



## ***K Nearest Neighbour***

- Rajesh Jakhotia

Earning is in Learning  
- Rajesh Jakhotia

# About K2 Analytics

*At K2 Analytics, we believe that skill development is very important for the growth of an individual, which in turn leads to the growth of Society & Industry and ultimately the Nation as a whole. For this it is important that access to knowledge and skill development trainings should be made available easily and economically to every individual.*

**Our Vision:** *“To be the preferred partner for training and skill development”*

**Our Mission:** *“To provide training and skill development training to individuals, make them skilled & industry ready and create a pool of skilled resources readily available for the industry”*

*We have chosen Business Intelligence and Analytics as our focus area. With this endeavour we make this presentation on “**KNN**” accessible to all those who wish to learn this technique using R / Python. We hope it is of help to you. For any feedback / suggestion feel free to write back to us at [ar.jakhotia@k2analytics.co.in](mailto:ar.jakhotia@k2analytics.co.in)*

*You can also write to us for job opportunities on analytics on our email [ar.jakhotia@k2analytics.co.in](mailto:ar.jakhotia@k2analytics.co.in)*

*Welcome to Logistic Regression using R!!!*



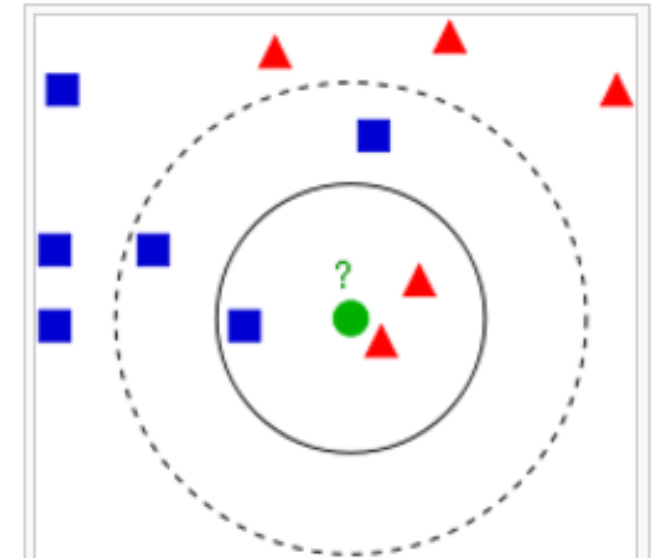
# Agenda


- K Nearest Neighbour
- Advantages and Disadvantages
- KNN Optimization Algorithms

# K Nearest Neighbour

## ***K Nearest Neighbour***

- Non Parametric Technique used for Classification & Regression
- **K-NN Classification** is done based on majority vote of its neighbours
- **K-NN Regression**, the output value is the average of the value of its k nearest neighbours
- K-NN is **Lazy Learning** Technique and simplest of all Machine Learning Techniques
- Neighbours can be given Weights. Common scheme is to give each neighbour a weight of  $1/d$ , where d is the distance to the neighbour



Example of  $k$ -NN classification.   
The test sample (green circle) should be classified either to the first class of blue squares or to the second class of red triangles. If  $k = 3$  (solid line circle) it is assigned to the second class because there are 2 triangles and only 1 square inside the inner circle. If  $k = 5$  (dashed line circle) it is assigned to the first class (3 squares vs. 2 triangles inside the outer circle).

# Distance

- Various distance measures
  - Euclidian Distance
  - Chebyshev Distance
  - Manhattan Distance ...and more

