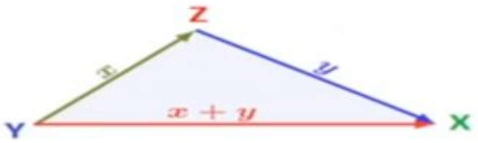
Semester : VII Subject : Big Data Analytics Academic Year: 2024 – 2025

Module 5:

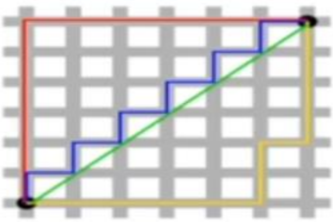
# Definition of a Distance Measure

A distance measure on this space is a function d(x, y) that takes two points in the space as arguments and produces a real number, and satisfies the following axioms:

1. $d(x, y) \geq 0$ (no negative distances).
2. $d(x, y) = 0$ if and only if $x = y$ (distances are positive, except for the distance from a point to itself).
3. $d(x, y) = d(y, x)$ (symmetric).
4. $d(x, y) \leq d(x, z) + d(z, y)$ (Triangle Inequality).




- L1-norm, or Manhattan distance
  - There, the distance between two points is the sum of the magnitudes of the differences in each dimension.



- L2-norm

$$d([x_1, x_2, \ldots, x_n],\ [y_1, y_2, \ldots, y_n]) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2}$$


$$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

# Lr-norm

A.K.A. Minkowski measure

The limit as $r$ approaches infinity of the $Lr - norm$.
As r gets larger, only the dimension with the largest difference matters, so formally, the $L\infty - norm$ is defined as the maximum of $|xi - yi|$ over all dimensions i.

$$d([x_1, x_2, \ldots, x_n], [y_1, y_2, \ldots, y_n]) = (\sum_{i=1}^{n} |x_i - y_i|^r)^{1/r}$$

# Numerical#01

Consider the two- dimensional Euclidean spaces (customary plane) and points (2,7) and (6,4).

The $L_1 - norm$ gives a distance of
$= |2-6| + |7-4|$
$= 4 + 3$
$= 7$
The $L_2 - norm$ gives a distance of
$= \sqrt{(2-6)^2+(7-4)^2}$
$= \sqrt{(4)^2+(3)^2}$
$= 5$
The $L_\infty - norm$ gives a distance of
$= \max(|2-6|, |7-4|)$
$= \max(4,3)$
$= 4$

# Jaccard Distance

1. d(x, y) is nonnegative because the size of the intersection cannot exceed the size of the union.
2. d(x, y) = 0 if x = y, because x ∪ x = x ∩ x = x. However, if x 6= y, then the size of x ∩ y is strictly less than the size of x ∪ y, so d(x, y) is strictly positive.
3. d(x, y) = d(y, x) because both union and intersection are symmetric; i.e., x ∪ y = y ∪ x and x ∩ y = y ∩ x.
4. Jaccard Similarity always satisfes triangular inequality, and therefore, so does Jaccard Distance.

## Numerical#01

Find the Jaccard distances between the following pairs of sets: {1, 2, 3, 4} and {2, 3, 4, 5}.

$$C1 = \{1, 2, 3, 4\}$$
$$C2 = \{2, 3, 4, 5\}$$
$$|C1 \cap C2| = 3$$
$$|C1 \cup C2| = 5$$
$$sim(C1, C2) = |C1 \cap C2| / |C1 \cup C2|$$
$$sim(C_1, C_2) = 3/5$$
$$d(C_1, C_2) = 1 - |C_1 \cap C_2| / |C_1 \cup C_2|$$
$$d(C_1, C_2) = 1 - 3/5$$
$$d(C_1, C_2) = 2/5 \text{ or } 40\% \text{ dissimilarity or distance}$$