AI Quest

KNN Algorithm For Pattern Recognition

www.aiquest.org

Application of KNN

- Classification
- Regression
- Clustering
- Anomaly detection
- Recommendation systems

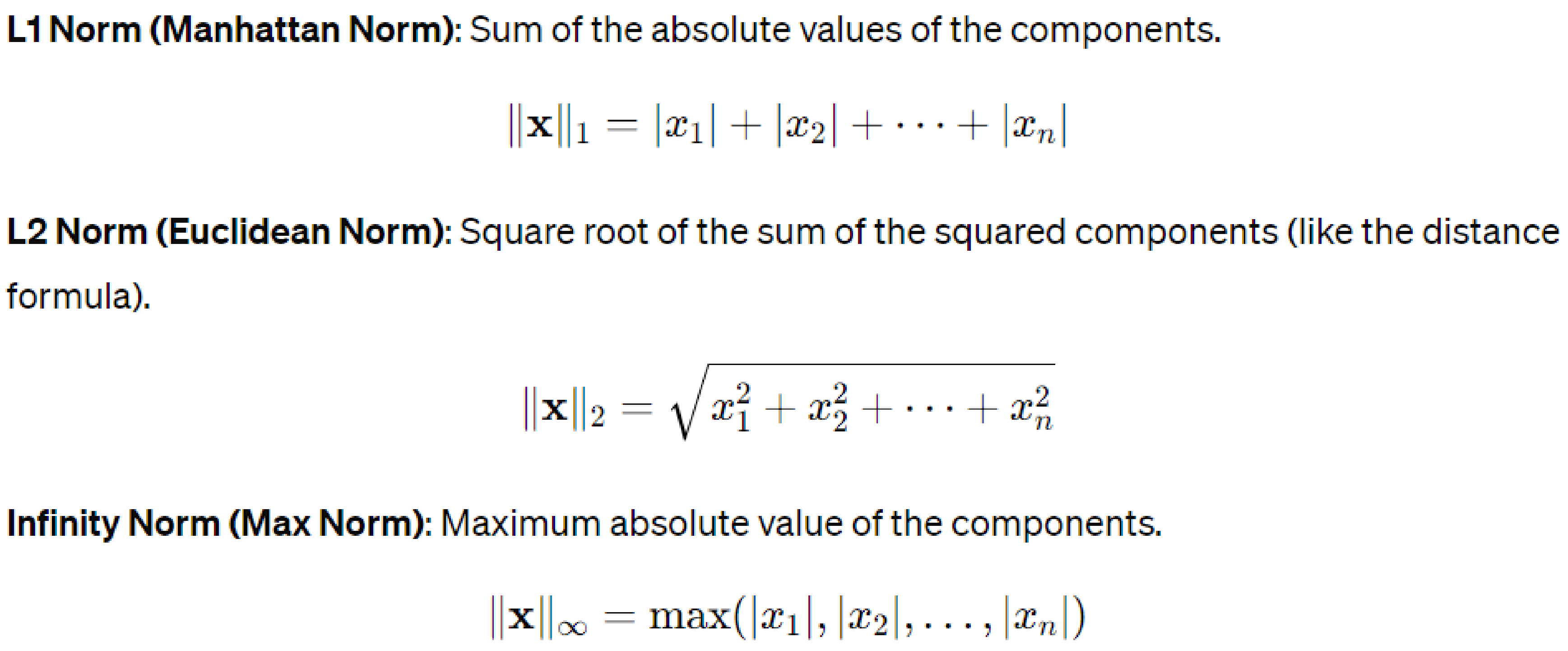
Note: KNN is considered a lazy learning algorithm because it does not perform an explicit training step but rather memorizes the training data and performs computations at the time of prediction. This allows KNN to quickly adapt to new data and handle dynamic environments since it does not require retraining the model when new data becomes available.

Norm Concepts in ML/DL

Norm: In machine learning (especially when we deal with vectors), "norm" refers to a **function** that measures the **length of a vector**, which can be a set of numbers representing features of data.

Norm Concepts in ML/DL

Norm: In machine learning (especially when we deal with vectors), "norm" refers to a **function** that measures the **length of a vector**, which can be a set of numbers representing features of data.



L1 Norm (Manhattan Norm): Sum of the absolute values of the components.

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

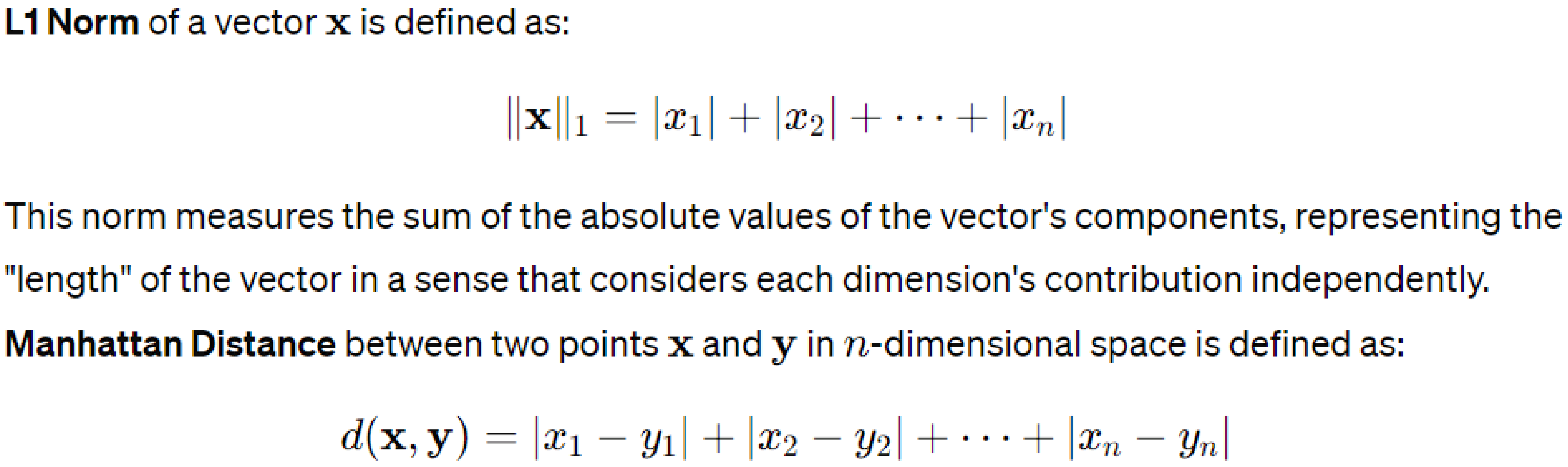
L2 Norm (Euclidean Norm): Square root of the sum of the squared components (like the distance formula).

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

Infinity Norm (Max Norm): Maximum absolute value of the components.

$$\|\mathbf{x}\|_\infty = \max(|x_1|, |x_2|, \ldots, |x_n|)$$

L1 Norm Concepts in ML/DL



L1 Norm of a vector $\mathbf{x}$ is defined as:

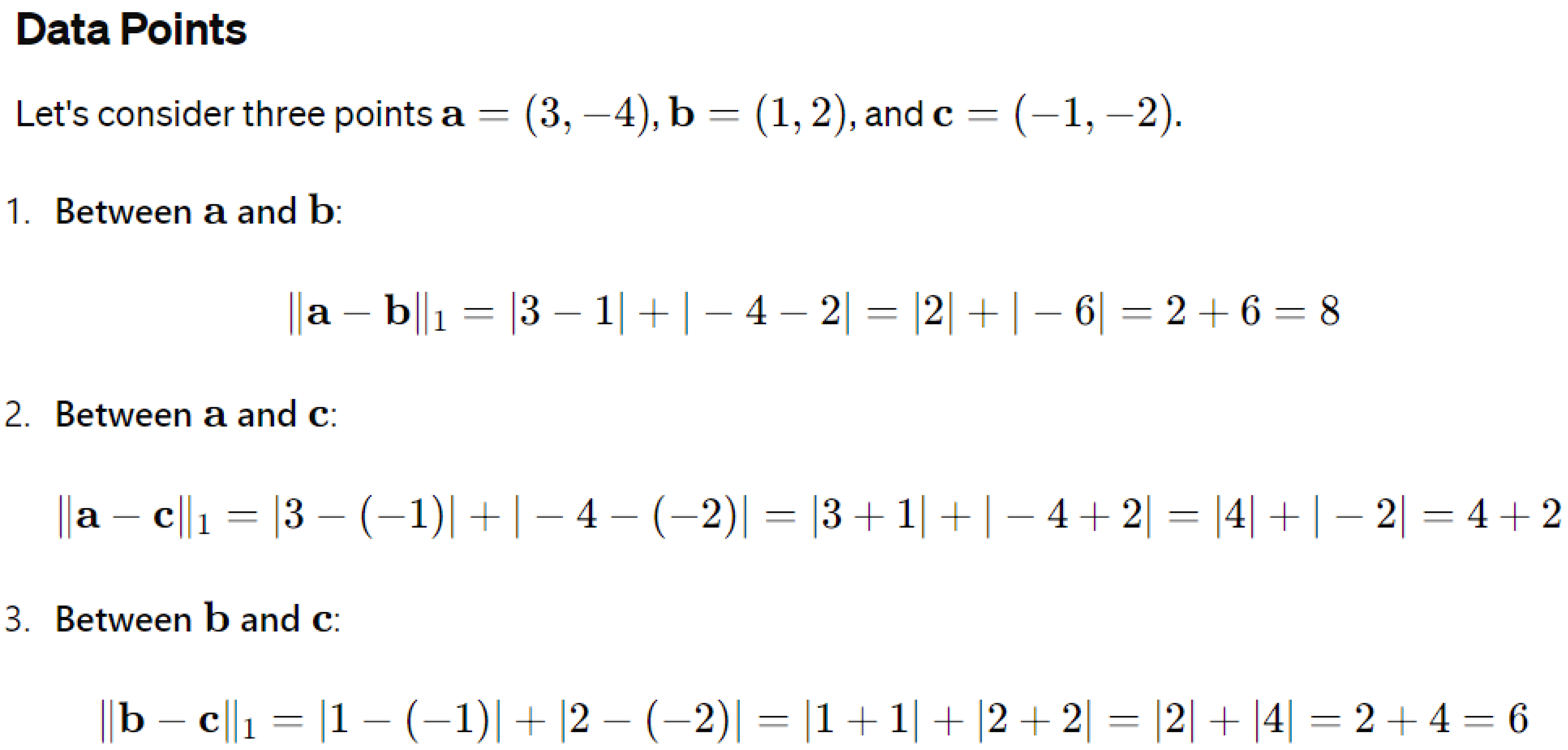
$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \cdots + |x_n|$$

This norm measures the sum of the absolute values of the vector's components, representing the "length" of the vector in a sense that considers each dimension's contribution independently.

