Lecture 5: K-nearest neighbors

Introduction to Machine Learning

Sophie Robert

L3 MIASHS — Semestre 2

2022-2023

Reminders on previous session

Question

Can anyone remind me of the definition of supervised learning ? Can anyone give me some kind of problems that can be solved with supervised learning ?

Main idea

K-nearest neighbors algorithm

The k-nearest neighbors algorithm is a **non-parametric supervised learning** method, which assigns to an incoming record the label issued from the plurality of votes of its k nearest neighbors.

Main idea

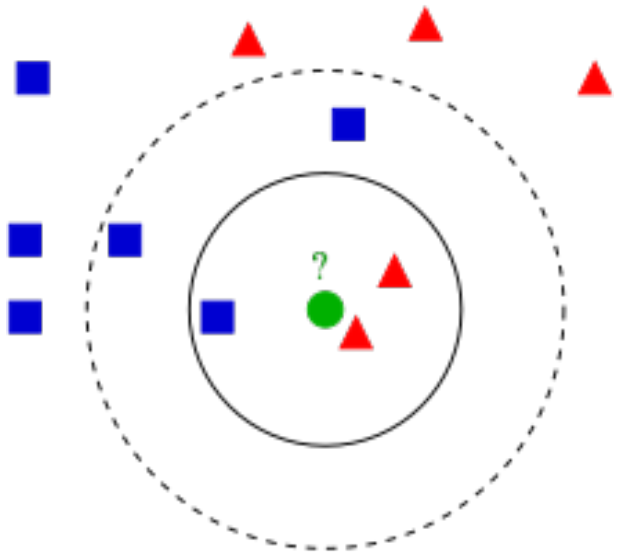
K-nearest neighbors algorithm

The k-nearest neighbors algorithm is a **non-parametric supervised learning** method, which assigns to an incoming record the label issued from the plurality of votes of its k nearest neighbors.

With an incoming data record:

- Find the $k \in \mathbb{N}$ nearest neighbors
- Assign the classification label of the most frequent labels among neighbors

Example



Can you identify a problem with certain values of k?

Example: Dog breed prediction

Training dataset:

Height	Weight	Label
45	30	Labradoodle
30	25	Labradoodle
40	35	Labradoodle
20	15	English cocker
22	18	English cocker
25	20	English cocker

Individual to classify using 1 NN and 3 NN (euclidean and manhattan distance)

Height	Weight	Label
25	31	?

Example: solution

Compute distance between dataset and individual to classify:

Distance	Label
20.02	Labradoodle
7.81	Labradoodle
15.52	Labradoodle
16.76	English cocker
13.34	English cocker
11.0	English cocker

Using 1NN: *Labradoodle*
Using 3NN: *English cocker*

Hyperparameters

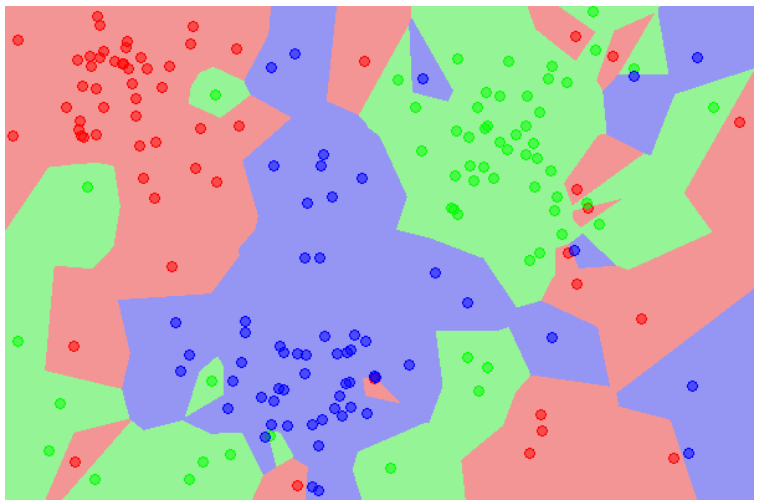
Hyperparameters

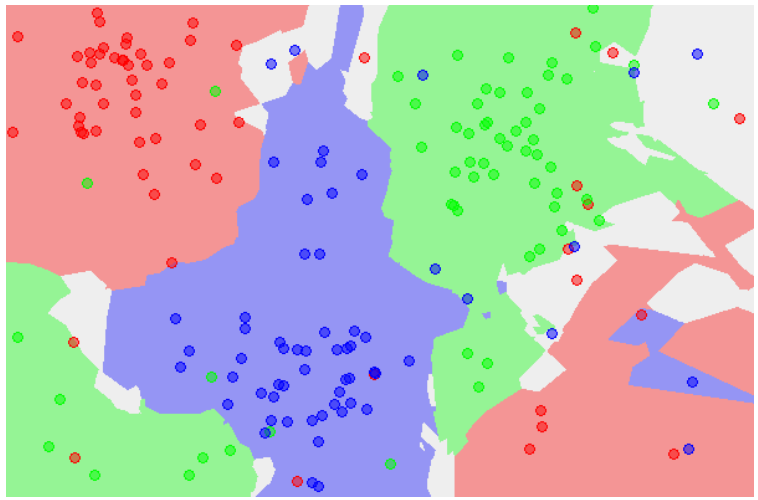
What **hyperparameters*** does the k-nearest neighbor algorithm require ?

Hyperparameters

Hyperparameters

What **hyperparameters*** does the k-nearest neighbor algorithm require ?





Hyperparameter selection

To select the optimum hyperparameters (distance to use, best number of neighbors), use **k-fold validation** and select the combination with the highest score (in its simplest version using a factorial design).

Advantages and limits

Advantages:

Advantages and limits

Advantages:

- Very easy to extend to multi-class classification
- Very easy to understand
- Non-parametric algorithm (no assumption regarding data distribution)
- No previous training

Advantages and limits

Advantages:

- Very easy to extend to multi-class classification
- Very easy to understand
- Non-parametric algorithm (no assumption regarding data distribution)
- No previous training

Limits:

Advantages and limits

Advantages:

- Very easy to extend to multi-class classification
- Very easy to understand
- Non-parametric algorithm (no assumption regarding data distribution)
- No previous training

Limits:

- Very sensitive to its hyperparametrization
- Very sensitive to noise (features with little to no impact on the dataset)
- Expensive to compute
- Difficult to interprete

Questions

Questions ?