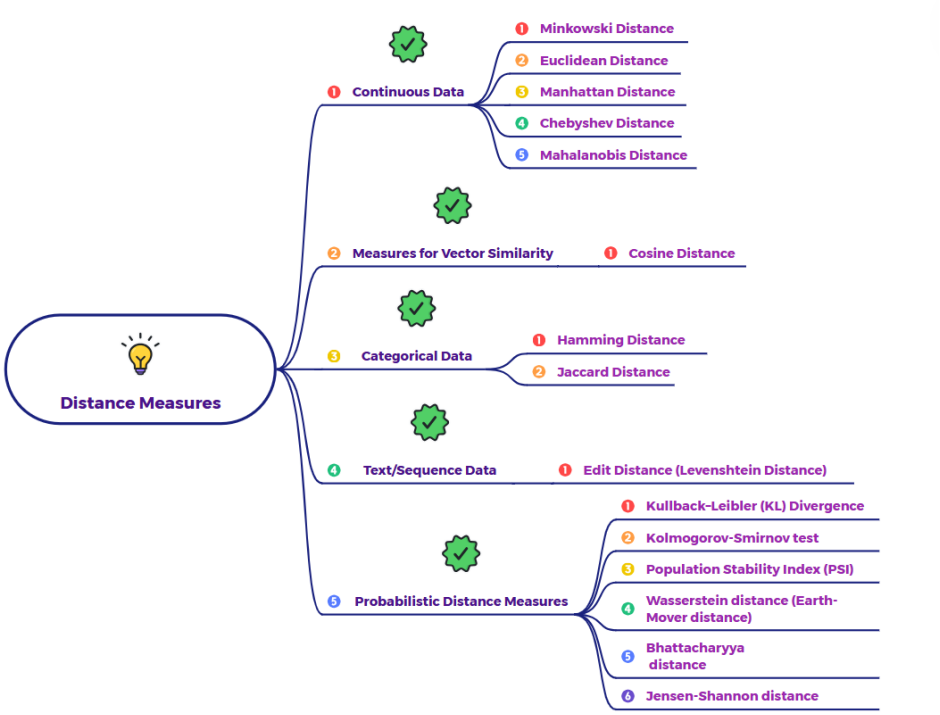
What are Distance Measures?

A **distance measure** is a function that quantifies the dissimilarity or distance between two objects. In machine learning, these "objects" are often data points represented as vectors. The smaller the distance, the more similar the objects are considered to be.



Key Points and Types

Distance measures are fundamental to many machine learning algorithms, including:

- **Clustering** (e.g., K-Means)
- **Classification** (e.g., K-Nearest Neighbors)
- **Dimensionality Reduction** (e.g., t-SNE)

There are many different types of distance measures, and the choice of which one to use depends heavily on the type of data and the specific problem you are trying to solve. The image you provided illustrates this well by categorizing them.

1. Continuous Data

These measures are used for numerical data.

- **Euclidean Distance**: The most common distance measure, representing the shortest straight-line distance between two points. It's often called the L2 norm.
- **Manhattan Distance**: Also known as the L1 norm or "taxicab" distance, it's the sum of the absolute differences of their Cartesian coordinates.
- **Minkowski Distance**: A generalization of both Euclidean and Manhattan distances.

2. Vector Similarity

These measures focus on the orientation of vectors rather than their magnitude.

- **Cosine Distance**: Measures the cosine of the angle between two vectors. It's often used for text analysis to determine how similar two documents are, regardless of their length.

3. Categorical Data

These are for non-numerical, or categorical, data.

- **Hamming Distance**: Counts the number of positions at which the corresponding symbols are different. Used for comparing two strings of equal length.
- **Jaccard Distance**: Measures dissimilarity between two sets, defined as one minus the Jaccard similarity coefficient, which is the size of the intersection divided by the size of the union of the sets.

4. Text/Sequence Data

- **Edit Distance**: The minimum number of single-character edits (insertions, deletions, or substitutions) required to change one word into the other.

5. Probabilistic Distance Measures

These are used to compare probability distributions.

- **KL Divergence**: An asymmetric measure of how one probability distribution differs from a second.
- **Wasserstein Distance**: Also known as the Earth Mover's Distance, it measures the minimum "cost" to transform one distribution into another. It is a true metric.
- **Jensen-Shannon Divergence**: A symmetric and smoothed version of KL divergence.

The distinction between these types is crucial, as using an inappropriate distance measure can lead to poor performance in a machine learning model.