1. Distance Measures in ML

These distances are used in various ML Algorithms. In machine learning, various types of distances or metrics are used to quantify the dissimilarity or similarity between data points. These distance metrics are fundamental in many machine learning algorithms, including clustering, nearest neighbor algorithms, and dimensionality reduction techniques. Here are some common types of distances used in machine learning:

Euclidean Distance: The Euclidean distance is the most common distance metric. It measures the straight-line distance between two points in Euclidean space. For two points, A and B, in n-dimensional space, the Euclidean distance is calculated as:

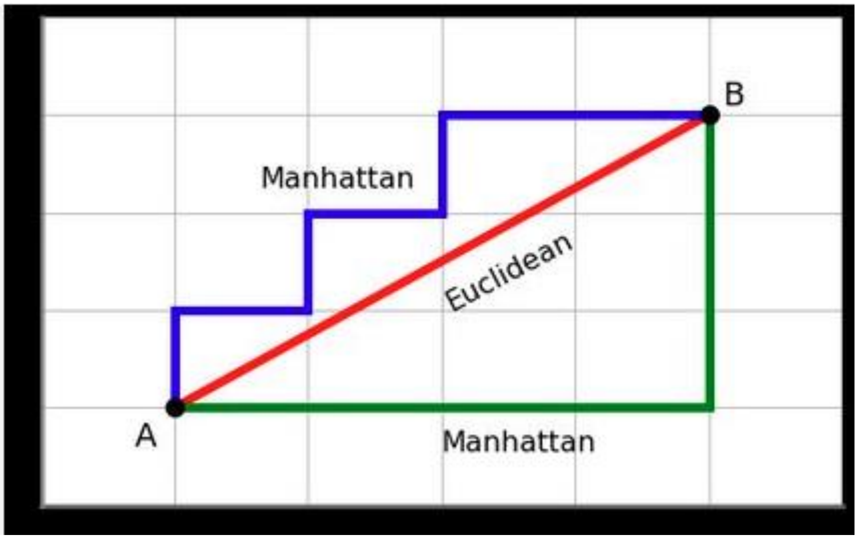
Euclidean Distance = sqrt((x2 - x1)^2 + (y2 - y1)^2 + ... + (xn - xn)^2)

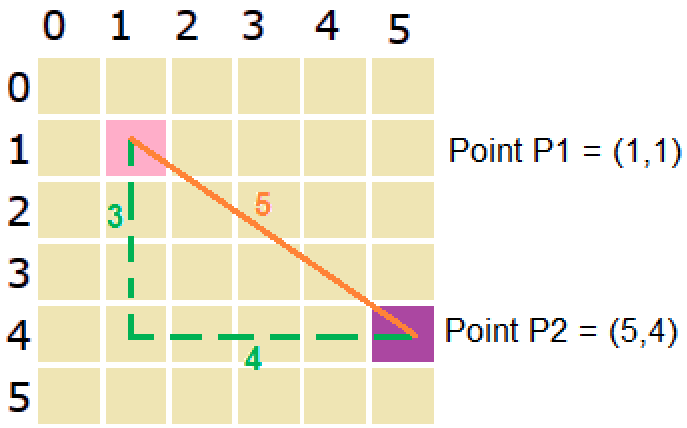
Euclidean distance is sensitive to differences in all dimensions.

Manhattan Distance: Also known as the L1 distance or taxicab distance, Manhattan distance measures the sum of the absolute differences between the coordinates of two points. For two points, A and B, in n-dimensional space, the Manhattan distance is calculated as:

