4 Distance Measures for Machine Learning

1. **Hamming Distance**
2. Euclidean Distance
3. Manhattan Distance (Taxicab or City Block)
4. **Minkowski Distance**
5. **Chebyshev Distance**

https://machinelearningmastery.com/distance-measures-for-machine-learning/#:~:text=The%20Manhattan%20distance%2C%20also%20called,a%20chessboard%20or%20city%20blocks.

Hamming

Hamming distance

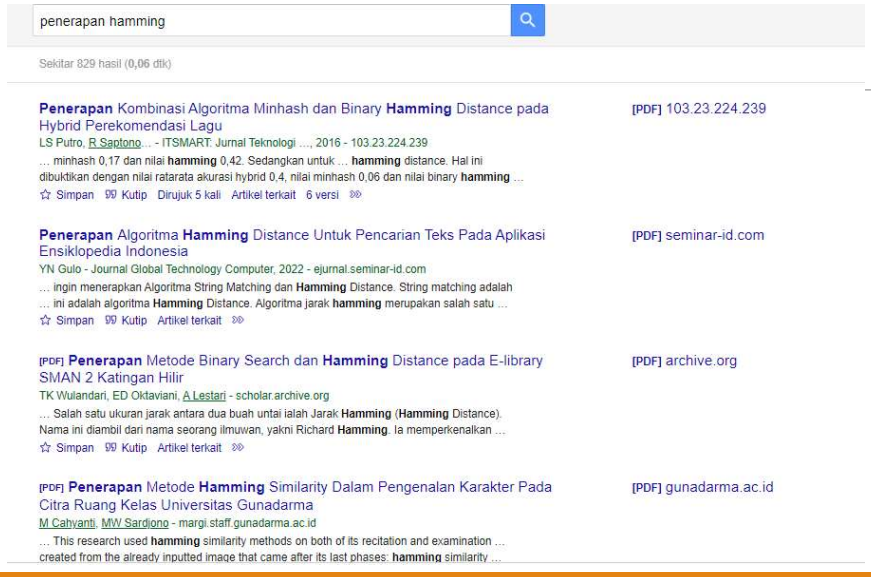
Jarak Hamming adalah metrik untuk membandingkan dua string data biner. Saat membandingkan dua string biner dengan panjang yang sama, jarak Hamming adalah jumlah posisi bit di mana kedua bit tersebut berbeda.

Jarak Hamming antara dua string, a dan b dilambangkan dengan d(a,b).

Ini digunakan untuk deteksi **kesalahan atau koreksi kesalahan ketika data ditransmisikan melalui jaringan komputer.** Ini juga digunakan dalam teori pengkodean untuk **membandingkan kata-kata data yang panjangnya sama**.

https://www.tutorialspoint.com/what-is-hamming-distance

Hamming untuk apa?



What Is Hamiing Distance Formula?

Hamming distance is a metric for comparing two binary data strings. While comparing two binary strings of equal length, Hamming distance is the number of bit positions in which the two bits are different.

The Hamming distance between two strings, a and b is denoted as d(a,b).

It is used for error detection or error correction when data is transmitted over computer networks. It is also using in coding theory for comparing equal length data words.

Hamming distance calculates the distance between two binary vectors, also referred to as binary strings or bitstrings for short.

You are most likely going to encounter bitstrings when you one-hot encode categorical columns of data.

For example, if a column had the categories '*red*,' '*green*,' and '*blue*,' you might one hot encode each example as a bitstring with one bit for each column.

- red = [1, 0, 0]
- green = [0, 1, 0]
- blue = [0, 0, 1]

The distance between red and green could be calculated as the sum or the average number of bit differences between the two bitstrings. This is the Hamming distance.

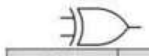
Examples Using Hamming Distance Formula

Example

Suppose there are two strings 1101 1001 and 1001 1101.

11011001 ⊕ 10011101 = 01000100. Since, this contains two 1s, the Hamming distance, d(11011001, 10011101) = 2.

XOR



INPUT		OUTPUT
A	B	
0	0	0
1	0	1
0	1	1
1	1	0

Minimum Hamming Distance

In a set of strings of equal lengths, the minimum Hamming distance is the smallest Hamming distance between all possible pairs of strings in that set.

Example

Suppose there are four strings 010, 011, 101 and 111.

010 ⊕ 011 = 001, d(010, 011) = 1.

010 ⊕ 101 = 111, d(010, 101) = 3.