Manhattan Distance between two points $\mathbf{x}$ and $\mathbf{y}$ in n-dimensional space is defined as:

$$d(\mathbf{x}, \mathbf{y}) = |x_1 - y_1| + |x_2 - y_2| + \cdots + |x_n - y_n|$$

L1 Norm Concepts in ML/DL



Data Points

Let's consider three points $\mathbf{a} = (3, -4)$, $\mathbf{b} = (1, 2)$, and $\mathbf{c} = (-1, -2)$.

1. Between $\mathbf{a}$ and $\mathbf{b}$:

$$\|\mathbf{a} - \mathbf{b}\|_1 = |3 - 1| + |-4 - 2| = |2| + |-6| = 2 + 6 = 8$$

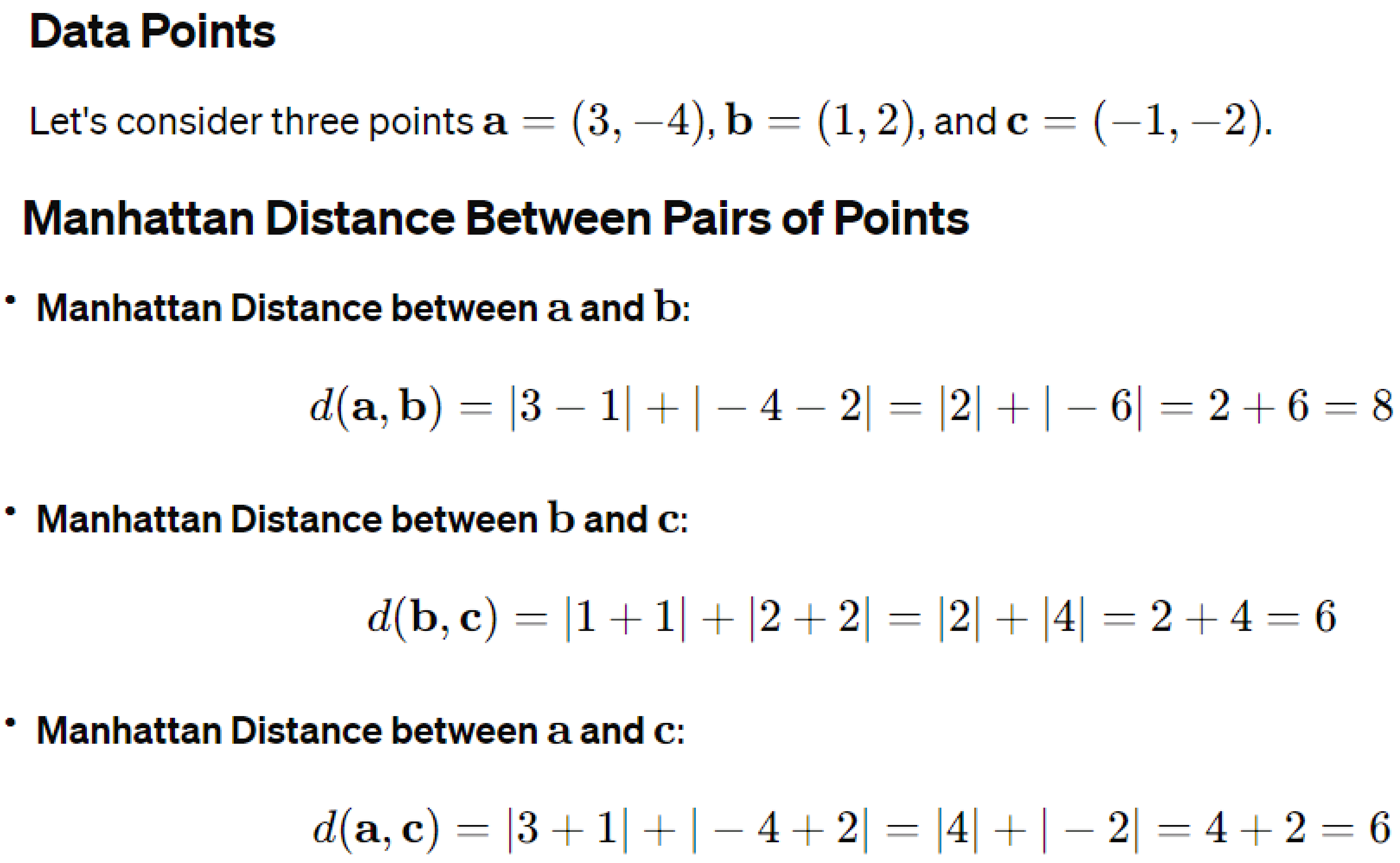
2. Between $\mathbf{a}$ and $\mathbf{c}$:

$$\|\mathbf{a} - \mathbf{c}\|_1 = |3 - (-1)| + |-4 - (-2)| = |3 + 1| + |-4 + 2| = |4| + |-2| = 4 + 2$$

3. Between $\mathbf{b}$ and $\mathbf{c}$:

$$\|\mathbf{b} - \mathbf{c}\|_1 = |1 - (-1)| + |2 - (-2)| = |1 + 1| + |2 + 2| = |2| + |4| = 2 + 4 = 6$$

L1 Norm Concepts in ML/DL



Data Points

Let's consider three points $\mathbf{a} = (3, -4)$, $\mathbf{b} = (1, 2)$, and $\mathbf{c} = (-1, -2)$.

Manhattan Distance Between Pairs of Points

- **Manhattan Distance between $\mathbf{a}$ and $\mathbf{b}$:**

$$d(\mathbf{a}, \mathbf{b}) = |3 - 1| + |-4 - 2| = |2| + |-6| = 2 + 6 = 8$$

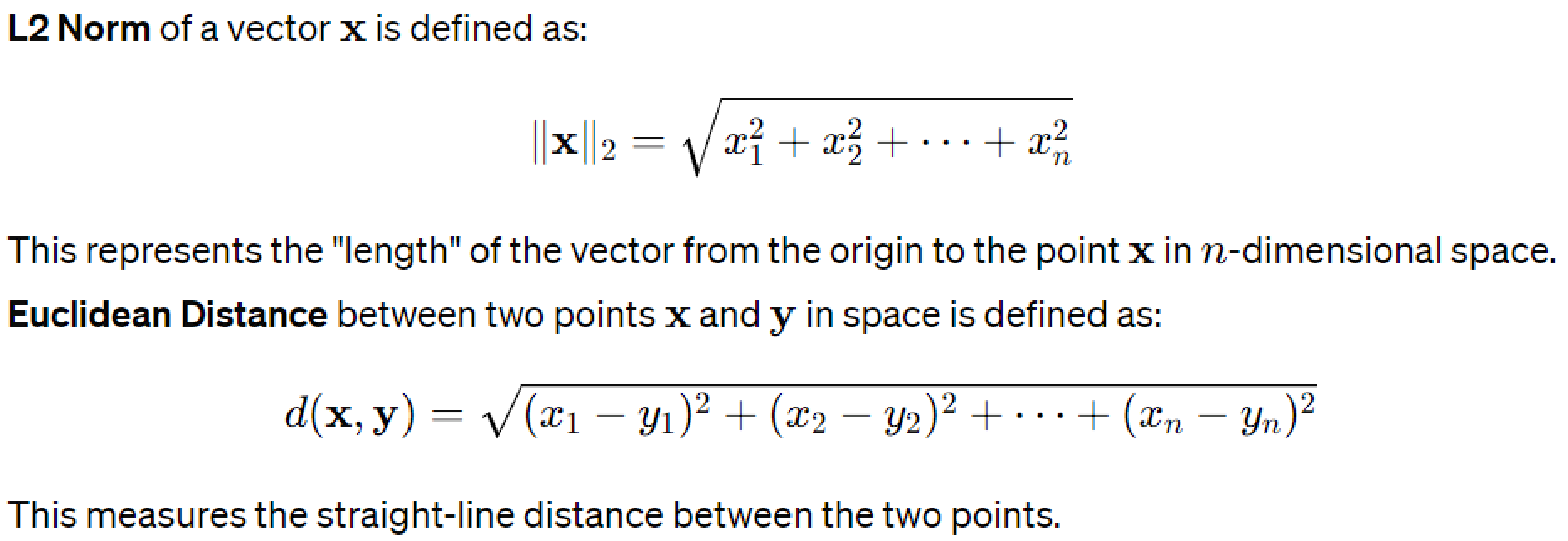
- **Manhattan Distance between $\mathbf{b}$ and $\mathbf{c}$:**

$$d(\mathbf{b}, \mathbf{c}) = |1 + 1| + |2 + 2| = |2| + |4| = 2 + 4 = 6$$

- **Manhattan Distance between $\mathbf{a}$ and $\mathbf{c}$:**

$$d(\mathbf{a}, \mathbf{c}) = |3 + 1| + |-4 + 2| = |4| + |-2| = 4 + 2 = 6$$

L2 Norm Concepts in ML/DL



L2 Norm of a vector $\mathbf{x}$ is defined as:

$$\|\mathbf{x}\|_2 = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}$$

This represents the "length" of the vector from the origin to the point $\mathbf{x}$ in n-dimensional space.