Manhattan Distance = |x2 - x1| + |y2 - y1| + ... + |xn - xn|

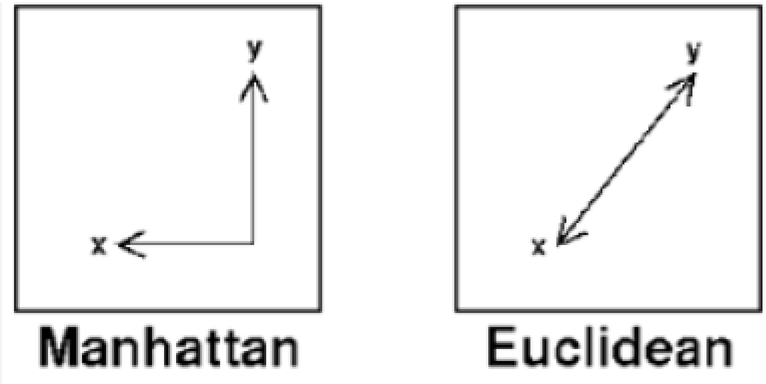
Manhattan distance is less sensitive to outliers and is commonly used in feature selection.





Euclidean distance = $\sqrt{(5-1)^2 + (4-1)^2}$ = 5

Manhattan distance = |5-1| + |4-1| = 7



Minkowski Distance: Minkowski distance is a generalization of various distance metrics used to measure the distance between two points in a multi-dimensional space. It is named after the German mathematician Hermann Minkowski, who made significant contributions to the field of geometry and the theory of numbers. The Minkowski distance between two points, denoted as P and Q, in a multi-dimensional space is defined as:

$$D(\mathbf{P},\mathbf{Q}) = \left(\sum_{i=1}^{n} |p_i - q_i|^p\right)^{\frac{1}{p}}$$

Here's what the variables and terms in the formula represent:

- D(P, Q) is the Minkowski distance between points P and Q.
- p is a parameter that determines the type of Minkowski distance. When p = 1, it corresponds to the Manhattan distance (L1 norm); when p = 2, it corresponds to the Euclidean distance (L2 norm); and for other values of p, it represents a generalized Minkowski distance.
- n is the number of dimensions in the multi-dimensional space.
- p_i and q_i are the respective coordinates of the points P and Q in each dimension.

Here are some specific cases of Minkowski distance based on the value of p:

- When p = 1, it calculates the Manhattan distance, which is the sum of the absolute differences between coordinates along each dimension. It is also known as the "taxicab distance" or "city block distance."
- When p = 2, it calculates the Euclidean distance, which is the straight-line distance between two points in the space.
- When p is greater than 2, it represents a generalized Minkowski distance, which is less common but can be used when different distance measures are needed based on the application.

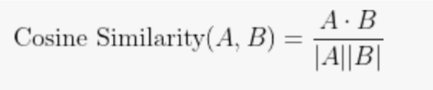
Minkowski distance is a versatile metric and can be applied in various fields, including machine learning, clustering, and pattern recognition, where measuring the distance or dissimilarity between data points in a multi-dimensional space is required. The choice of p depends on the specific problem and the characteristics of the data.

Cosine Similarity: Cosine similarity measures the cosine of the angle between two vectors in multi-dimensional space. It is often used in natural language processing (NLP) for comparing documents or text data. It ranges from -1 (perfectly dissimilar) to 1 (perfectly similar).

Cosine similarity is a measure of similarity between two non-zero vectors in an inner product space. In the context of text analysis, each vector typically represents the frequency of terms (words) in a document or a data point.

The cosine similarity between two vectors A and B is calculated as the cosine of the angle between these vectors in the multi-dimensional space. It is represented by the formula:

Cosine Similarity Formula



$$\text{Cosine Similarity}(A, B) = \frac{A \cdot B}{|A||B|}$$

Here, "A · B" represents the dot product of vectors A and B, and "||A||" and "||B||" represent the Euclidean norms (lengths) of vectors A and B, respectively.

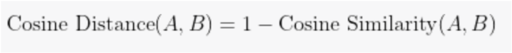
Cosine similarity returns a value between -1 and 1, where 1 indicates that the vectors are perfectly similar (pointing in the same direction), 0 indicates orthogonality (no similarity), and -1 indicates perfect dissimilarity (pointing in opposite directions).

Cosine Distance:

Cosine distance is essentially the complement of cosine similarity. It measures the dissimilarity or distance between two vectors in the same space.

