Mikailalsys Journal of Mathematics and Statistics

e-ISSN : 3030-816X
p-ISSN : 3030-8399

Index : Harvard, Boston, Sydney University, Dimensions, Lens, Scilit, Semantic, Google, etc

https://doi.org/10.58578/MJMS.v1i1.1990

DISTANCE METRICS FOR MACHINE LEARNING AND IT'S RELATION WITH OTHER DISTANCES

Dipendra Prasad Yadav[1], Nand Kishor Kumar[2], Suresh Kumar Sahani[3]
[1]Thakur Ram Campus, Birgunj, Nepal; [2]Trichandra Campus, Tribhuvan University, Nepal
[3]M.I.T. Campus, T.U, Janakpur, Nepal
dipendra.yadav2032@gmail.com; nandkishorkumar2025@gmail.com

Article Info:

Submitted:	Revised:	Accepted:	Published:
Sep 27, 2023	Oct 17, 2023	Oct 23, 2023	Oct 27, 2023

Abstract

In machine learning, distance metrics play a crucial role in measuring the degree of dissimilarity among data points. When creating and optimizing machine learning models, data scientists and machine learning practitioners can make more informed choices by understanding the features of popular distance metrics and their relationships. The effectiveness and interpretability of the model's output can be greatly influenced by selecting the appropriate distance metric. We explain distance metrics and their relevance in machine learning with various examples of metrics, including Minkowski distance, Manhattan distance, Max Metric for R^n, Taxicab distance, Relative distance, and Hamming distance.

Keywords: Distance Metrics, Minkowski Distance, Manhattan Distance, Max Metric for R^n, Taxicab Distance, Relative Distance, Hamming Distance

Introduction

Distance metrics play a fundamental role in various aspects of machine learning, particularly in tasks involving clustering, classification, and recommendation systems. These metrics quantify the dissimilarity between data points and are essential for algorithms like k-means clustering, hierarchical clustering, k-nearest neighbors, and more. Understanding different distance metrics and their relationships with each other is crucial for building effective machine learning models.

In the context of machine learning, a distance metric (also known as a similarity metric) is a function that quantifies the "distance" or dissimilarity between two data points in a feature space. This distance is a numerical representation of how far apart or dissimilar the data points are. The choice of distance metric can significantly impact the performance of a machine learning model, as it influences the clustering, classification, or recommendation results [1].

What is the Janakpur to Birgunj distance in meters?

How similar are two siblings?

Which Bhajan ought to be suggested by my software?

Each of these inquiries appears to be looking for a very distinct insight. But when they are replied from the data domain, they all have one thing in common: they can all be answered with the same family of metrics ⟶Distance Metrics.

Objectives:

The following are the aims of this paper:

(a) What distance measurements are used in the field of machine learning?
(b) What makes them pertinent?
(c) How do they impact models and algorithms?
(d) How do other lengths relate to distance metrics?

Google maps route between Janakpur and Birgunj

To compare two data points or observations and determine how comparable they are, we must compute some kind of metric. Distance metrics compute the distance between two places to numerically assess their similarity. A little result from the distance metric computation indicates similarity between the two places; a large result indicates difference. Easy, huh? You may think, Okay, that's cool. What, though, is the fuss all about? Two items are always the same distance apart. Why is this such a big deal? My buddy, the solution is sometimes more complicated than what you might naturally assume. On seeing the above Google Maps as an illustration. However, Birgunj is 150 km from Janakpur but the taxtcab distance between them is very shorter than road distance via Chandrapur and Nijgadh.

The distance we cover when traveling from Janakpur to Birgunj yields varying results. Why does this occur? It indicates that there are various methods for measuring distance, each producing a unique set of outcomes. Why does machine learning explain this well and why is it significant to us?

Need of Distance Metrics in Machine Learning

These distance metrics are utilized in many of our favorite algorithms, such as K-Nearest Neighbors, Classification, K-Means Clustering, and Self-Organizing Maps (SOM), which is why we should be concerned about this. Certain Kernel Algorithms, such as Support Vector Machine, employ computations that are also categorized as "distance calculations."

Because of this, it is important to comprehend the reasoning behind each distance measure in order to choose when to utilize it in conjunction with another, as this can significantly affect the outcomes of earlier algorithms or models.

It is helpful to at least have a rudimentary understanding of what each distance metric signifies because we may choose to utilize one over another depending on the type of data, the method we are using, and the thinking behind it. Let's examine the most popular metrics now that we understand what distance measures are and why they are so important.