Euclidean Distance between two points $\mathbf{x}$ and $\mathbf{y}$ in space is defined as:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

This measures the straight-line distance between the two points.

L2 Norm Concepts in ML/DL

Data Points

Let's take three points $\mathbf{a} = (1, 2)$, $\mathbf{b} = (3, 4)$, and $\mathbf{c} = (-1, -1)$.

Euclidean Distance Between Pairs of Points

- **Euclidean Distance between $\mathbf{a}$ and $\mathbf{b}$:**

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{(1-3)^2 + (2-4)^2} = \sqrt{(-2)^2 + (-2)^2} = \sqrt{4+4} = \sqrt{8} = 2\sqrt{2}$$

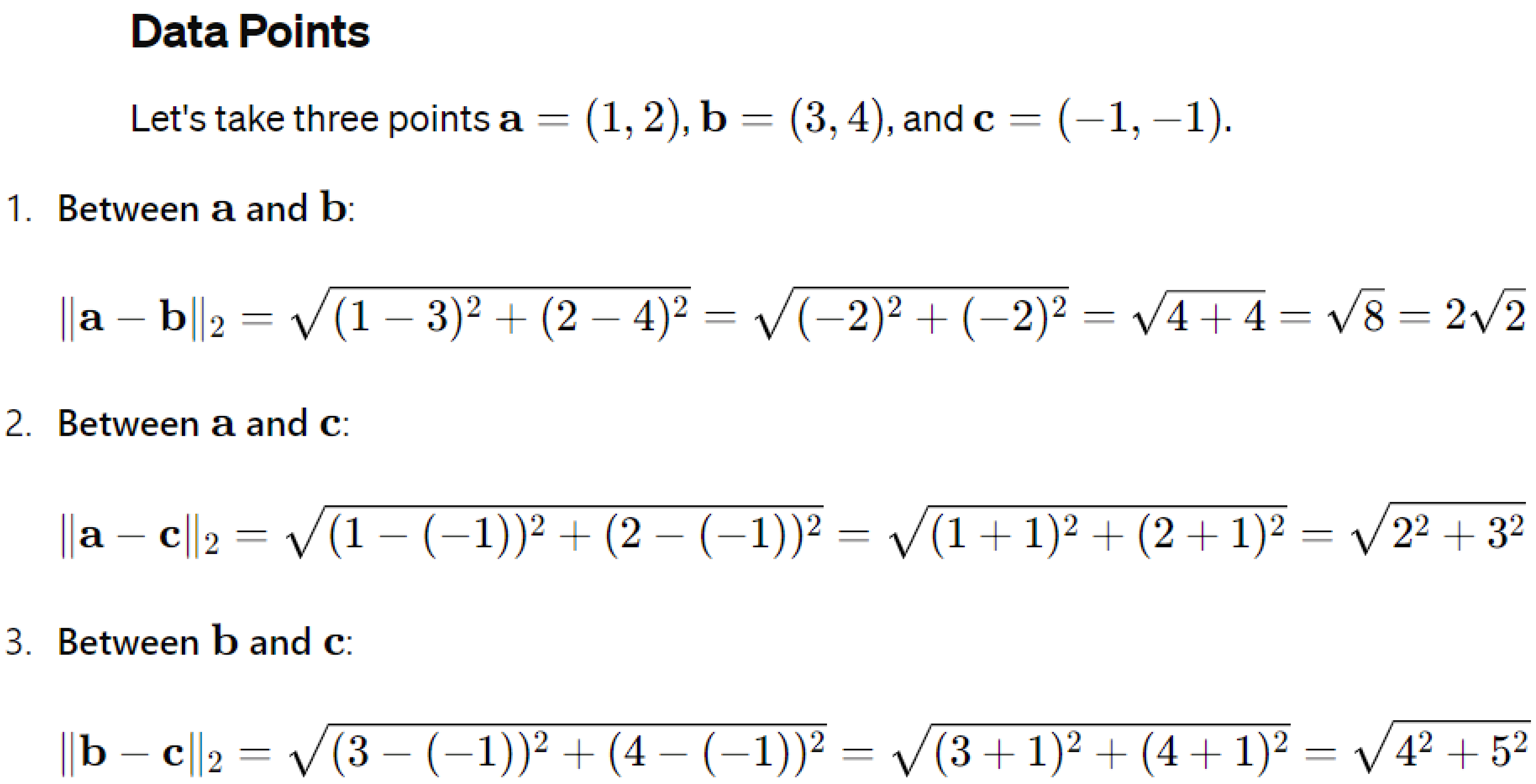
- **Euclidean Distance between $\mathbf{b}$ and $\mathbf{c}$:**

$$d(\mathbf{b}, \mathbf{c}) = \sqrt{(3+1)^2 + (4+1)^2} = \sqrt{4^2 + 5^2} = \sqrt{16+25} = \sqrt{41}$$

- **Euclidean Distance between $\mathbf{a}$ and $\mathbf{c}$:**

$$d(\mathbf{a}, \mathbf{c}) = \sqrt{(1+1)^2 + (2+1)^2} = \sqrt{2^2 + 3^2} = \sqrt{4+9} = \sqrt{13}$$

L2 Norm Concepts in ML/DL



Data Points

Let's take three points $\mathbf{a} = (1, 2)$, $\mathbf{b} = (3, 4)$, and $\mathbf{c} = (-1, -1)$.

1. Between $\mathbf{a}$ and $\mathbf{b}$:

$$\|\mathbf{a} - \mathbf{b}\|_2 = \sqrt{(1-3)^2 + (2-4)^2} = \sqrt{(-2)^2 + (-2)^2} = \sqrt{4+4} = \sqrt{8} = 2\sqrt{2}$$

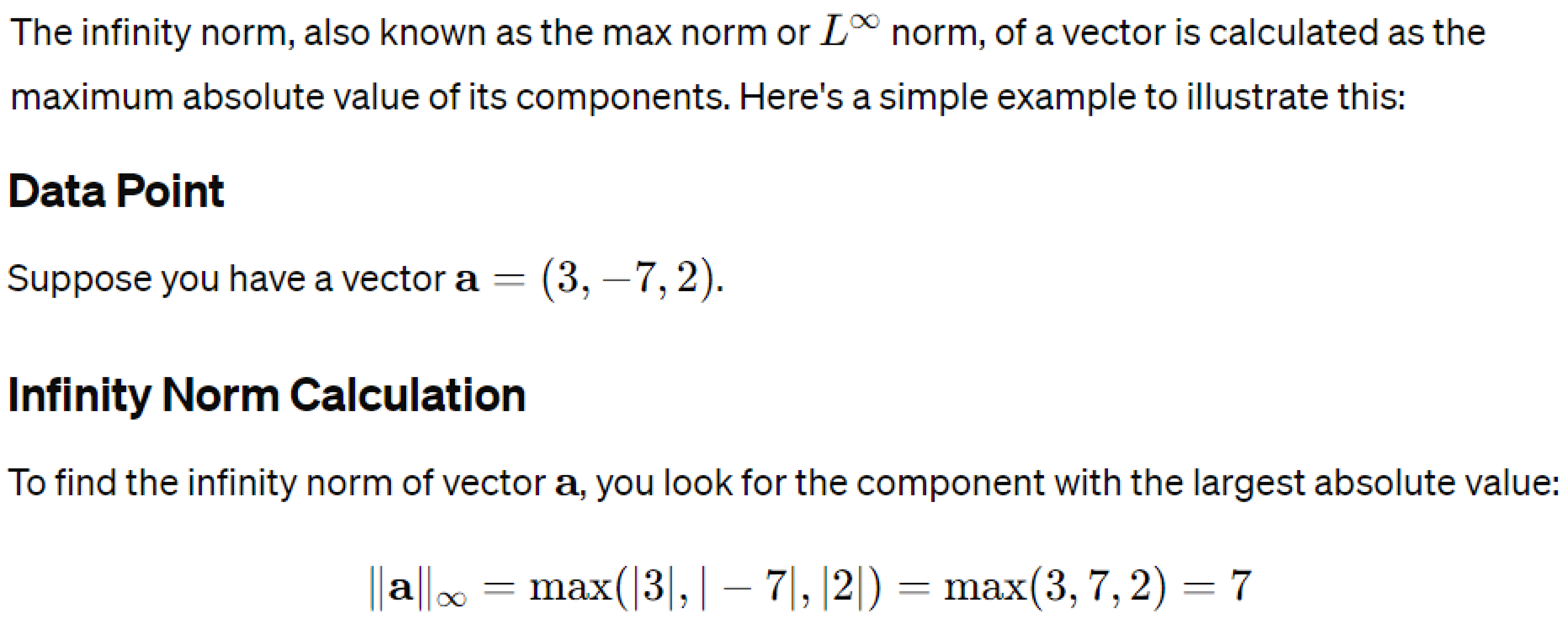
2. Between $\mathbf{a}$ and $\mathbf{c}$:

$$\|\mathbf{a} - \mathbf{c}\|_2 = \sqrt{(1-(-1))^2 + (2-(-1))^2} = \sqrt{(1+1)^2 + (2+1)^2} = \sqrt{2^2 + 3^2}$$

3. Between $\mathbf{b}$ and $\mathbf{c}$:

$$\|\mathbf{b} - \mathbf{c}\|_2 = \sqrt{(3-(-1))^2 + (4-(-1))^2} = \sqrt{(3+1)^2 + (4+1)^2} = \sqrt{4^2 + 5^2}$$

∞ Norm Concepts in ML/DL



The infinity norm, also known as the max norm or L^∞ norm, of a vector is calculated as the maximum absolute value of its components. Here's a simple example to illustrate this:

Data Point

Suppose you have a vector $\mathbf{a} = (3, -7, 2)$.