To calculate cosine distance, you subtract the cosine similarity value from 1:

Cosine Distance Formula:



$$\text{Cosine Distance}(A, B) = 1 - \text{Cosine Similarity}(A, B)$$

Cosine distance returns a value between 0 and 2, where 0 indicates that the vectors are perfectly similar, and 2 indicates that they are maximally dissimilar.

In practice, cosine similarity is often used when you want to compare the similarity between two vectors or documents, while cosine distance is used when you want to measure their dissimilarity. These concepts are particularly useful in text analysis, recommendation systems, and clustering algorithms to find similarities or dissimilarities between textual data or feature vectors.

Note: If we have to check the alignment between data points then cosine similarity makes more sense. Or for data points in higher dimensional spaces. Else we can use Manhattan or Euclidian distance.

Jaccard Similarity: Jaccard similarity is used for comparing sets. It calculates the size of the intersection of two sets divided by the size of their union. It is commonly used in recommendation systems and text analysis.

Hamming Distance: Hamming distance is primarily used for comparing binary data or strings of equal length. It calculates the number of positions at which two binary strings differ.

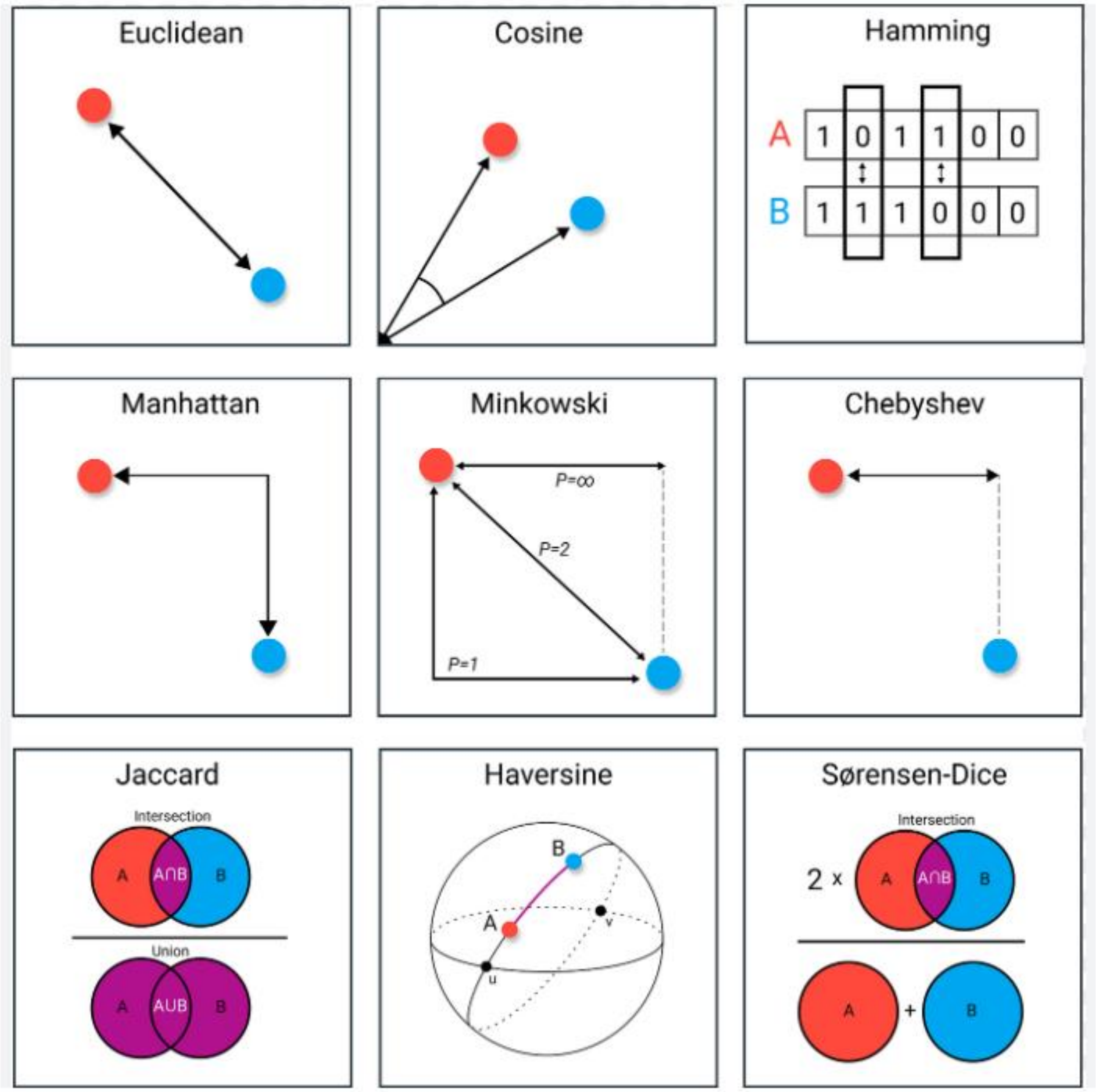
Mahalanobis Distance: Mahalanobis distance takes into account the correlations between dimensions of the data. It is used in multivariate statistics and is particularly useful when dealing with data with different scales and correlations.