010 ⊕ 111 = 101, d(010, 111) = 2.

011 ⊕ 101 = 110, d(011, 101) = 2.

011 ⊕ 111 = 100, d(011, 111) = 1.

101 ⊕ 111 = 010, d(011, 111) = 1.

Hence, the Minimum Hamming Distance, d_{min} = 1.

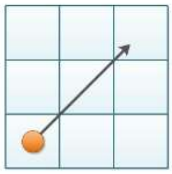
Examples Using Hamming Distance Formula

- "karolin" and "kathrin" is 3.
- "karolin" and "kerstin" is 3.
- 1011101 and 1001001 is 2.
- 2173896 and 2233796 is 3.

- For binary strings a and b the Hamming distance is equal to the number of ones in **a XOR b**.

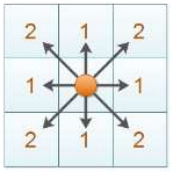
Minkowski

Euclidean Distance



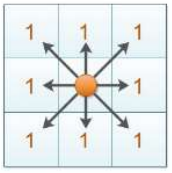
$\sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$

Manhattan Distance

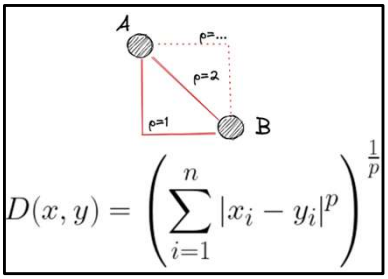


$|x_1 - x_2| + |y_1 - y_2|$

Chebyshev Distance



$\max(|x_1 - x_2|, |y_1 - y_2|)$



Minkowski Distance

Minkowski distance adalah metrik ukur yang digunakan untuk menghitung jarak antara dua vektor bernilai bilangan riil. Metrik ini adalah bentuk generalisasi dari Euclidean dan Manhattan distance dengan tambahan parameter yang disebut "order" atau p, yang memungkinkan pengukuran jarak yang berbeda untuk dihitung.

Perhitungan Minkowski distance adalah sebagai berikut:

$$d(x, y) = (\sum_{i=1}^{n} |x_i - y_i|^p)^{(\frac{1}{p})}$$

Dimana "p" adalah parameter order

Ketika p diatur ke 1, perhitungannya sama dengan jarak Manhattan. Ketika p diatur ke 2, akan sama dengan jarak Euclidean.

- p=1: jarak Manhattan.
- p=2: Jarak Euclidean.

Minkowski

- Minkowski Distance merupakan generalisasi dari Euclidean Distance

$$d(\mathbf{x}, \mathbf{y}) = \left(\sum_{k=1}^{n} |x_k - y_k|^r \right)^{1/r}$$

r = parameter

r = 1 = city block

r = 2 = Euclidean distance

r = ∞ = Chebyshev distance

- Lebih mudahnya
- $d(A, B) = \sqrt[r]{\sum_{k=1}^{n} |A_k - B_k|^r}$

Minkowski Distance

$$d(A, B) = \sqrt[r]{\sum_{k=1}^{n} |A_k - B_k|^r}$$

Object	Feature 1	Feature 2	Feature 3	Feature 4
A	0	3	4	5
B	7	6	3	-1

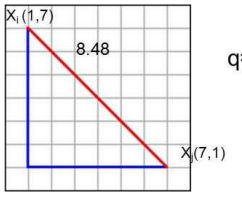
- Dengan menggunakan r sebesar 3, maka minkowski distance adalah
- $d_{BA} = \sqrt[3]{|0-7|^3 + |3-6|^3 + |4-3|^3 + |5+1|^3}$
- $d_{BA} = \sqrt[3]{343 + 27 + 1 + 216} = \sqrt[3]{587} = 8.373$

Minkowski

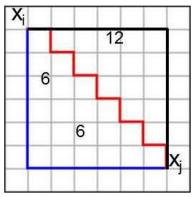
- Minkowski distance: a generalization

$$d(i,j) = \sqrt[q]{|x_{i1} - x_{j1}|^q + |x_{i2} - x_{j2}|^q + \ldots + |x_{ip} - x_{jp}|^q} \quad (q > 0)$$