Infinity Norm Calculation

To find the infinity norm of vector $\mathbf{a}$, you look for the component with the largest absolute value:

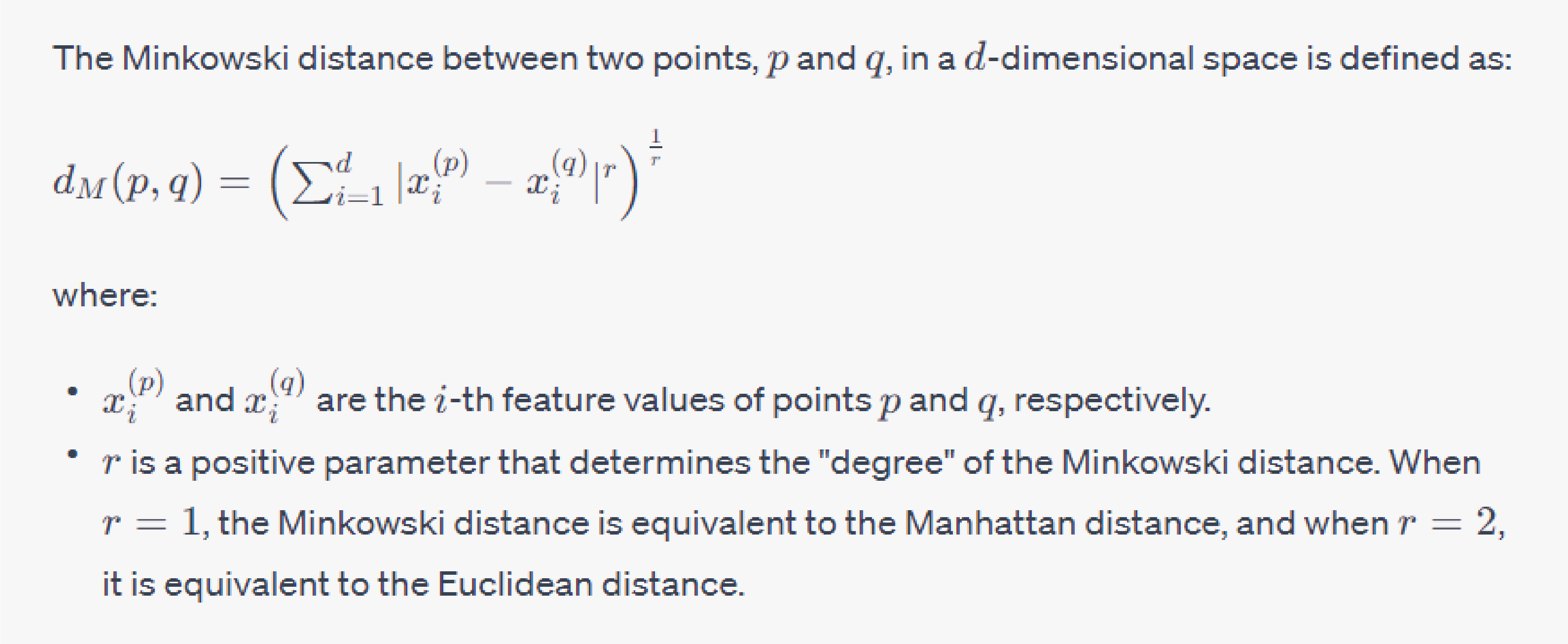
$$||\mathbf{a}||_\infty = \max(|3|, |-7|, |2|) = \max(3, 7, 2) = 7$$

Norm Concepts in ML/DL

Norms are used in several key contexts in machine learning when dealing with vectors:

- **Magnitude Measurement:** Norms measure the size or length of vectors.
- **Optimization:** They help minimize errors in models, such as in least squares optimization.
- **Regularization:** Norms control model complexity in regularization techniques like L1 and L2, helping prevent overfitting.
- **Distance Calculation:** In algorithms like k-NN and k-means clustering, norms determine the distances between data points.
- **Error Evaluation:** Norms quantify the error in predictions, like calculating Mean Squared Error (MSE).
- **Data Normalization:** They are used to scale data vectors uniformly in preprocessing.

The Minkowski Distance



The Minkowski distance between two points, p and q, in a d-dimensional space is defined as:

$$d_M(p, q) = \left(\sum_{i=1}^{d} |x_i^{(p)} - x_i^{(q)}|^r\right)^{\frac{1}{r}}$$

where:

- $x_i^{(p)}$ and $x_i^{(q)}$ are the i-th feature values of points p and q, respectively.
- r is a positive parameter that determines the "degree" of the Minkowski distance. When $r = 1$, the Minkowski distance is equivalent to the Manhattan distance, and when $r = 2$, it is equivalent to the Euclidean distance.

Perform K-NN Algorithm (Visually)

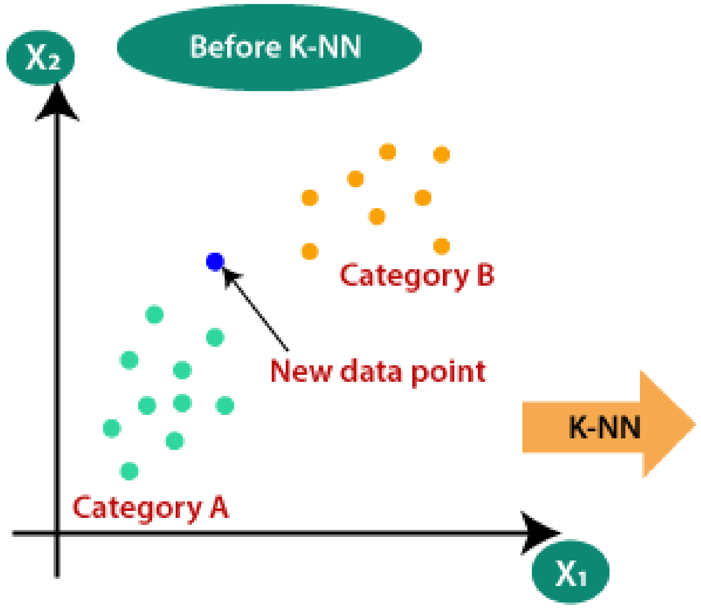


Image: javatpoint.com

Perform K-NN Algorithm (Visually)

