Euclidean Distance

Euclidean distance is one of the most commonly used distance metrics in KNN. It measures the straight-line distance between two points in Euclidean space. For two points $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$, the Euclidean distance $d(\mathbf{x}, \mathbf{y})$ is calculated as:

$$d(\mathbf{x}, \mathbf{y}) = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

Application in KNN

In the context of KNN:

1. **Distance Calculation**: For a given input instance $\mathbf{x}$, you calculate the Euclidean distance between $\mathbf{x}$ and every instance in the training dataset.
2. **Finding Nearest Neighbors**: You then sort these distances and identify the k smallest ones. The corresponding instances in the training dataset are the nearest neighbors.
3. **Prediction**: For classification, the predicted class is usually determined by majority voting among the k nearest neighbors. For regression, it could be the average of the values of the nearest neighbors.

Example

Suppose you have the following 2D feature space with points and you want to classify a new point (x, y) with $k = 3$:

Point	x_1	x_2	Label
A	1.0	1.0	Class1
B	2.0	2.0	Class2
C	3.0	3.0	Class1
D	6.0	6.0	Class2

New point to classify: $\mathbf{p} = (2.5, 2.5)$

Calculate the Euclidean distance between $\mathbf{p}$ and each point:

- $d(\mathbf{p}, A) = \sqrt{(2.5 - 1.0)^2 + (2.5 - 1.0)^2} = \sqrt{2.25 + 2.25} = \sqrt{4.5} \approx 2.12$
- $d(\mathbf{p}, B) = \sqrt{(2.5 - 2.0)^2 + (2.5 - 2.0)^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5} \approx 0.71$
- $d(\mathbf{p}, C) = \sqrt{(2.5 - 3.0)^2 + (2.5 - 3.0)^2} = \sqrt{0.25 + 0.25} = \sqrt{0.5} \approx 0.71$
- $d(\mathbf{p}, D) = \sqrt{(2.5 - 6.0)^2 + (2.5 - 6.0)^2} = \sqrt{12.25 + 12.25} = \sqrt{24.5} \approx 4.95$

The three closest points are B, C, and A with distances 0.71, 0.71, and 2.12 respectively. If $k = 3$:

- B (Class2)

- C (Class1)
- A (Class1)

Majority voting would classify the new point $\mathbf{p}$ as Class1.

So, in summary, the Euclidean distance metric is a core part of the KNN algorithm when it comes to determining the "closeness" of instances, and it forms the basis for selecting the nearest neighbors.