DISTANCE METRIC

A **distance metric** is a way of measuring the "distance" or similarity between two data points in a dataset. In machine learning, distance metrics are essential because they allow algorithms (like K-Nearest Neighbors) to find points that are "closest" together in feature space.

Common Distance Metrics

1. **Euclidean Distance**:
    - Think of it as the "straight-line" distance between two points.
    - It's calculated as the square root of the sum of squared differences between each feature.

Formula for points $A = (x_1, y_1)$ and $B = (x_2, y_2)$:

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

2. **Manhattan Distance (Taxicab or City Block Distance)**:
    - Measures the distance by summing the absolute differences along each feature dimension.
    - Imagine navigating a city where you can only move in grid-like paths.

Formula:

$$d(A, B) = |x_2 - x_1| + |y_2 - y_1|$$

3. **Minkowski Distance**:
    - Generalized form that includes both Euclidean and Manhattan distances.
    - With a parameter p, if p=2 it's Euclidean distance, and if p=1 it's Manhattan distance.

Formula:

$$d(A, B) = \left(\sum_{i=1}^{n} |x_i - y_i|^p \right)^{\frac{1}{p}}$$

4. **Cosine Similarity**:
    - Measures the angle between two vectors, focusing on direction rather than magnitude.
    - It's especially useful in high-dimensional spaces, like text analysis.

Formula:

$$\text{Cosine Similarity} = \frac{A \cdot B}{\|A\| \times \|B\|}$$

5. **Hamming Distance**:
    - Counts the number of positions at which two binary strings differ.
    - Common in text comparison and binary feature spaces.

Formula: For binary vectors A and B,

$$d(A, B) = \text{count of differing bits}$$

Choosing the Right Distance Metric

The right metric depends on the data type and the problem:

- **Euclidean** is often used for continuous features.
- **Manhattan** suits grid-like structures or sparse data.
- **Cosine similarity** is great for high-dimensional, directional data like text.
- **Hamming** works well with categorical or binary data.