Let a function is defined on real line **R.** The distance d which links

d(x, y) = | x- y| for every x, y ∈**R.**

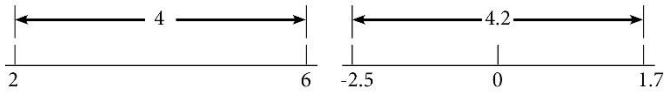


Fig.1. Distance on **R**

In functional analysis, spaces and functions are defined and studied.We replace the set of real numbers underlying **R** by an abstract set X and introduce on X a "distance function" which has only a few of the most fundamental properties of the distance function on **R** [1].

Definition and Some Examples [1]

A metric space is a (X, d) where X is a set and d is a metric on X (or distance function on X), defined on X × X such that for all x , y , z ∈X

- d is real- valued, finite and nonnegative.
- d (x, y) = 0 if and only if x = y.
- d (x,y) = d (y,x) (**Symmetry**)
- d(x,y) ≦ d (x, z) + d(z,y) (**Triangle inequality**)

Let X be a set and d: X × X → $\mathbb{R}^+$ a function from X × X to the set $\mathbb{R}^+$of non-negative real numbers satisfying the following axioms for x, y, z ∈ X :

- d (x, y) = 0 if and only if x = y.
- d (x, y) = d (y, x) (**Symmetry**)
- d(x, z) ≦ d (x, y) + d(y, z) (**Triangle inequality**)

Then d is called a metric or distance function on X and d(x, y) is called the distance from x to y. The set X with metric d is called a metric space and represented by (X, d).

In this article, we will discuss about different metrics and how they are related. The distance between two points A and B, from Pythagoras theorem, the length of x and y-axis are

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

Distance function is a mathematical formula used by distance metrics. The distance function can differ across different distance metrics. Now the different distance metrics and their relations are described here.

Minkowski Distance [6]

The distance which is calculated from the formula

$$(\sum_{i=1}^{n} |x_i - y_i|^p)^{\frac{1}{p}} \quad \text{(i)}$$

is known as Minkowski distance. It is generalized distance metrics. This generalization is the manipulation of formula (i)to calculate the distance between two points in different ways.

When p = 1, the equation (i) yields the expression for Manhattan distance.

When p = 2, the equation () yields the expression for Euclidean distance.

When p = ∞, the equation (i) yields the expression for Chebychev distance.

Manhattan distance

The distance between two data points x and y in a grid of path from

$$d = \sum_{i=1}^{n} |x_i - y_i| \quad \text{(ii)}$$

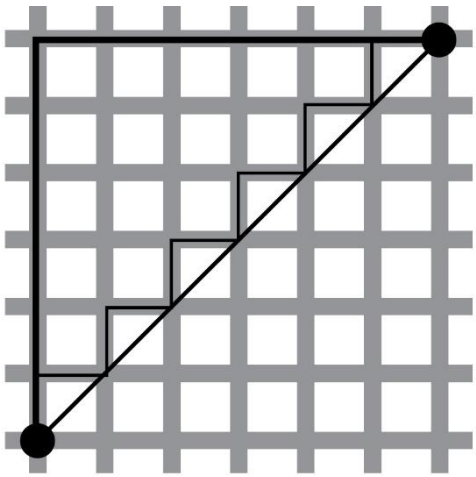
where n is the number of variables $x_i \in x$, $y_i \in y$ defined for

Fig 2

x = $x_1, x_2, x_3 \cdots\cdots$& y = $y_1, y_2, y_3 \cdots\cdots$

$$d = (x_1 - y_2) + (x_2 - y_2) + \cdots\cdots\cdots + (x_n - y_n) \quad \text{(iii)}$$

This visualization of Manhattan Distance justifies the Taxicab geometry, City block distance [3]



Source: Taxicab geometry Wikipedia

Fig 3

Taxicab Metric for $\mathbb{R}^n$

The equation (ii) is defined for a function d' on $\mathbb{R}^n \times \mathbb{R}^n$ as

$$d'(x, y) = \sum_{i=1}^{n} |x_i - y_i| \text{ where } x_i \in x \ , \ y_i \in y \quad \text{(iv)}$$

This is called taxicab metric for $\mathbb{R}^n$. Form figure (ii) , in the plane , the distance from x to y is the sum of length of horizontal segment and vertical segment joining x to y[2].

