01

3 mins read

DIFFERENT DISTANCE MEASURES FOR MACHINE LEARNING

@learn.machinelearning

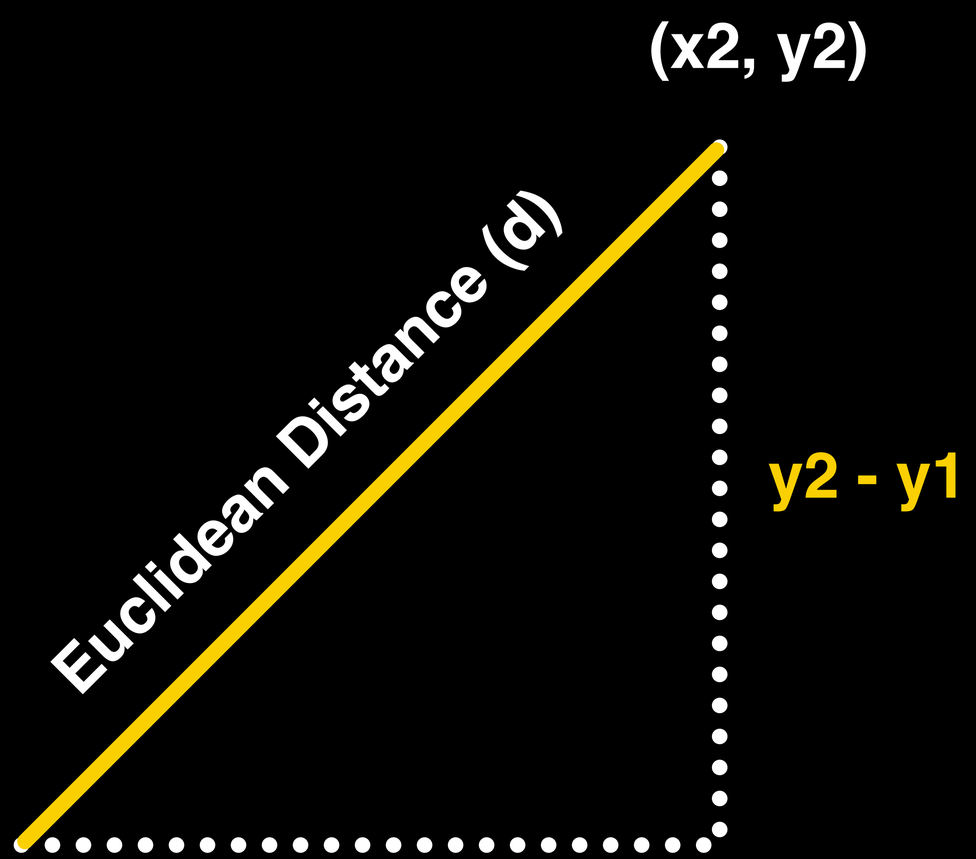
Why distance metrics?

Distance metrics play an important role in machine learning. Both supervised and unsupervised algorithms in machine learning use Distance Metrics to understand patterns in the/ similarity between input data. These metrics are also used for the identification of similarities between results.

An effective distance measure improves the performance of our machine learning model, whether that's for classification tasks or clustering. It is very important to know which distance measure to use for a given data. In the next few slides, we will look at different distance measures.

Algorithms which use distance measures at their core.

- K-Nearest Neighbors
- Learning Vector Quantization (LVQ)
- Self-Organizing Map (SOM)
- K-Means Clustering