A

Block

Manhattan Distance =  $8 + 4 = 12$

Block

Chebyshev Distance =  $\text{Max}(8, 4) = 8$

Block

Euclidian Distance =  $\text{sqrt}(8^2 + 4^2) = 8.94$

Block

Block

Block

Block

Block

Block

Block

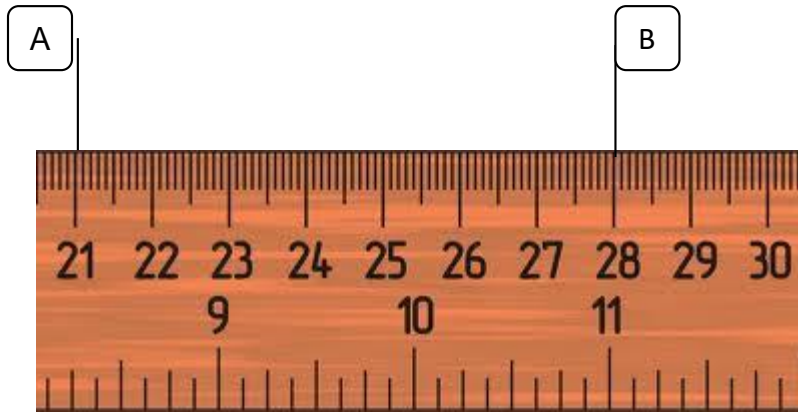
Block

B

Google Search

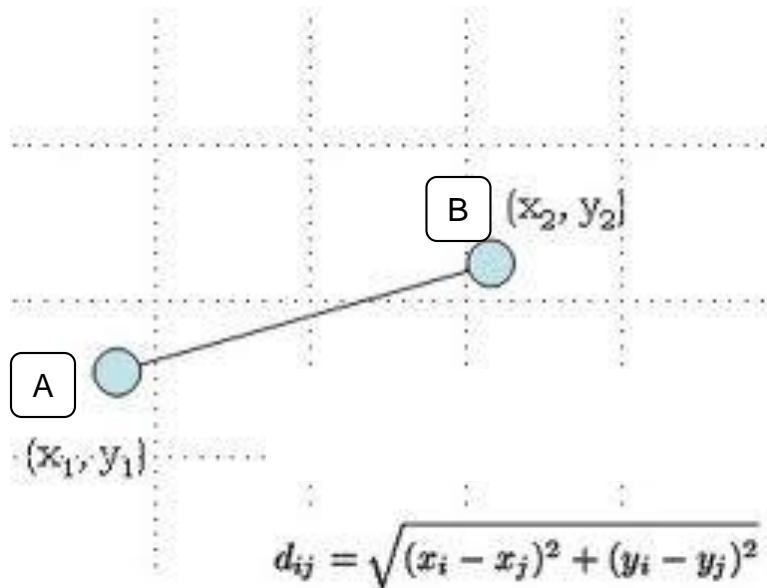
“sklearn Distance Measures”

# Distance Computation



What is the distance between Point A and B?

Ans: 7

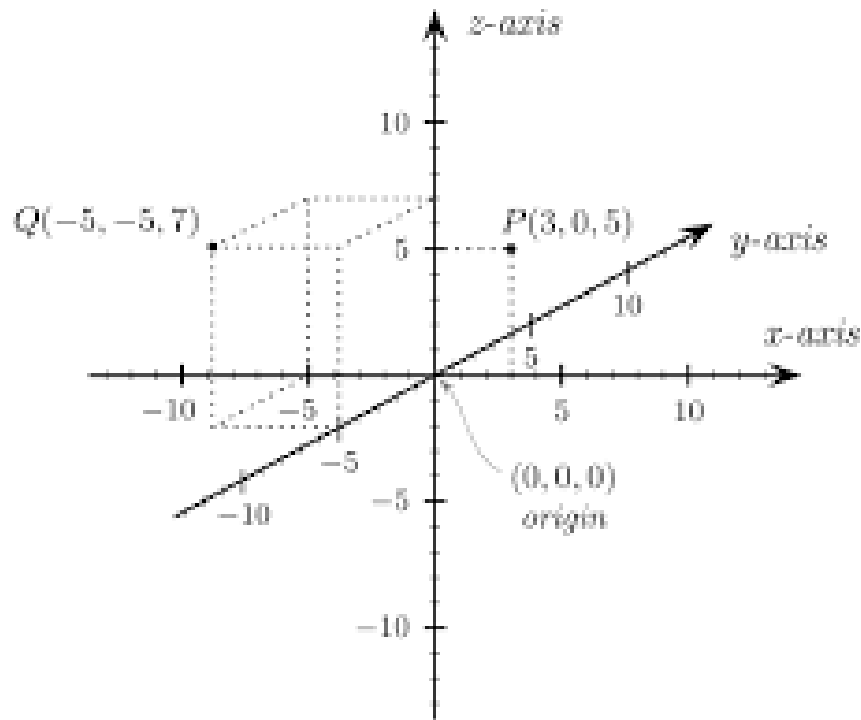


What is the distance between Point A and B?

Ans:  $\sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2]}$

(Remember the Pythagoras Theorem)

# Eucledian Distance



- What is the distance between Point A and B in n-Dimension Space?
- If A  $(x_1, y_1, \dots, z_1)$  and B  $(x_2, y_2, \dots, z_2)$  are cartesian coordinates
- By using **Euclidean Distance** we get Distance AB as
- $D_{AB} = \sqrt{[(x_2 - x_1)^2 + (y_2 - y_1)^2 + \dots + (z_2 - z_1)^2]}$

# Chebyshev Distance

- In mathematics, **Chebyshev distance** is a metric defined on a vector space where the distance between two vectors is the greatest of their differences along any coordinate dimension
- Assume two vectors: A ( $x_1, y_1, \dots, z_1$ ) & B ( $x_2, y_2, \dots, z_2$ )
- Chebyshev Distance
$$= \text{Max} ( |x_2 - x_1| , |y_2 - y_1| , \dots , |z_2 - z_1| )$$
- Application: Survey / Research Data where the responses are Ordinal

Reference Link : [https://en.wikipedia.org/wiki/Chebyshev\\_distance](https://en.wikipedia.org/wiki/Chebyshev_distance)