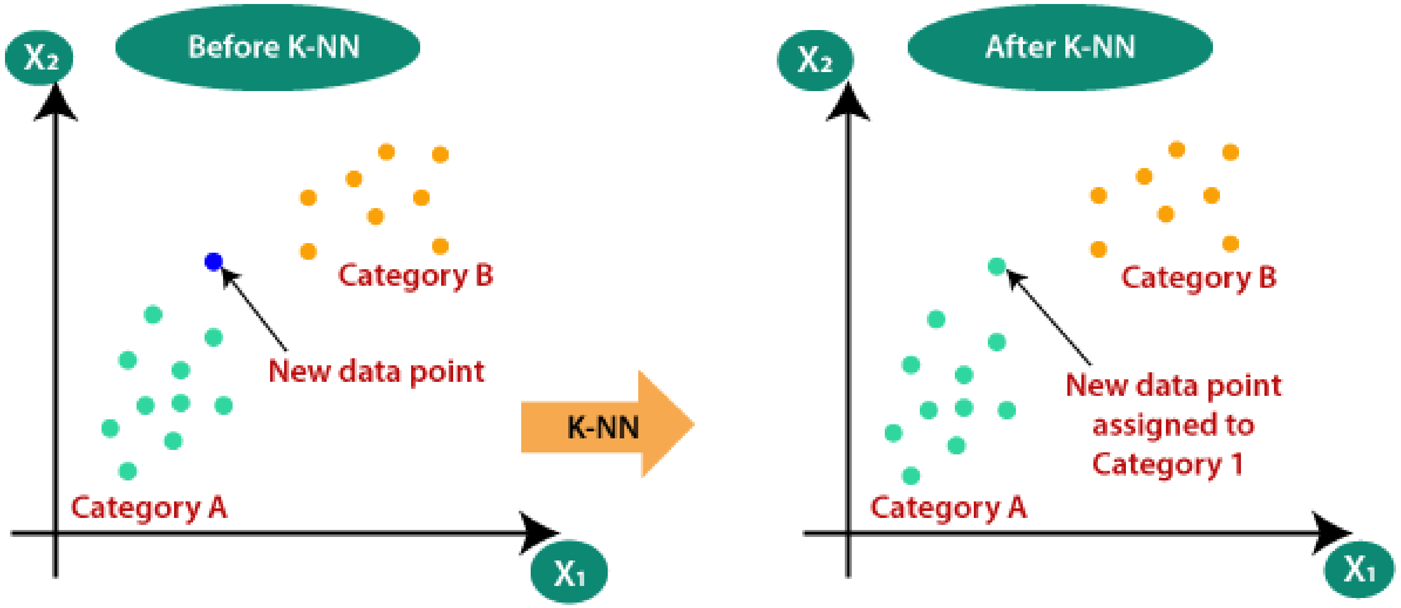


Image: javatpoint.com

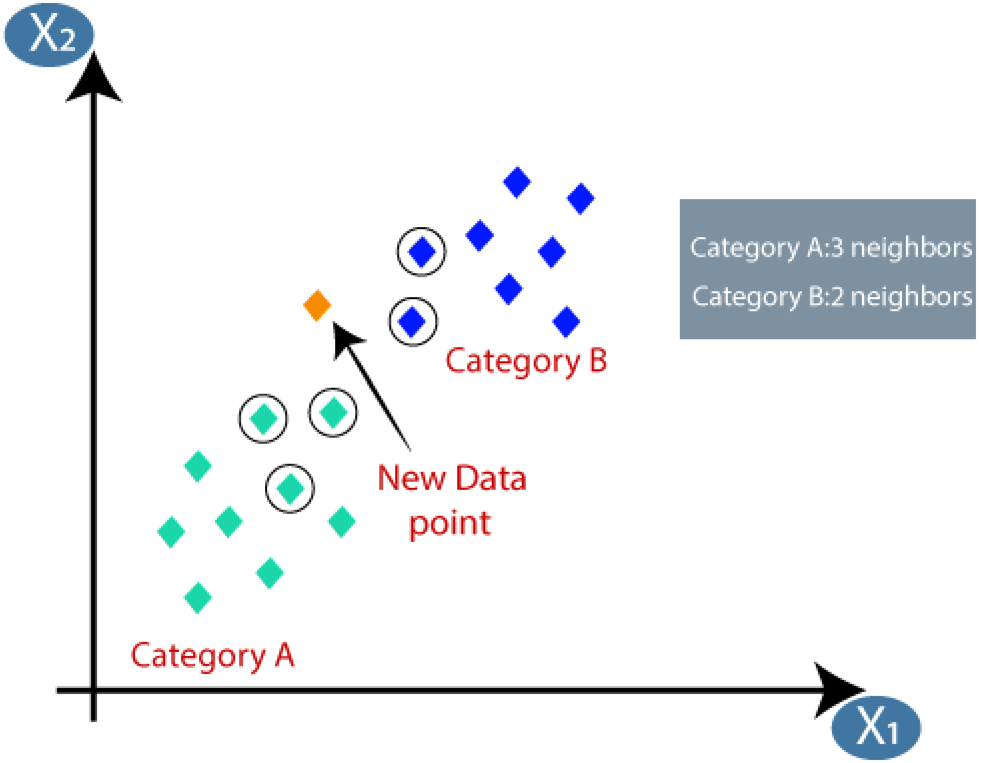
Steps: K-NN Algorithm

Explanation of the K-Nearest Neighbors (KNN) algorithm:

1. **Collect the data:** Obtain a labeled dataset with instances and their corresponding class labels.
2. **Choose the value of K:** Determine the number of nearest neighbors to consider.
3. **Preprocess the data if needed:** Perform any necessary data preprocessing steps, such as normalization or handling missing values.
4. **Calculate distances:** Measure the distance between the new instance and all instances in the training dataset using a distance metric, commonly the Euclidean distance.
5. **Find K nearest neighbors:** Select the K instances with the shortest distances to the new instance.
6. **Determine the class:** For classification tasks, assign the class label that appears most frequently among the K nearest neighbors as the predicted class for the new instance.
7. Make predictions: Repeat steps 4 to 6 for all new instances in the test dataset to generate predictions.

That summarizes the core steps of the KNN algorithm. Please note that this simplified explanation omits the steps related to tuning the algorithm's parameters and evaluating its performance.

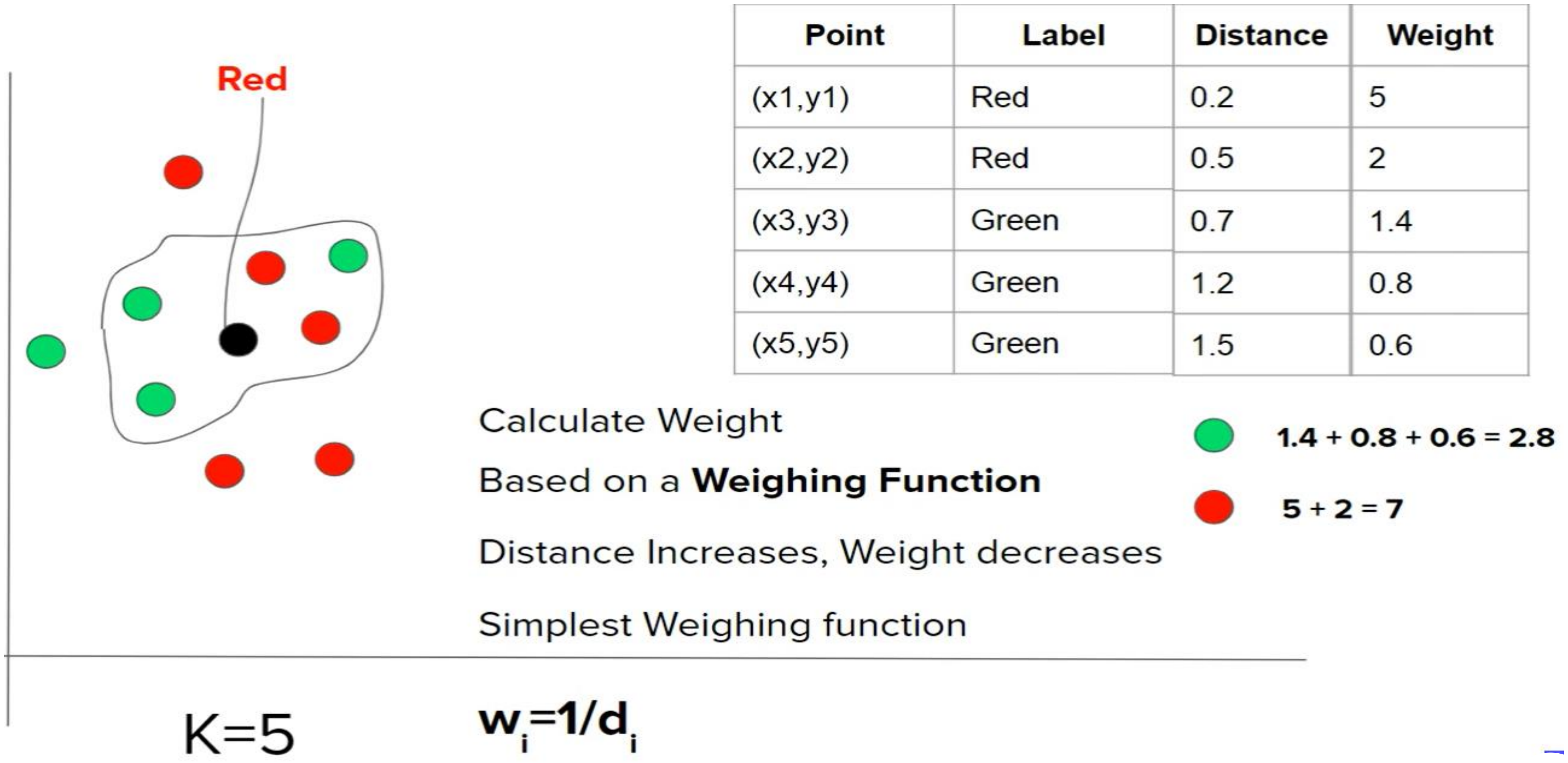
Perform K-NN for Classification (Visually)



www.aiquest.org

Image: javatpoint.com

K-NN Parameters: weights{‘uniform’, ‘distance’}



Point	Label	Distance	Weight
(x1,y1)	Red	0.2	5
(x2,y2)	Red	0.5	2
(x3,y3)	Green	0.7	1.4
(x4,y4)	Green	1.2	0.8
(x5,y5)	Green	1.5	0.6