Max Metric for $\mathbb{R}^n$

The other metric d'' for $\mathbb{R}^n$is defined maximum of the absolute values of the differences of the coordinates of x and y

d'' (x, y) = max $\{|x_i - y_i|\}_{i=1}^{n}$ where $x_i \in x$, $y_i \in y$

(v)

d'' satisfies all the properties defined in the definition, also satisfies triangle inequality [2]

Euclidean Distance [3,5]

Euclidean distance is mostly and safely used and calculated between two given points. It makes sense when all the dimensions have same units (like meter or kilometer). It is calculated by using Minkoski Distance formula when p=2 and got the form,

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

(vi)

$\boxed{x \leftrightarrow y}$

Euclidean

Cosine Distance [3,4]

In cosine metric, the degree of angle is measured between two vectors and is defined from the equation of dot products as

a.b= $\|\vec{a}\|\|\vec{b}\| \cos\theta$

or, $\cos\theta = \frac{\vec{a}.\vec{b}}{\|\vec{a}\|\|\vec{b}\|}$

where cos0'=1; for same direction and having similarity between two points.

Cos90'= 0 ; for orthogonal vectors and unrelated.

Cos180' = -1; for opposite directions and have no similarity.

Vector Version of Distance

Vector version of distance is a distance with both magnitude (size) and direction.

Mahalanobis Distance

Mahalanobis distance is used to select the models of both men and women to decide which is closer and probable. It is used for calculating distance between two in multivariate space. The d,istance between an observation and mean by using following formula as:

$$D_M(\vec{x}) = \sqrt{(\vec{x}-\vec{\mu})^T S^{-1}(\vec{x}-\vec{\mu})}$$

where S is covariance metrics. The variance-normalized distance equation is derived from inverse of covariance [3].

Lavrentiev[7] described relative distance function in mapping between the frontier under a conformal mapping of a simply connected domain onto the unit circle.

Relative distance [7]

Suppose Ω be a connected domain and let F is the frontier of Ω. Let a,b ∈ Ω and defined $\rho(a,b)$ is the greatest lower bound (glb) of the lengths of all polygonal paths connecting a to b in Ω. It is understood that (a,b) is a metric c in Ω and $\rho(a,b) \geq |a-b|$ with equality if a,b lie in some convex sun-domain of Ω.

For every, a ∈ Ω and c ∈ F then $\rho(a,b)$ to be the infimum of lim $\rho(a,\mathcal{Z}_n)$ of all sequences

$\{\mathcal{Z}_n\} \to c.$

Haming Distance [8]

Haming distance is the one of several string metrics for measuring the edit distance between two sequences. This distance is named after the American mathematician Richard W. Haming.It is applied in coding theory and in also in block codes, where the equal length string are vectors over a finite field.

Let X be the set of all ordered triples of zeros and ones. X consists of eight elements and a metric d on X is defined by

d(x,y) = No. of places where x and y have different entries. It is symbolized by letters, bits or decimal digits as:

Haming distance between 0000 and 1111 is 4. The haming distance "Nepal" and "Nepal" is 3.The formula for the haming distance

d_H (x,y) = $\sum i$ = 0k-1$\sum j$= 0, j≠ i, K-1a_{ij}. This summation describes all the off-diagonal elements of A that indicate the positions where x and y differ.

Selecting the Right Distance Metric

Selecting the most appropriate distance metric for our machine learning task is a critical step in model development. The choice should be guided by the nature of the data, the problem we are trying to solve, and the specific algorithm we intend to use. It may involve experimenting with different metrics and evaluating their impact on the model's performance.

Conclusions

Distance metrics are essential tools in machine learning for quantifying dissimilarity between data points. Understanding the characteristics of common metrics and their relationships with each other allows data scientists and machine learning practitioners to make informed decisions when developing and fine-tuning machine learning models. The choice of the right distance metric can significantly influence the effectiveness and interpretability of the model's results.

References

[1] Kreyszing, E. (1973). *Introductory Functional Analysis with Applications*. New York: John Wiley & Sons, 1-7.

[2] Croom, F. H. (1989). *Principles of Topology*. USA: Rinehart and Winston, Ins. 4,55-60.

[3] Sharma, N. (2019). *Importance of Distance Metrics in Machine Learning Modeling*. Towards Data Science.

[4]Cosine Similarity- Sklearn TDS article Wikipedia Example.

[5]Distance Metrics- Math.net, Wiki

[6]Minkowski Distance Metric- Wiki, Blog, Famous Metrics.

[7] Lavrentiev,M. (1929). Sur la correspondence entre les frontiers dans la representation conform. *Rec.Math*,36,112-115.

[8] Haming, R.W. (1950). Error detecting and error correcting codes. The Bell System Technical Journal,29(2),147-160.

.