Chebyshev Distance: Also known as the maximum value distance, Chebyshev distance measures the maximum absolute difference between corresponding coordinates of two points. It is sensitive to the largest difference along any dimension.

Correlation Distance: Correlation distance measures the similarity between two vectors by considering their correlation coefficients. It is used when the magnitude of data is not important, and only the relationships between variables matter.

Geodesic Distance: Geodesic distance is used on curved surfaces, such as the Earth's surface. It measures the shortest path between two points on a curved surface.

The choice of distance metric depends on the nature of the data, the problem you are trying to solve, and the characteristics of the machine learning algorithm you are using. Selecting an appropriate distance metric is essential for obtaining meaningful results in various machine learning tasks.



2. When to use Manhattan vs Euclidean Distance?

The choice between Manhattan distance and Euclidean distance depends on the characteristics of your data and the specific problem you are trying to solve in machine learning or data analysis. Here are some guidelines on when to use each distance metric:

Use Manhattan Distance When:

Data Lies on a Grid: Manhattan distance is particularly suitable for data that lies on a grid, such as when working with images, grids of pixels, or city block data.

Features Have Different Units or Scales: When the features in your dataset have different units or scales, Manhattan distance can be more appropriate because it treats each dimension equally and is less sensitive to the differences in scales.

Categorical or Binary Data: Manhattan distance is often used with categorical or binary data because it measures the "taxicab" distance or number of changes needed to convert one data point into another.

Feature Selection: In feature selection, where you want to identify the most relevant features, Manhattan distance can be used as a criterion for selecting the most important features, as it tends to be less influenced by outliers than Euclidean distance.

Grid-Based Search Algorithms: When performing grid-based search or optimization algorithms, Manhattan distance can be useful for defining neighborhoods or regions to explore in the parameter space.

Use Euclidean Distance When:

Data Lies in Euclidean Space: Euclidean distance is appropriate when your data naturally lies in Euclidean space, such as in most real-world physical measurements like height, weight, and temperature.

Continuous Numeric Data: Euclidean distance is well-suited for continuous numeric data where the magnitude and relationships between values are meaningful.

Features Are Scaled Equally: If your features are on similar scales and there is no significant reason to treat any dimension differently, Euclidean distance is a reasonable choice.

Clustering or Nearest Neighbor Algorithms: Euclidean distance is commonly used in clustering algorithms like K-means and in nearest neighbor algorithms for classification and regression.

Principal Component Analysis (PCA): Euclidean distance is often used in dimensionality reduction techniques like PCA, where it assumes that features are normally distributed and have similar variances.

In many cases, it's a good practice to try both distance metrics and evaluate their impact on your specific machine learning task through cross-validation or other performance metrics. The choice between Manhattan and Euclidean distance should be driven by the characteristics of your data and the goals of your analysis.

Note: Euclidean Distance becomes more complex in higher dimensional spaces. Manhattan distance is easy to use.