Calculate Weight

Based on a **Weighing Function**

Distance Increases, Weight decreases

Simplest Weighing function

$$\mathbf{w_i=1/d_i}$$

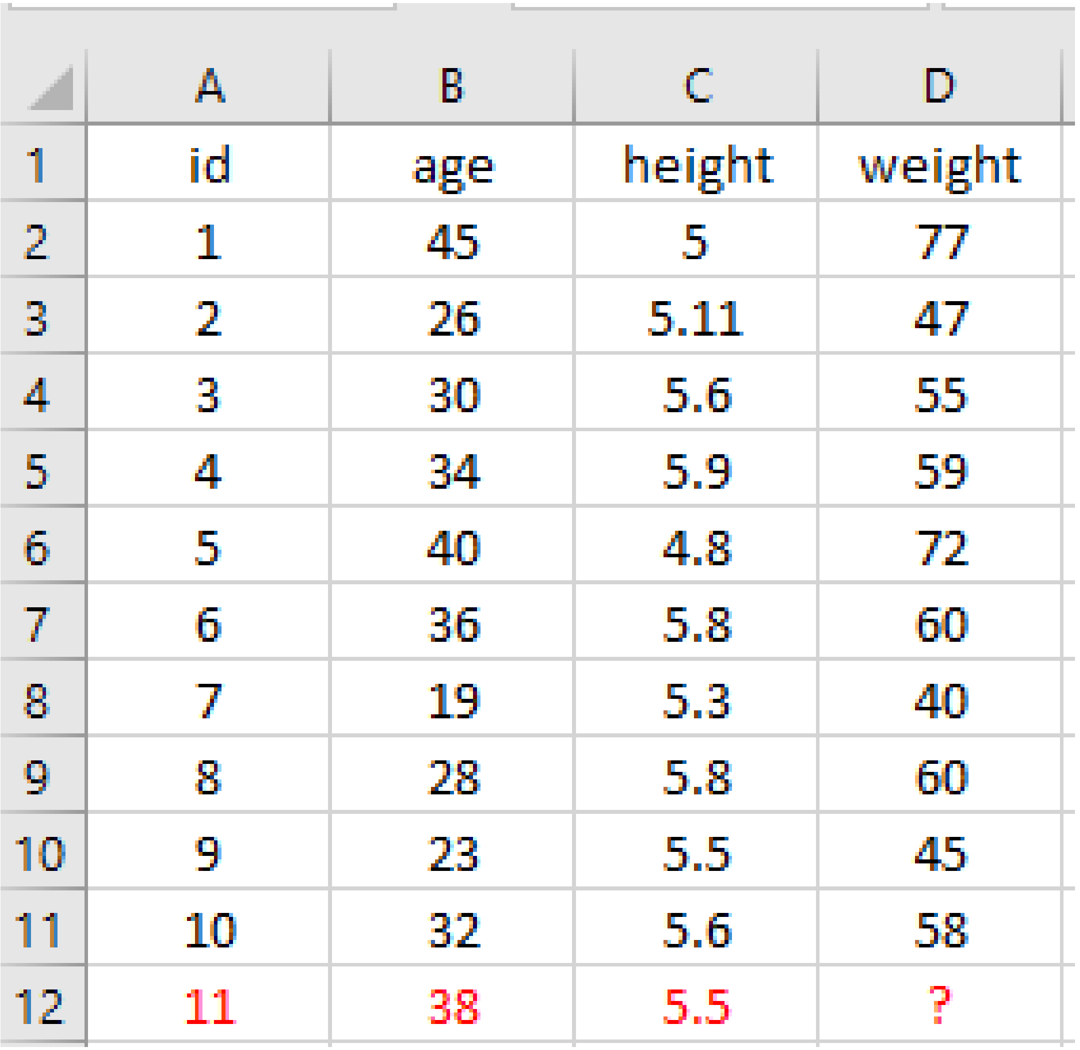
🟩 **1.4 + 0.8 + 0.6 = 2.8**

🟥 **5 + 2 = 7**

Image: Google.com

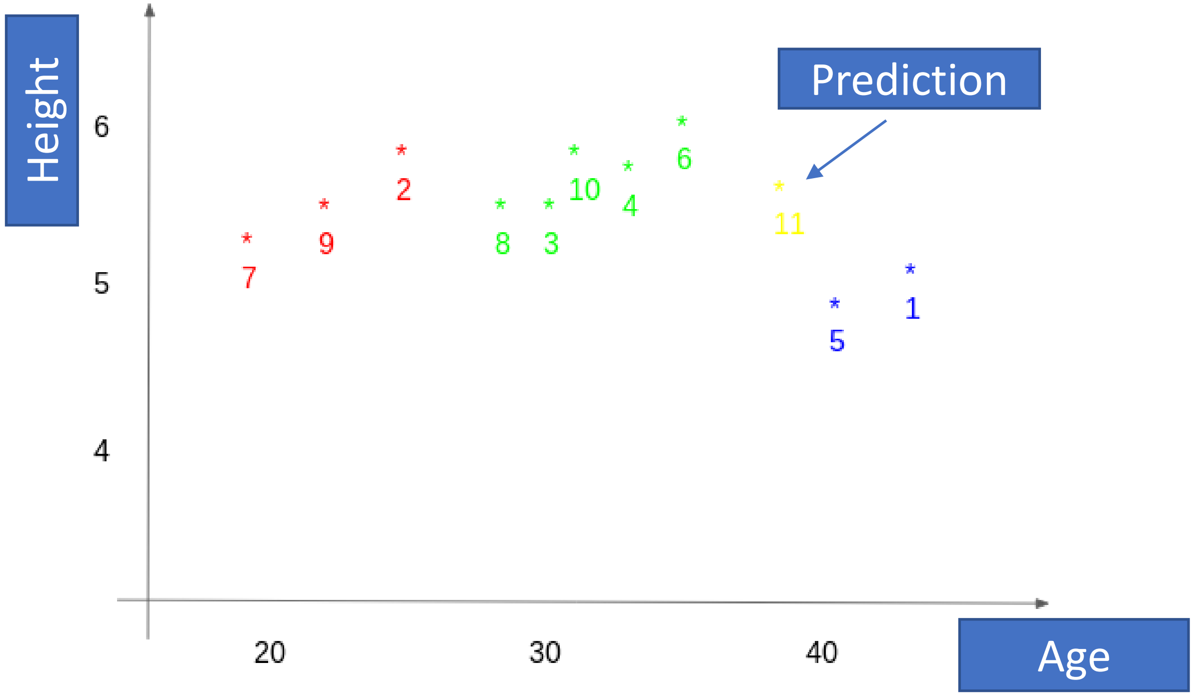
Let's Calculate for Regression!

K-NN Algorithm: Regression



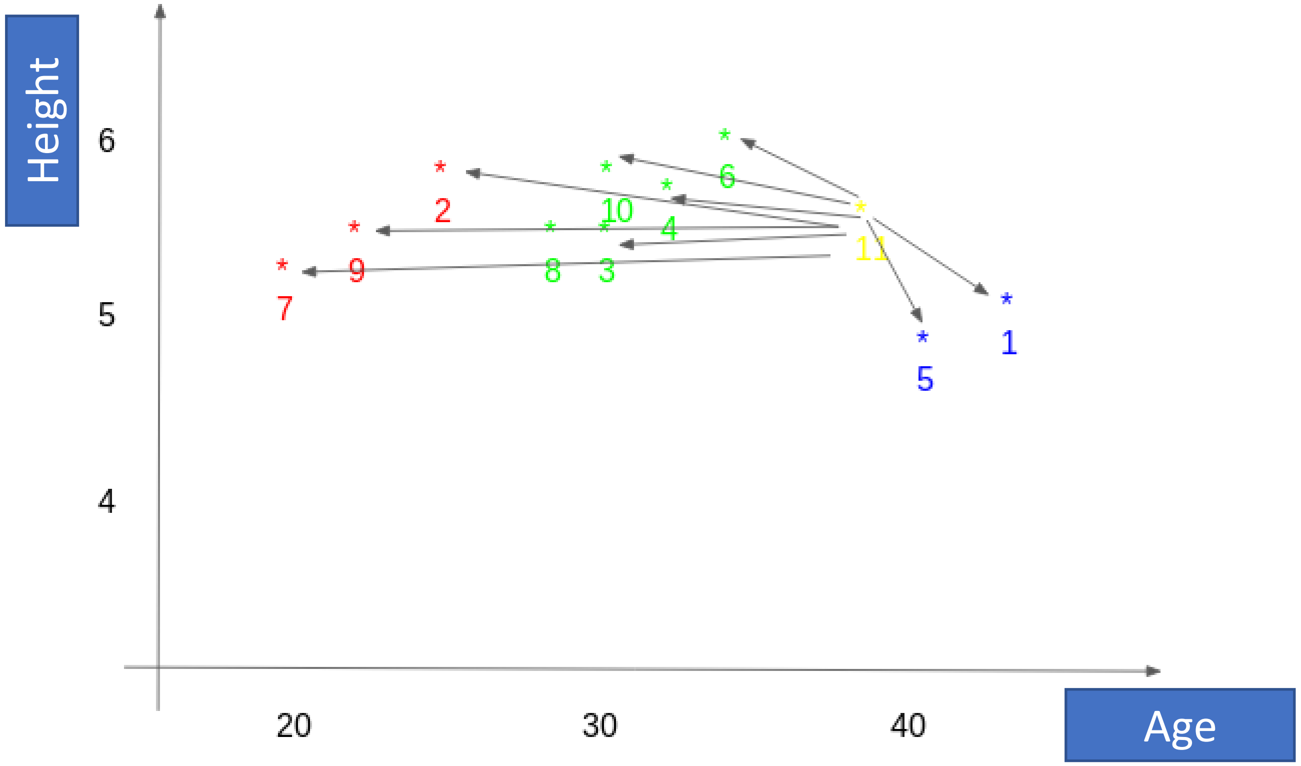
	A	B	C	D
1	id	age	height	weight
2	1	45	5	77
3	2	26	5.11	47
4	3	30	5.6	55
5	4	34	5.9	59
6	5	40	4.8	72
7	6	36	5.8	60
8	7	19	5.3	40
9	8	28	5.8	60
10	9	23	5.5	45
11	10	32	5.6	58
12	11	38	5.5	?