$$D(x, y) = \sqrt{\sum_{i=1}^{n}(x_i - y_i)^2)}$$

EUCLIDEAN DISTANCE

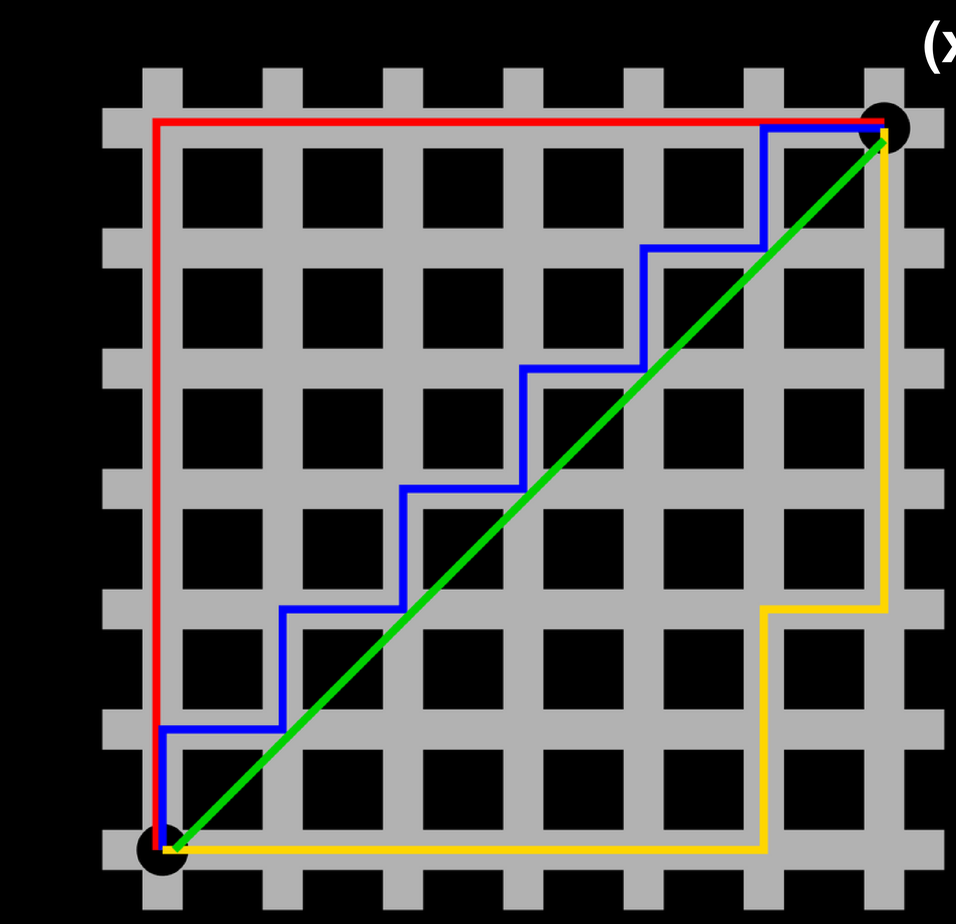
Euclidean Distance represents the shortest distance between two points. It is calculated as the square root of the sum of differences between each point. In simple words, Euclidean distance is the length of the line segment connecting the points.

Disadvantages

- It is better to normalize the data because this metric might be skewed when there is a difference in the units of features.
- It also fails when the dimensionality increases because of the curse of dimensionality.

When to use

- It works best when we have low dimensional data with methods like KNN and HBDSCAN



Source - geeksforgeeks

$$D(x, y) = \sum_{i=1}^{k} |x_i - y_i|$$

MANHATTAN DISTANCE

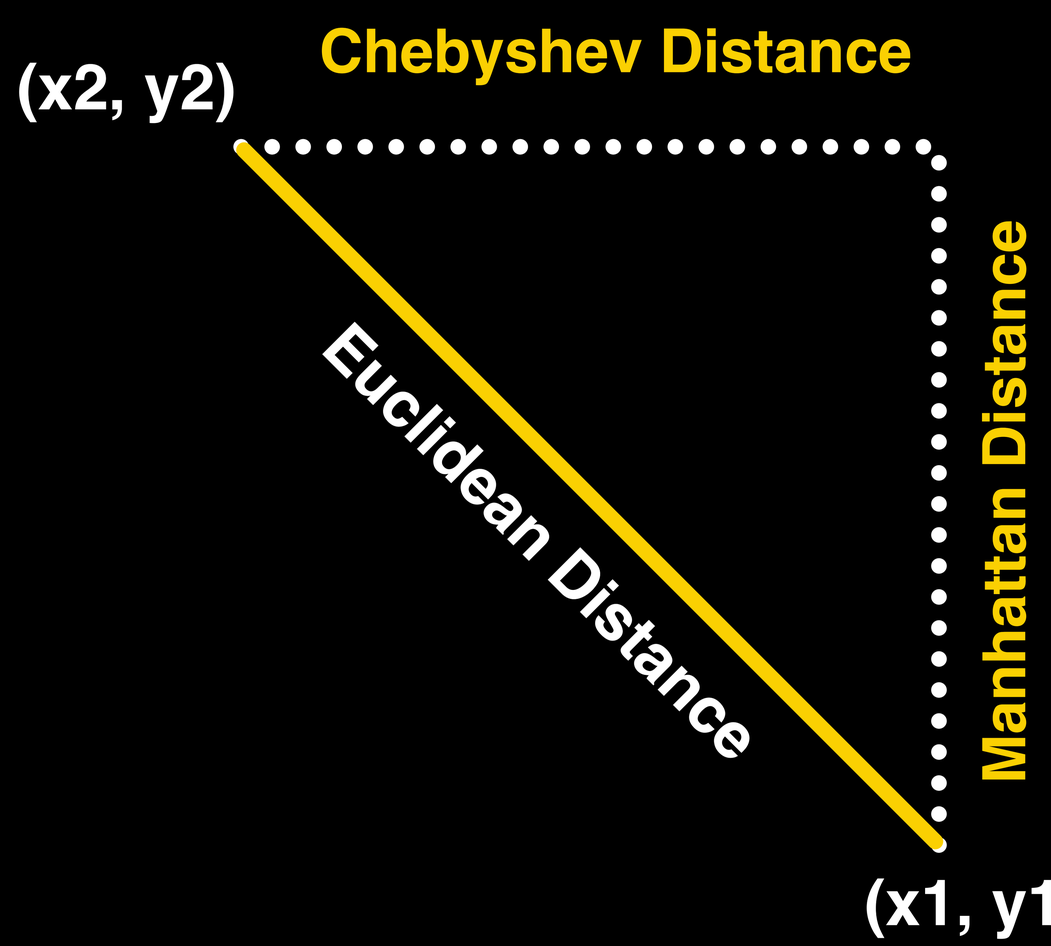
Often called Taxicab distance or City Block distance. It is the sum of absolute differences between points across all the dimensions. It is perhaps more useful to vectors that describe objects on a uniform grid, like a chessboard or city blocks. There is no diagonal movement involved in calculating the distance.

