Machine Learning and Intelligent Systems

K-Nearest Neighbors

Maria A. Zuluaga

October 6 2023

EURECOM - Data Science Department

Table of contents

Definition

Intuition

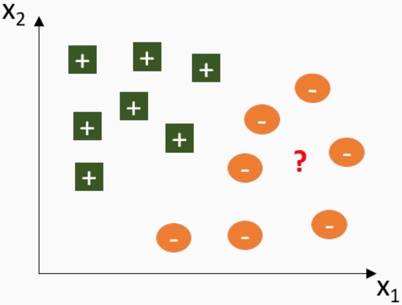
Nearest neighbors is among the simplest, oldest (1967) yet effective learning methods.

Let's try to understand how it works through an exercise.

Intuition

Nearest neighbors is among the simplest, oldest (1967) yet effective learning methods.

Let's try to understand how it works through an exercise.

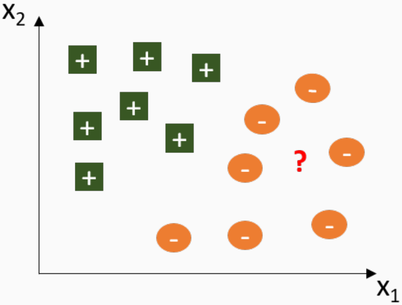


- How would you classify the test point (?) ?
- What drove you to make that choice? What is the assumption behind your choice?

Intuition

Nearest neighbors is among the simplest, oldest (1967) yet effective learning methods.

Let's try to understand how it works through an exercise.



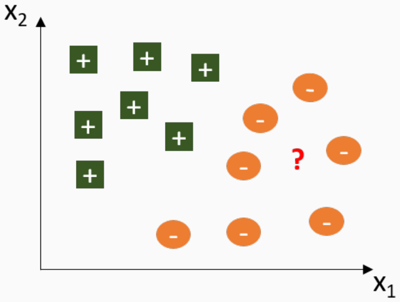
- How would you classify the test point (?) ?
- What drove you to make that choice? What is the assumption behind your choice?

Assumption: Similar inputs have similar outputs

k Nearest Neighbors: Formalization

Data Assumptions:

Similar $\mathbf{x}$ have similar y



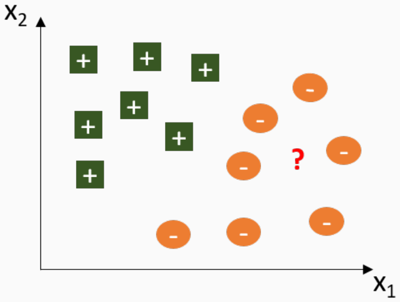
k Nearest Neighbors: Formalization

Data Assumptions:

Similar $\mathbf{x}$ have similar y

$$y \in \mathbb{R} \text{ - regression}$$

$$y \in \mathcal{C} \text{ - classification}$$



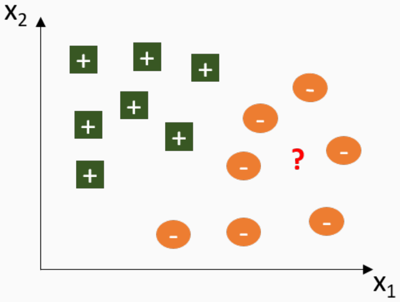
k Nearest Neighbors: Formalization

Data Assumptions:

Similar $\mathbf{x}$ have similar y

$$y \in \mathbb{R} \text{ - regression}$$

$$y \in \mathcal{C} \text{ - classification}$$



Definition of the k nearest neighbors:

Given a test point $\mathbf{x}^*$, let us denote $\mathcal{S}_{\mathbf{x}^*}$ as its set of k nearest neighbors. Formally:

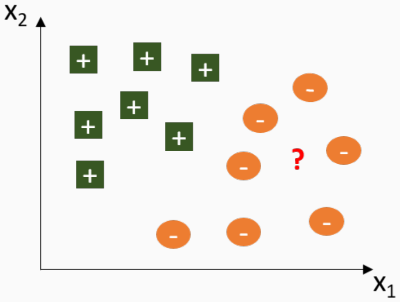
k Nearest Neighbors: Formalization

Data Assumptions:

Similar $\mathbf{x}$ have similar y

$$y \in \mathbb{R} \text{ - regression}$$

$$y \in \mathcal{C} \text{ - classification}$$



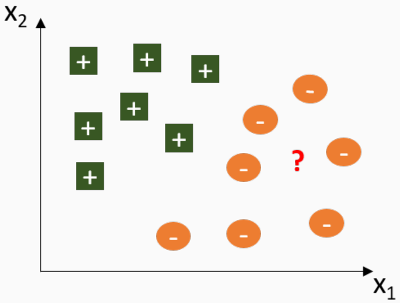
Definition of the k nearest neighbors:

