K-Nearest Neighbor

Data Mining

Lê Công Minh Khôi – 519H0181

Introduction

Simple algorithm

Classification/Regression

Based on a similarity measure

1970s

Statistical estimation

Pattern recognition

1990s

Popular in machine learning

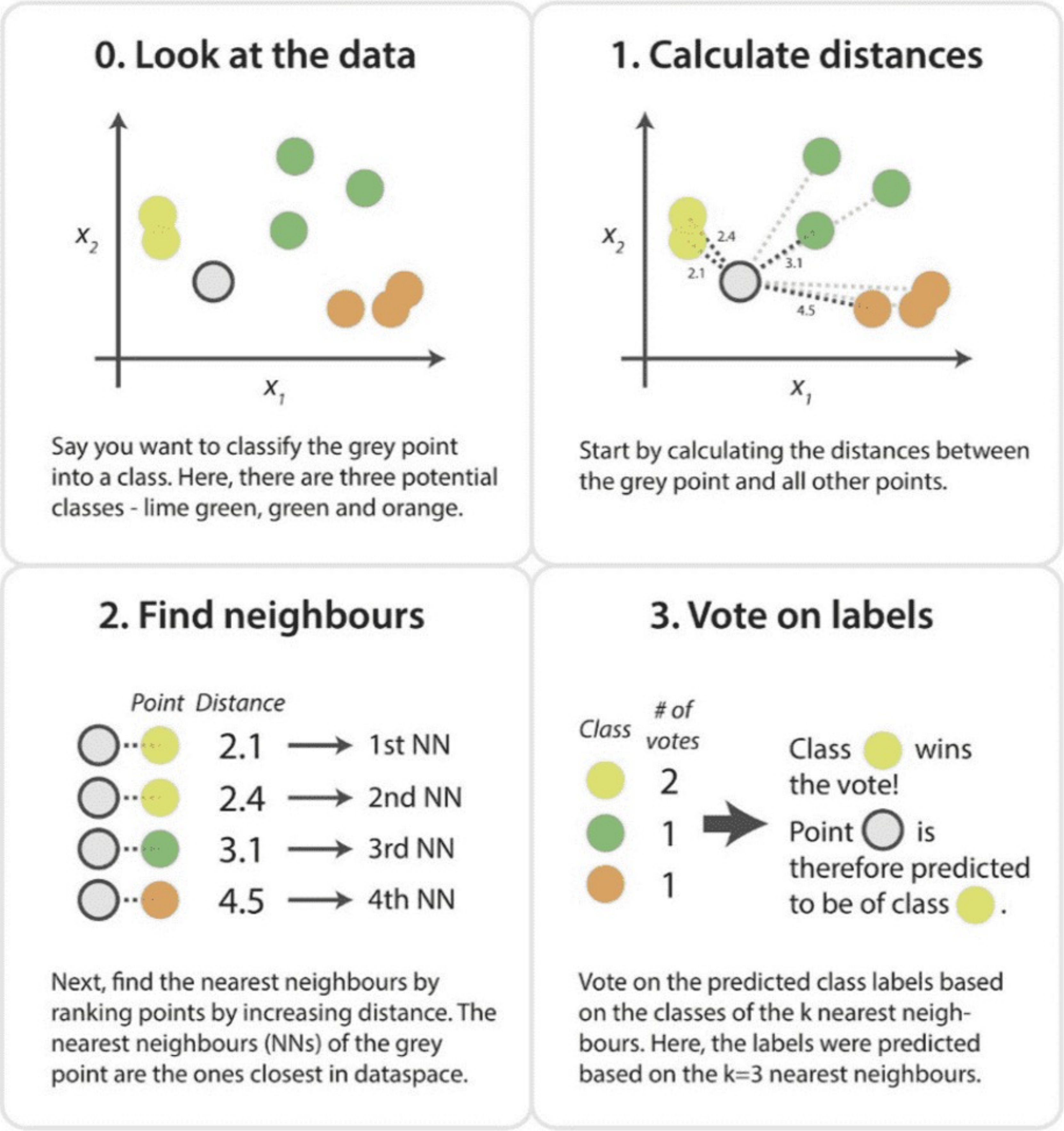
Why do we need a kNN algorithm?

kNN

Simple To Implement

Versatile

Effective



How does the kNN algorithm work?

The kNN algorithm assumes that similar things exist in close proximity. In other words, similar things are near to each other.

Step by step explanation

1. Load the data
2. Choose k
3. For each point in the data
    1. Calculate the distance between current point and the point we try to predict
    2. Add the distance and the index of the example to an ordered collection
4. Sort the ordered collection of distances and indices from smallest to largest (in ascending order) by the distances
5. Pick the first k entries from the sorted collection
6. Get the labels of the selected k entries
7. Return prediction
    1. If regression, return the mean of the k labels
    2. If classification, return the mode of the k labels

How to choose the value of k?