Disadvantages

- It also fails when the dimensionality increases.
- it is more likely to give a higher distance value than euclidean distance since it does not the shortest path possible.

When to use

- When your dataset has discrete and/or binary attributes, Manhattan seems to work quite well



$$D(x, y) = \max_i \left(|x_i - y_i|\right)$$

CHEBYSHEV DISTANCE

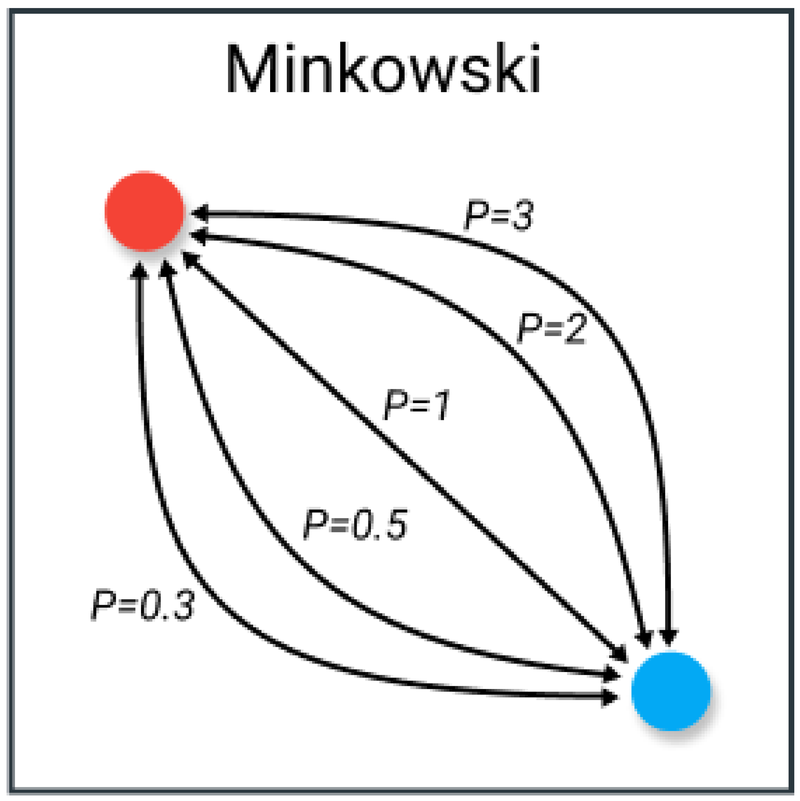
It is calculated as the maximum of the absolute difference between the elements of the vectors.

Disadvantages

- We cannot use it in all use cases as it is specific to few use-cases

When to use

- Chebyshev distance is often used in warehouse logistics as it closely resembles the time an overhead crane takes to move an object.



Source - medium (marteen)

$$Minkowski = \sqrt[p]{\sum_{i=1}^{n} |x_i - y_i|^p}$$

MINKOWSKI DISTANCE

It is the generalized form of the Euclidean and Manhattan Distance Measure. Here, p represents the order of the norm.

- For, p=1, the distance measure is the Manhattan measure.
- p=2, the distance measure is the Euclidean measure.
- p = ∞, the distance measure is the Chebyshev measure.

Disadvantages

- It has the same disadvantages as Euclidean and Manhattan.
- Finding the right P value can be quite computationally inefficient

When to use

- Because of different P value we can find the distance measure that works best for your use-case.

HAMMING DISTANCE

We use hamming distance if we need to deal with categorical attributes. Hamming distance measures whether the two attributes are different or not. When they are equal, the distance is 0; otherwise, it is 1. We can use hamming distance only if the strings are of equal length.

Example

- Let's consider two one-hot encoded vectors of 2 strings.
- Euclidean - [1,0,0,0,0]
- Manhattan - [0,0,1,0,0]
- The distance between 2 vectors could be calculated as the sum or the average number of bit differences between the two strings
- So, here the distance is 1

Disadvantages

- It will not work when the two vectors are of different lengths.
- It does not take the actual value into account as long as they are different or equal.

When to use

- It is mostly used to measure the distance between categorical variables.