Given a test point $\mathbf{x}^*$, let us denote $\mathcal{S}_{\mathbf{x}^*}$ as its set of k nearest neighbors. Formally:

$$\mathcal{S}_{\mathbf{x}^*} \subseteq \mathcal{D} \text{ .s.t. } |\mathcal{S}_{\mathbf{x}^*}| = k$$

$$\forall (\mathbf{x}', y') \in \mathcal{D} \setminus \mathcal{S}_{\mathbf{x}^*} \ \mathrm{dist}(\mathbf{x}^*, \mathbf{x}') \geq \max_{(\mathbf{x}'', y'') \in \mathcal{S}_{\mathbf{x}^*}} \mathrm{dist}(\mathbf{x}^*, \mathbf{x}'')$$

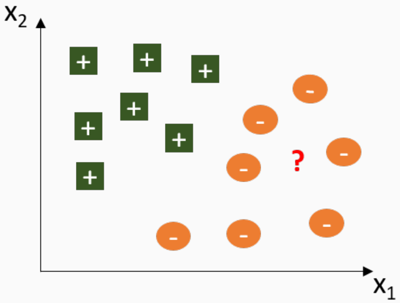
k Nearest Neighbors: Hypothesis



k Nearest Neighbors: Hypothesis

Regression: The output is the average of the values of k nearest neighbors

$$\hat{y} = h(\mathbf{x}^*) = \frac{1}{k} \sum_{(\mathbf{x}_i'', y_i'') \in \mathcal{S}_{\mathbf{x}^*}} y_i'' \tag{1}$$



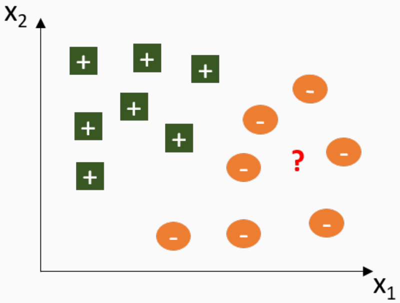
k Nearest Neighbors: Hypothesis

Regression: The output is the average of the values of k nearest neighbors

$$\hat{y} = h(\mathbf{x}^*) = \frac{1}{k} \sum_{(\mathbf{x}_i'', y_i'') \in \mathcal{S}_{\mathbf{x}^*}} y_i'' \qquad (1)$$

Classification: An unseen point $\mathbf{x}^*$ is classified by a majority vote of its k nearest neighbors:

$$\hat{y} = h(\mathbf{x}^*) = \text{mode}(\{y_i'' : (\mathbf{x}_i'', y_i'') \in \mathcal{S}_{\mathbf{x}^*}\}) \qquad (2)$$



Distance Metric

The definition of the set $\mathcal{S}_{\mathbf{x}^*}$ is strongly dependent on the distance metric (dist) that is used. The better the chosen metric reflects the similarity among points the better the resulting model will be.

Distance Metric

The definition of the set $\mathcal{S}_{\mathbf{x}^*}$ is strongly dependent on the distance metric (dist) that is used. The better the chosen metric reflects the similarity among points the better the resulting model will be.

One of the most common metrics used is the Minkowski distance:

$$\text{dist}(\mathbf{x}, \mathbf{z}) = \left(\sum_{j=1}^{D} |x_j - z_j|^p \right)^{1/p} \tag{3}$$

Distance Metric

The definition of the set $\mathcal{S}_{\mathbf{x}^*}$ is strongly dependent on the distance metric (dist) that is used. The better the chosen metric reflects the similarity among points the better the resulting model will be.

One of the most common metrics used is the Minkowski distance:

$$\text{dist}(\mathbf{x}, \mathbf{z}) = \left(\sum_{j=1}^{D} |x_j - z_j|^p \right)^{1/p} \tag{3}$$

The main advantage of this metric is its generality. It contain many well-known distances as special cases.

Question: Can you identify what case is $p = 2$?

The k-NN algorithm differs from the ”general” setup of supervised learning that we have seen so far.

Instance-Based Learning

The k-NN algorithm differs from the ”general” setup of supervised learning that we have seen so far.

It does not attempt to construct a general internal model, but simply stores instances of the training data.

Instance-Based Learning

The k-NN algorithm differs from the "general" setup of supervised learning that we have seen so far.

It does not attempt to construct a general internal model, but simply stores instances of the training data.

This type of approach is referred as **instance-based learning** or **non-generalizing learning**.

Question: Can you spot the differences?

Placement Exam

Wrap-Up

Summary: Notation

Symbol	Reads as
X	Input variable ($\mathbb{R}^D$)
$\mathbf{x}_i$	i^{th} feature vector. Observed value of X.
$\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N)^T$	Matrix of N input $D-$dimensional vectors $\mathbf{x}_i$
x_j	j^{th} element of the i^{th} input vector $\mathbf{x}_i$, i.e. x_i^j
Y	Output variable ($\mathcal{C}$)
y_i	i^{th} output label
$\mathbf{y} = (y_1, \ldots, y_N)^T$	Observed vector of outputs y_i
$\mathbf{x}^*$	Test point (unseen data)
$\hat{y}$	Prediction for $\mathbf{x}^*$

Table 1: Different notation for the input and output variables

Further Reading and Useful Material

Source	Notes
The Elements of Statistical Learning	Ch. 2