What is Euclidean distance? What is Manhattan/Hamming Distance?

Data mining procedures are usually based on identifying the things that are similar and grouping them based on the similarity. This similarity can often be helpful in identifying solutions to many business problems. Below are few examples of business tasks that use similarity.A paper company ‘A’ can identify business similar to their top clients for helping their sales staff to identify a potential customer.Hewlett-Packard maintains many highly configured servers for clients which is aided by a tool that can give servers with similar configuration when a server configuration is passed as input.Similarity between the objects can also be identified based on the distance between them when they are plotted as data by representing each object as a vector. The closer the objects are, the similar their features tend to be. There are few ways to measure the distance or similarity between the objects such as Euclidean distance and Manhattan distance.

If an object is described by ‘n’ features i.e n dimensions d1,d2,d3….. dn. The general equation for Euclidean distance between two objects A and B can be defined as.



Let us look at a simple example that calculates distance between two objects. In this case, similarity between the persons with multiple attributes from a simplified credit application program.

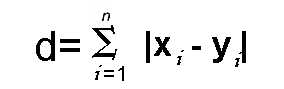
|  |  |  |
| --- | --- | --- |
| **Attribute** | **Person A** | **Person B** |
| Age | 25 | 30 |
| Work experience | 5 | 10 |
| sex (male 1 female 2) | 2 | 1 |

From the above definition for the Euclidean distance, the similarity between the two persons can be calculated as

d(A,B) = √ ((30-25)2+(10-5)2+ (1-2)2) = 25 + 25 + 1 = √25+25+1= √51=7.141

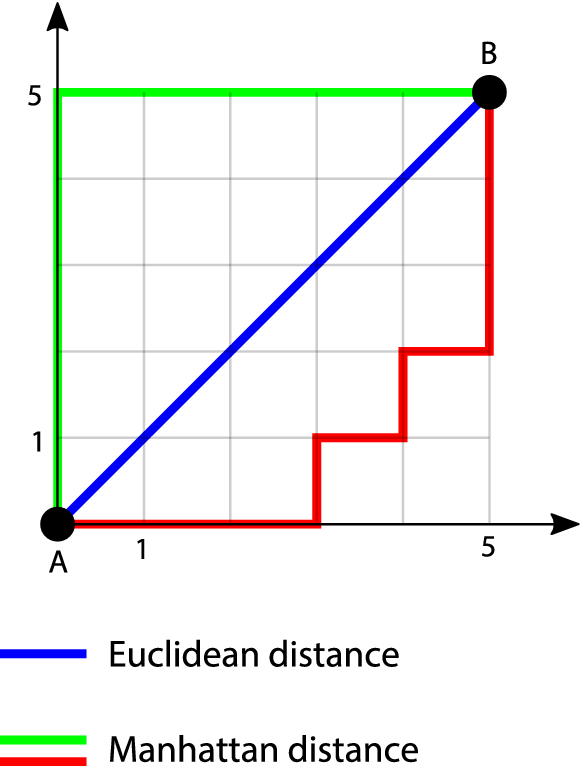
Manhattan distance which is also called Taxicab distance, or the city block distance is primarily used when we want to calculate the distance between two points in a grid like path. It is named as Manhattan distance because it represents the total distance one has to travel in a place like Manhattan between two points which is arranged in a grid like structure. Let’s assume we want to calculate distance, d, between two data points x and y. Distance d will be calculated using an absolute sum of difference between its co-ordinates.

d=(x1 - y1) + (x2 - y2) + (x3 - y3) + … + (xn - yn).



Below illustration shows the Euclidean distance and the Manhattan distance in a two-dimensional area.

Euclidean distance is calculated as √ 52 + 52= 7.07. Manhattan distance= 5+ 5 =10



Hamming distance is another metric which can be used to compare two binary strings. When comparing two binary strings of same length, hamming distance is the number of bit positions where the bits are different.

To calculate the hamming distance between two binary strings, we need to perform XOR operation and then count the number of 1s in the resultant string. For example, hamming distance between two binary strings 00100110 and 10101100 is

00100110 ⊕10101100=10001010. As this output has a total of three 1s,

the Hamming distance d(00100110,10101100) = 3.

Euclidean distance is the most popular and widely used distance metric in data science. But Manhattan distance is preferred over Euclidean distance where there are many dimensions in the data.