Usually odd

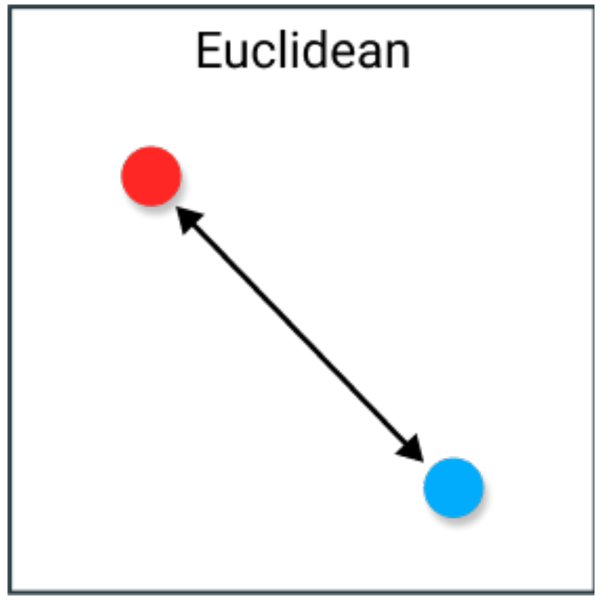
k
sensitive to noise

Try different k value

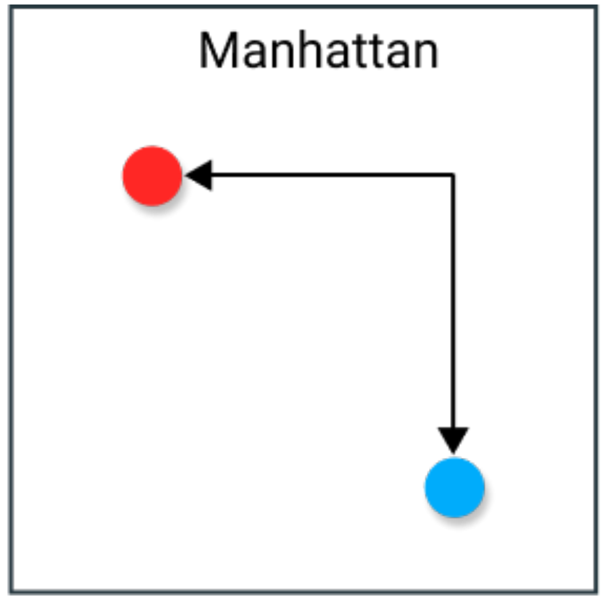
Pick the best result

Calculating the distance

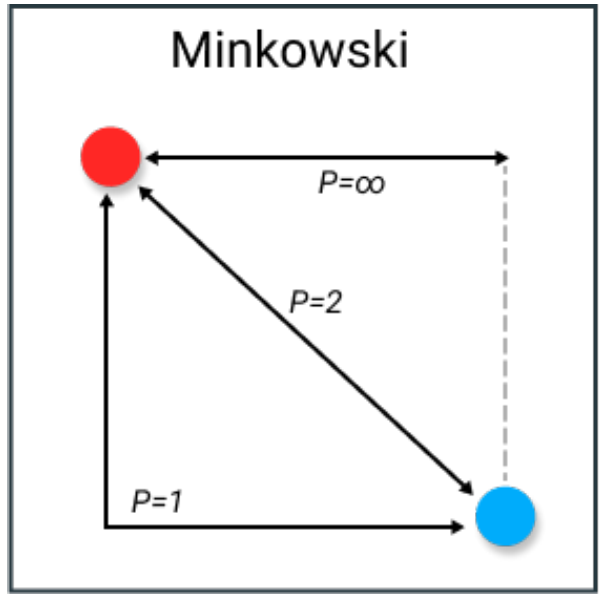
Euclidean



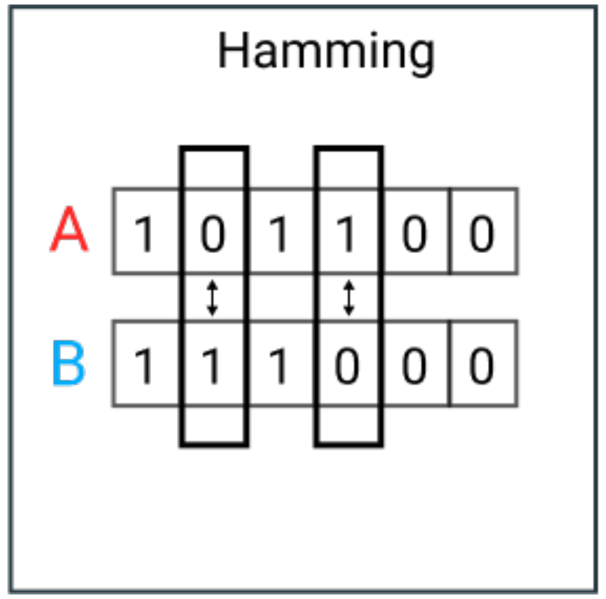
Manhattan



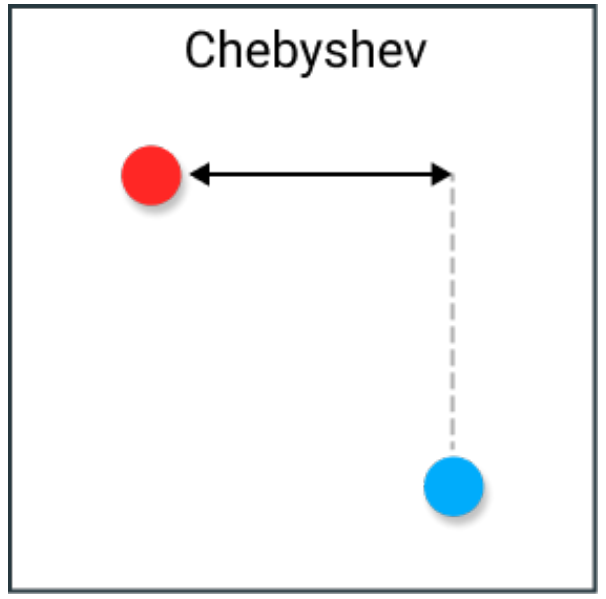
Minkowski



Hamming



Chebyshev



Advantages

Simple to implement

Versatile

Training is trivial

Works with any number of classes

Effective if the training data is large

Disadvantages

Computationally expensive

High memory requirement

Prediction stage might be slow with a large testing dataset

Does not work well with high dimensional data

One-hot encoding is required for categorical features

Thank you

Please keep in mind your questions