# Manhattan Distance

- Manhattan Distance also called City Block Distance
- Assume two vectors: A ( $x_1, y_1, \dots, z_1$ ) & B ( $x_2, y_2, \dots, z_2$ )
- Manhattan Distance

$$= |x_2 - x_1| + |y_2 - y_1| + \dots + |z_2 - z_1|$$

A

Block

Manhattan Distance =  $8 + 4 = 12$

Block

Chebyshev Distance =  $\text{Max}(8, 4) = 8$

Block

Euclidean Distance =  $\text{sqrt}(8^2 + 4^2) = 8.94$

Block

Block

Block

Block

Block

Block

Block

Block

B

# Key things to remember

- Feature Selection / Dimension Reduction
- Variable Scaling
- Optimal Value of K to be determined
- In Classification, K should be set as Odd Number to avoid tie
- High Memory & CPU cost at time of classifying

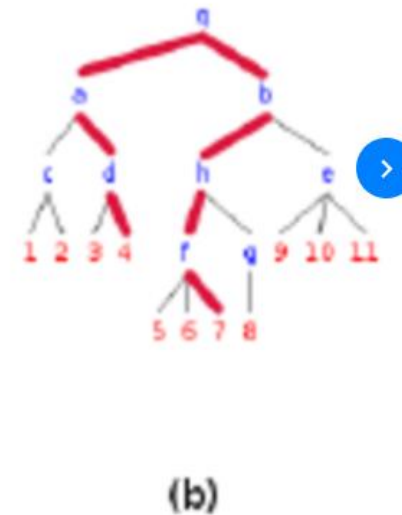
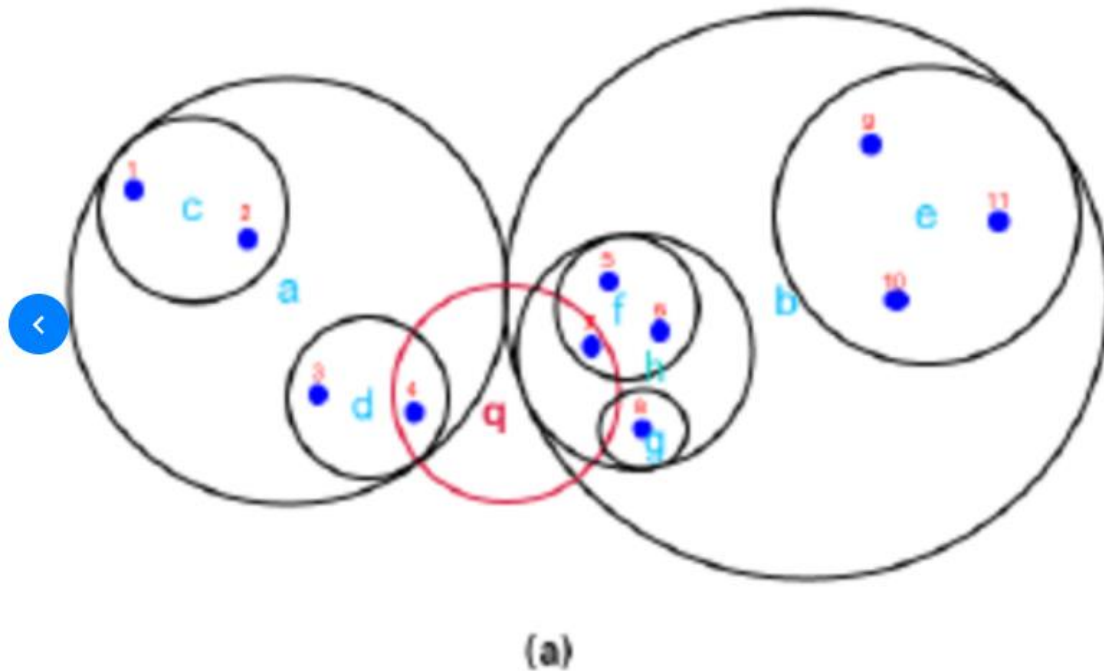
# KNN – Advantages & Disadvantages

- Advantages
  - Very Simple to Understand
  - Makes no assumption about the distribution of data
  - Not impacted by outliers
- Dis-advantages
  - Finding Optimal K is some what a challenge
  - Too much processing time at time of prediction. Performance can be improved by KDTree and BallTree algorithms
  - Ineffective when the class distributions overlap and there is too much of noise in data
  - Does not build any model. It is a Lazy Learner technique

# KNN Optimization Algorithms

**BallTree** – It creates Multi-Dimension Clusters (each resembling balls)

**KDTree** – K Dimensional Tree - Data is separated based on variable having the highest variance



[https://www.researchgate.net/figure/a-Ball-tree-partitions-b-Corresponding-search-tree\\_fig2\\_283471105](https://www.researchgate.net/figure/a-Ball-tree-partitions-b-Corresponding-search-tree_fig2_283471105)

<https://ashokharnal.wordpress.com/tag/ball-tree-explained-in-simple-manner/>



***Thank you***

Contact us:  
[ar.jakhotia@k2analytics.co.in](mailto:ar.jakhotia@k2analytics.co.in)