Distance Metrics in K-Nearest Neighbors (KNN)

1 Introduction

In K-Nearest Neighbors (KNN) and other machine learning algorithms, distance metrics play a crucial role in determining the similarity or dissimilarity between data points. This document provides an overview of various distance metrics including Euclidean, Manhattan, Minkowski, Hamming, Cosine, and Jaccard distances.

2 Euclidean Distance

The Euclidean distance is the straight-line distance between two points in Euclidean space. For two points $\mathbf{x}_1 = (x_{11}, x_{12}, \ldots, x_{1d})$ and $\mathbf{x}_2 = (x_{21}, x_{22}, \ldots, x_{2d})$ in $\mathbb{R}^d$, the Euclidean distance is computed as:

$$\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{d} (x_{1i} - x_{2i})^2}$$

Example:
For points $\mathbf{x} = (3, 4)$ and $\mathbf{y} = (4, 3)$:

$$\text{Euclidean Distance} = \sqrt{(3-4)^2 + (4-3)^2} = \sqrt{(-1)^2 + (1)^2} = \sqrt{1+1} = \sqrt{2} \approx 1.414$$

3 Manhattan Distance

The Manhattan distance, also known as the City Block Distance or L1 norm, is the sum of the absolute differences of their coordinates. For two points $\mathbf{A} = (a_1, a_2, \ldots, a_d)$ and $\mathbf{B} = (b_1, b_2, \ldots, b_d)$:

$$\text{Manhattan Distance} = \sum_{i=1}^{d} |a_i - b_i|$$

Example:
For blocks $\text{block1} = (1, 2, 3, 4)$ and $\text{block2} = (5, 6, 7, 8)$:

$$\text{Manhattan Distance} = |1-5| + |2-6| + |3-7| + |4-8| = 4+4+4+4 = 16$$

Manhattan Lengths:
For blocks $\text{block1} = (5, 2, -3, 4)$ and $\text{block2} = (1, 6, -7, 8)$:

$$\text{Manhattan Length of block1} = |5| + |2| + |-3| + |4| = 5+2+3+4 = 14$$

$$\text{Manhattan Length of block2} = |1| + |6| + |-7| + |8| = 1+6+7+8 = 22$$

4 Hamming Distance

The Hamming distance measures the number of positions at which the corresponding symbols differ. For two binary strings $\mathbf{u}$ and $\mathbf{v}$:

$$\text{Hamming Distance} = \sum_{i=1}^{k} |u_i - v_i|$$

Example:
For $\mathbf{x}_1 = (0,0,0,1,0,1,1,0,1,1,1,0,0,0,1)$ and $\mathbf{x}_2 = (0,1,0,1,0,1,0,1,0,1,0,1,0,1,0)$:

$$\text{Hamming Distance} = \sum_{i=1}^{15} |x_{1i} - x_{2i}| = 4$$

5 Cosine Similarity and Distance

Cosine similarity measures the cosine of the angle between two non-zero vectors $\mathbf{a}$ and $\mathbf{b}$. The cosine distance is 1 minus the cosine similarity.

$$\text{Cosine Similarity} = \frac{\mathbf{a} \cdot \mathbf{b}}{\|\mathbf{a}\|\|\mathbf{b}\|}$$

Example:
For vectors $\mathbf{a} = (2, 3)$ and $\mathbf{b} = (2, 2)$:

$$\text{Cosine Similarity} = \frac{(2 \times 2 + 3 \times 2)}{\sqrt{2^2+3^2} \times \sqrt{2^2+2^2}} = \frac{10}{\sqrt{13} \times \sqrt{8}} \approx 0.89$$

$$\text{Cosine Distance} = 1 - 0.89 = 0.11$$

6 Jaccard Distance

The Jaccard distance is defined as 1 minus the Jaccard index. It measures dissimilarity between two sets.

Jaccard Index:

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|}$$

Jaccard Distance:

$$\text{Jaccard Distance} = 1 - J(X, Y)$$

Example:
For sets $X = \{a, b, c\}$ and $Y = \{b, c, d\}$:

$$|X \cap Y| = 2$$

$$|X \cup Y| = 4$$

$$J(X, Y) = \frac{|X \cap Y|}{|X \cup Y|} = \frac{2}{4} = 0.5$$

$$\text{Jaccard Distance} = 1 - 0.5 = 0.5$$

7 Distance Example Calculation

Given:
- $\mathbf{a} = (a, b, c, d, f)$ - $\mathbf{b} = (a, g, d, d, f)$

Euclidean Distance Calculation:
For numerical vectors, replace with actual values and apply the Euclidean distance formula.

Manhattan Distance Calculation:
For numerical vectors, replace with actual values and apply the Manhattan distance formula.