K-NN Algorithm: Regression



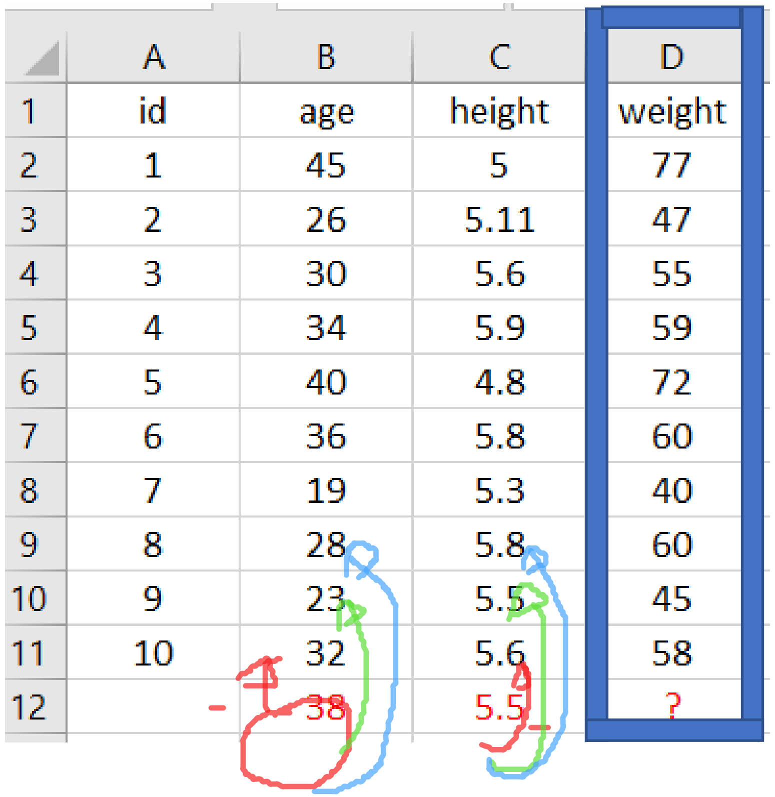
Analyticsvidhya.com

K-NN Algorithm: Regression



Analyticsvidhya.com

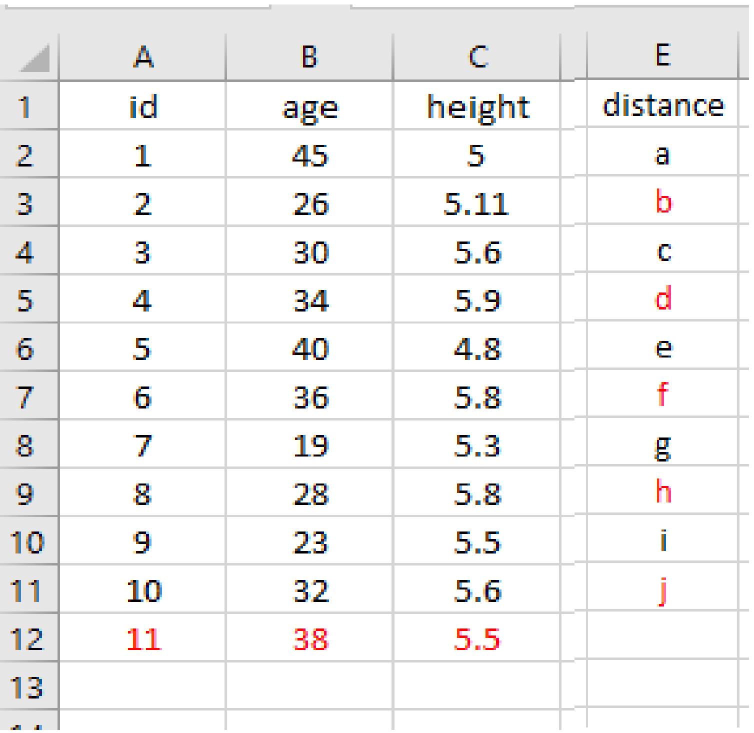
K-NN Algorithm: Regression



	A	B	C	D
1	id	age	height	weight
2	1	45	5	77
3	2	26	5.11	47
4	3	30	5.6	55
5	4	34	5.9	59
6	5	40	4.8	72
7	6	36	5.8	60
8	7	19	5.3	40
9	8	28	5.8	60
10	9	23	5.5	45
11	10	32	5.6	58
12		38	5.5	?

Distance,

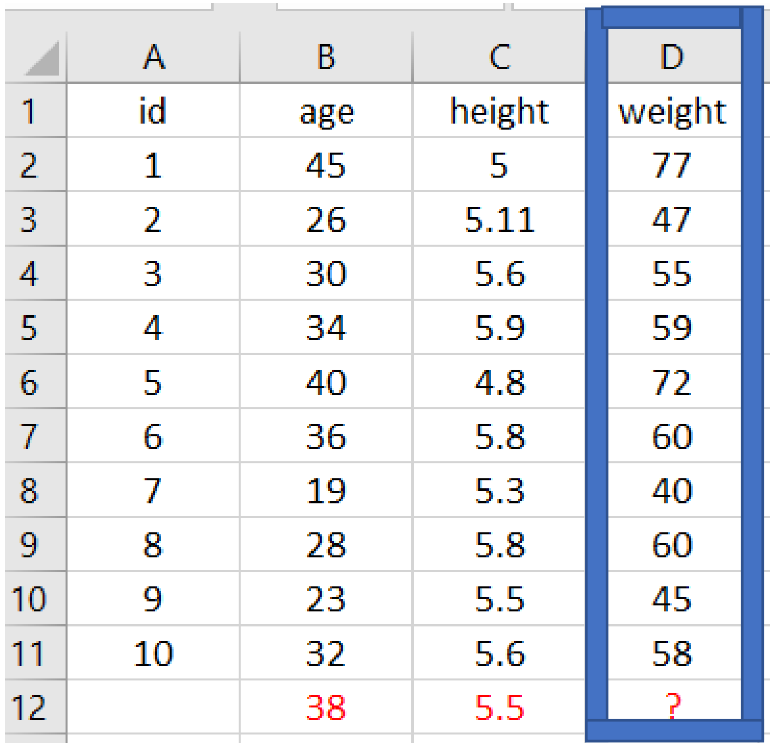
$$d(p,q)^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2$$



	A	B	C	E
1	id	age	height	distance
2	1	45	5	a
3	2	26	5.11	b
4	3	30	5.6	c
5	4	34	5.9	d
6	5	40	4.8	e
7	6	36	5.8	f
8	7	19	5.3	g
9	8	28	5.8	h
10	9	23	5.5	i
11	10	32	5.6	j
12	11	38	5.5	
13				

এবার আর ধাপে ১২ এর রেকর্ড তার থেকে প্রতিটা D এর distance বের করব।

K-NN Algorithm: Regression



	A	B	C	D
1	id	age	height	weight
2	1	45	5	77
3	2	26	5.11	47
4	3	30	5.6	55
5	4	34	5.9	59
6	5	40	4.8	72
7	6	36	5.8	60
8	7	19	5.3	40
9	8	28	5.8	60
10	9	23	5.5	45
11	10	32	5.6	58
12		38	5.5	?

Distance,

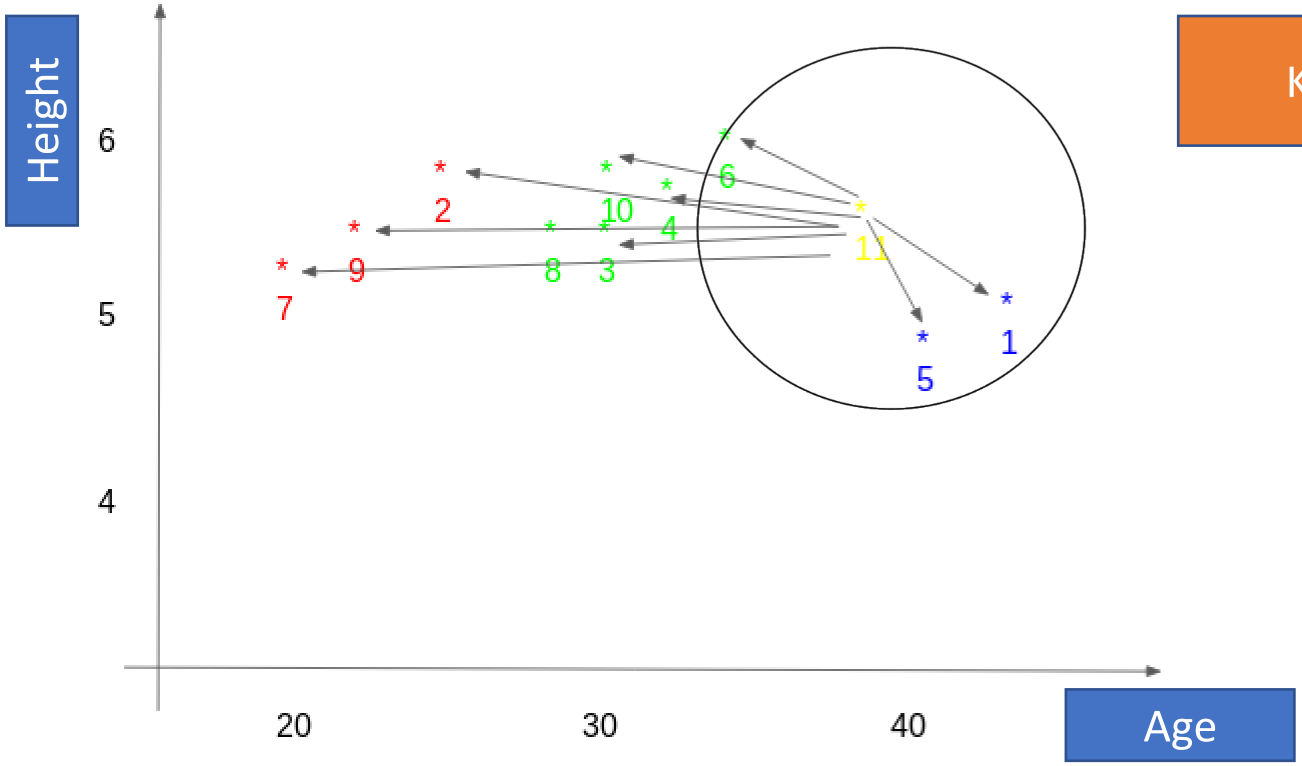
$$d(p,q)^2 = (q_1 - p_1)^2 + (q_2 - p_2)^2$$