- If q = 2, d is Euclidean distance
- If q = 1, d is Manhattan distance



q=2



q=1

Chebyshev

Chebyshev distance

In case of $q \to \infty$, the distance equals to the maximum difference of the attributes. Useful if the worst case must be avoided:

$$d_\infty(X,Y) = \lim_{q\to\infty}\left(\sum_{i=1}^{n}|x_i - y_i|^q\right)^{1/q}$$

$$D_{\text{Chebyshev}} = \max\left(|x_2 - x_1|, |y_2 - y_1|\right).$$

Example:

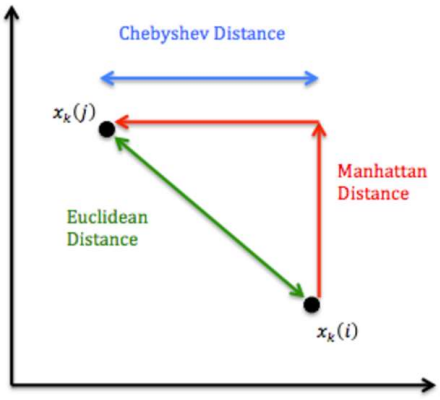
$$d_\infty((2,8),(6,3)) = \max(|2-6|,|8-3|) = \max(4,5) = 5$$

Chebyshev Distance

- Berbeda dengan kedua teknik pengukuran jarak sebelumnya, Chebyshev distance didefinisikan sebagai perbedaan terbesar antara dua vektor di sepanjang dimensi koordinat apa pun.
- Maksudnya, teknik ini hanyalah jarak maksimum dari suatu sumbu.

$$d(P,Q) = \max|P - Q| = max|p_i - q_i|$$

- Pada rumus tersebut kita bisa melihat bahwa cara kerja Chebyshev distance adalah mencari selisih maksimum di antara kedua vektor.
- Selisih tersebut kemudian diabsolutkan untuk mencegah terjadinya minus. Karena tidak mungkin ada jarak yang minus dalam kehidupan nyata.



The Chebyshev Distance is also called the Chessboard Distance or the L Infinity Distance. The most intuitive understanding of the Chebyshev distance is the movement of the King on a chessboard: it can go one step in any direction (up, down, left, right and verticals).

CHEBYSHEV (CHESSBOARD) DISTANCE

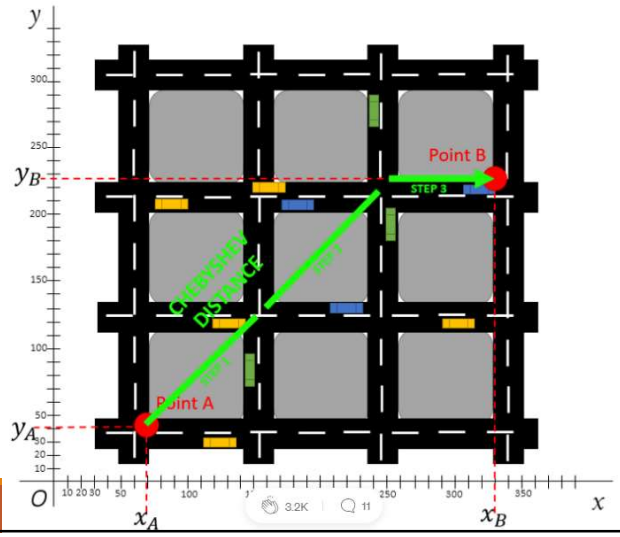
$$dist(A,B) = \max(|x_A - x_B|, |y_A - y_B|)$$

$$dist(A,B) = \max(|70 - 330|, |40 - 220|)$$

$$dist(A,B) = \max(|-260|, |-188|)$$

$$dist(A,B) = \max(260, 188)$$

$$dist(A,B) = 260$$



Computing the Chebyshev Distance from point A to point B.

Chebyshev Distance

The Chebyshev distance calculation, commonly known as the "maximum metric" in mathematics, measures distance between two points as the maximum difference over any of their axis values. In a 2D grid, for instance, if we have two points (x1, y1), and (x2, y2), the Chebyshev distance between is *max(y2 - y1, x2 - x1)*.

(0, 3)	(3, 1)	(3, 2)	(3, 3)
(0, 2)	(2, 1)	(2, 2)	(2, 3)
(0, 1)	(1, 1)	(1, 2)	(1, 3)
(0,0)	(0, 1)	(0, 2)	(0, 3)

On the above grid, the difference in the x-value of the two red points is *2-0=**2***, and the difference in the y-values is *3-0=**3***. The maximum of 2 and 3 is 3, and thus the Chebyshev distance between the two points is 3 units.