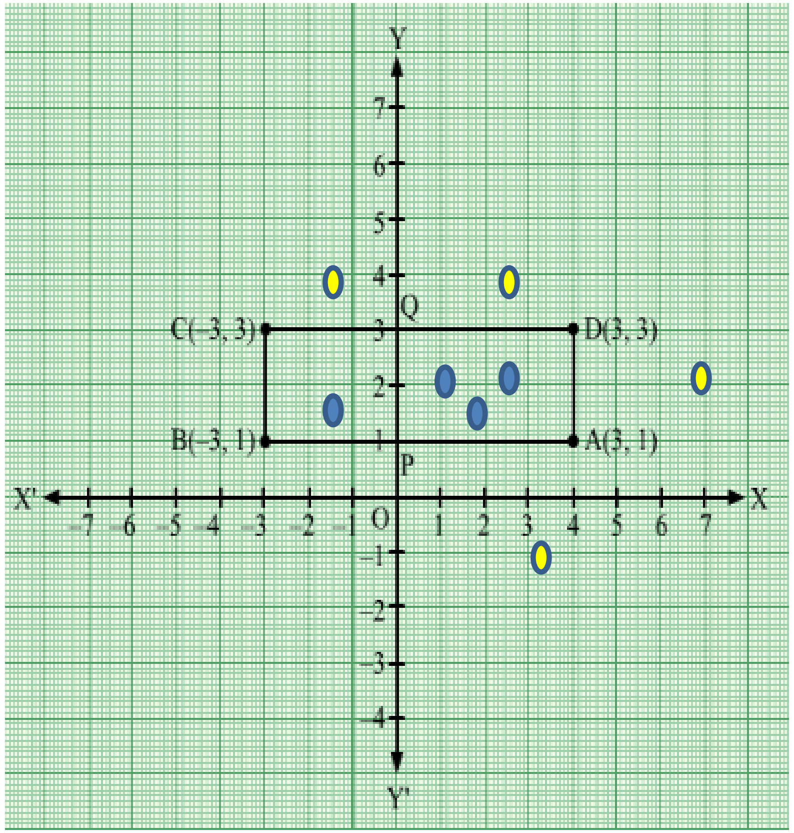
Rectangle, Hyperrectangle, Square, Cube, Hypercube

- If x>=-3 && x<=3

  If y>=1 && y<=3

  then point p(x,y)

  lies within the rectangle



For Your Valuable Time.