Sequence:

a > e > f > d > j > c > h > b > i > g

k=3 k=5

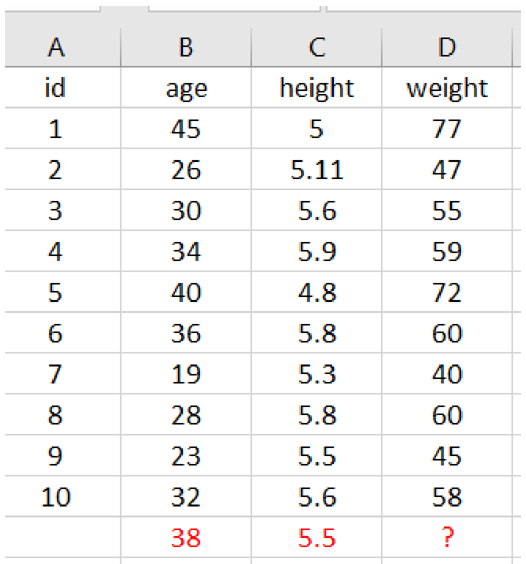
K-NN Algorithm: Regression



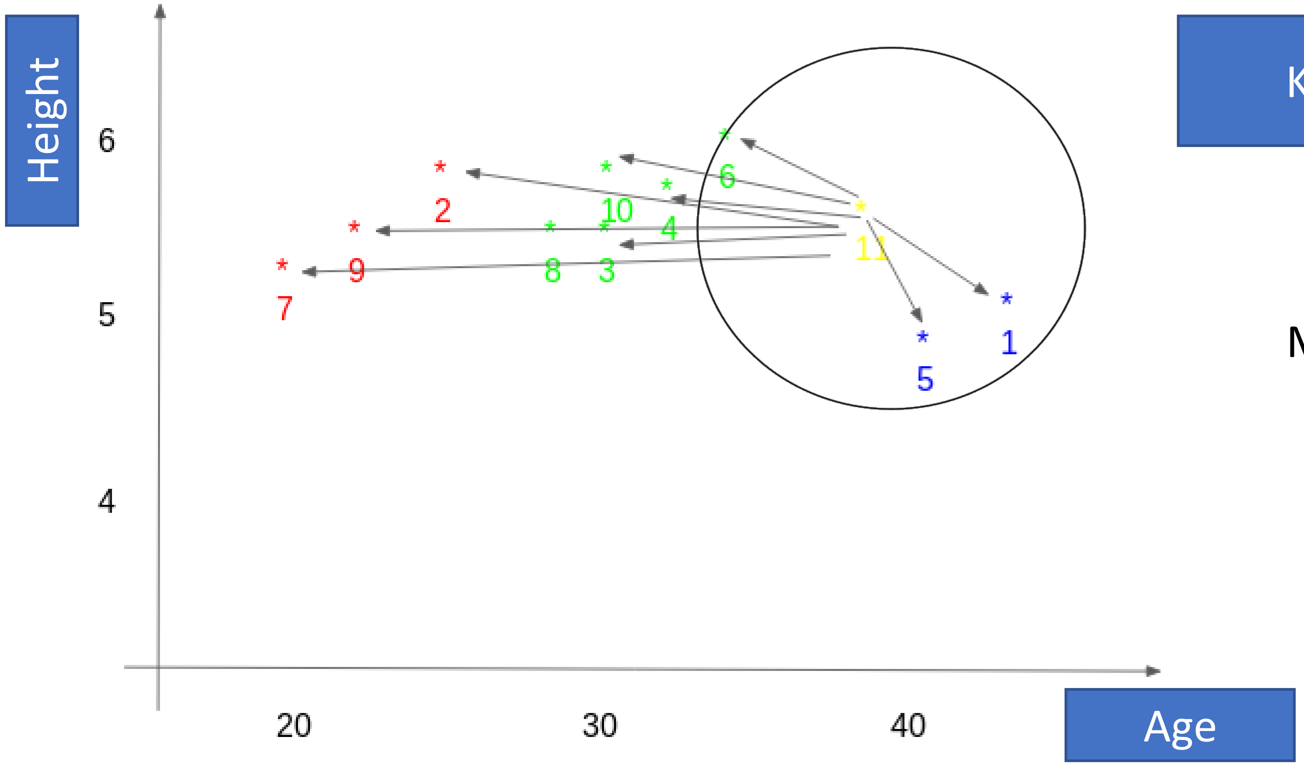
K = 3

Analyticsvidhya.com

K-NN Algorithm: Regression



A	B	C	D
id	age	height	weight
1	45	5	77
2	26	5.11	47
3	30	5.6	55
4	34	5.9	59
5	40	4.8	72
6	36	5.8	60
7	19	5.3	40
8	28	5.8	60
9	23	5.5	45
10	32	5.6	58
	38	5.5	?

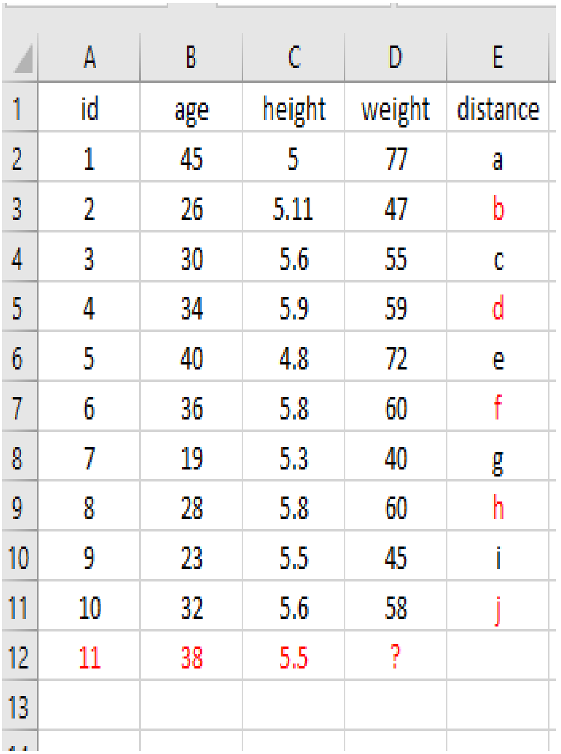


K = 3

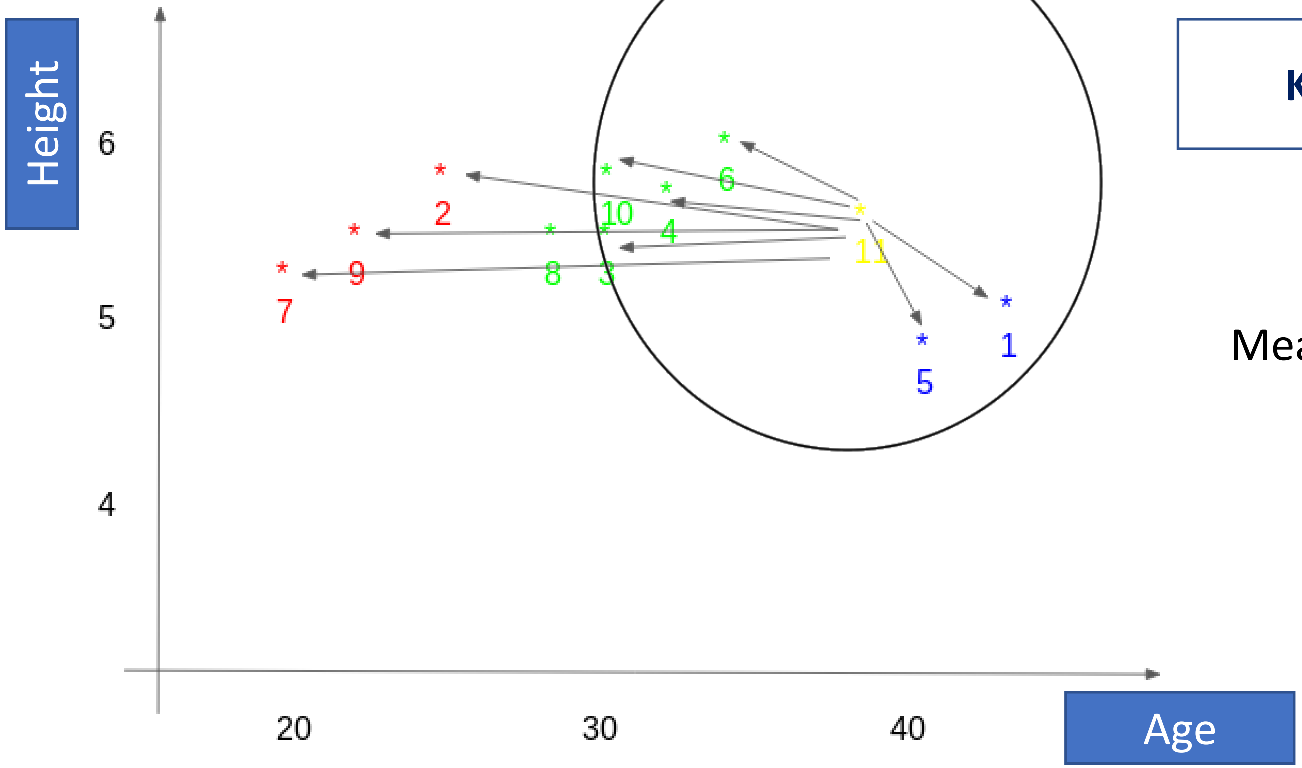
Mean: (77+72+60)/3
= 69.6 kg

K-NN Algorithm: Regression

Sequence:
a > e > f > d > j > c > h > b > i > g



	A	B	C	D	E
1	id	age	height	weight	distance
2	1	45	5	77	a
3	2	26	5.11	47	b
4	3	30	5.6	55	c
5	4	34	5.9	59	d
6	5	40	4.8	72	e
7	6	36	5.8	60	f
8	7	19	5.3	40	g
9	8	28	5.8	60	h
10	9	23	5.5	45	i
11	10	32	5.6	58	j
12	11	38	5.5	?	
13					



K = 5

Mean: (77+72+60+59+58)/5
= 65.2 kg

Let’s Do Self Assessment!