K-Nearest Neighbor (K-NN) algorithm

Prof. Junghyun Kim
Director, Engineering Systems Design Laboratory (ESDL)
Assistant Professor, School of Applied Artificial Intelligence
Handong Global University

1

Course objectives

- By the end of this module, you will be able to answer the following questions:
  - What is the K-NN algorithm?
  - How does the K-NN algorithm work?
  - What are the pros and cons of the K-NN algorithm?

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

2

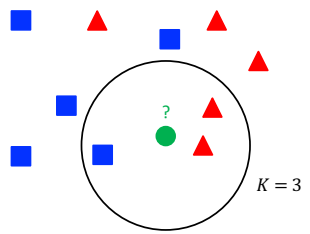
What is the K-NN algorithm?

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

3

What is the K-NN algorithm?

- The K-NN algorithm is a type of instance-based learning (i.e., lazy learning algorithm)
  - The algorithm assumes that similar things are near to each other
  - A class label is assigned based on a majority vote



Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

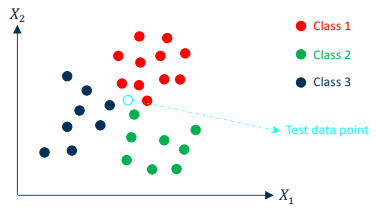
4

How does the K-NN algorithm work?

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

How does the K-NN algorithm work?

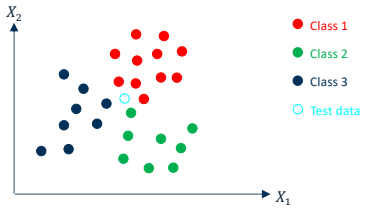
- Suppose that we would like to classify the test data point into one of the classes
  - In other words, should we classify the test data point into either Class 1 or Class 2, or Class 3?



Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

How does the K-NN algorithm work?

- STEP 1. Specify the number of K
  - For example, if $K = 5$, the algorithm will be looking for five points close to the test data point

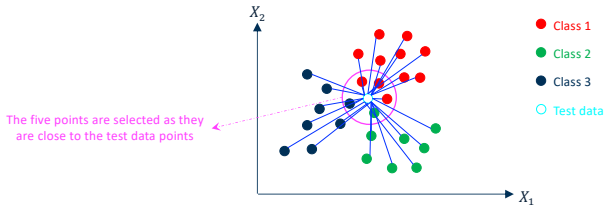


How to identify five points close to the test data point?

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

How does the K-NN algorithm work?

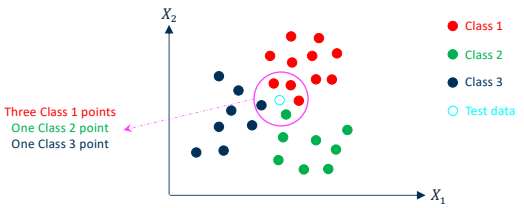
- STEP 2. Calculate the distance between the test data point and the other points
  - The Euclidean distance is typically used; however, it is important to note that there are different types of distance metrics such as the Manhattan distance



Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

How does the K-NN algorithm work?

- STEP 3. Determine a class membership based on the selected points
  - In other words, among the K neighbor points, count the number of the data points in each class

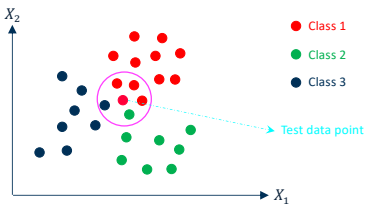


  - Note that we may select the nearest point if there is an identical score among class membership

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

How does the K-NN algorithm work?

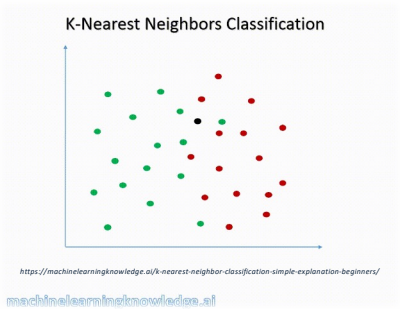
- STEP 4. Assign the test data point to the class where the number of points is maximum
  - In this case, the test data point is likely to assign to the Class 1



Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

Demo

- Animated explanation of the KNN-based classification



Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

What are the pros and cons of the K-NN algorithm?

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

Pros and cons of the K-NN algorithm

13

- Advantages
  - It is easy to implement
  - It is a robust algorithm, especially if it deals with large volumes of training data
  - It does not require generalizing a model (i.e., instance-based learning)
  - It only requires one hyperparameter, which is a K value
  - …
- Disadvantages
  - It is challenging to choose the best *K* parameter in a proper manner
  - It does not perform well with high-dimensional data inputs (i.e., **the curse of dimensionality**)
  - It is prone to overfitting
  - It is computationally expensive, especially if the number of data points increases
  - …

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

Pros and cons of the K-NN algorithm

14

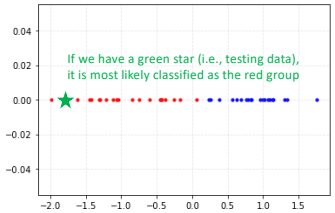
- Discussion
  - Why is the K-NN algorithm vulnerable to the curse of dimensionality?

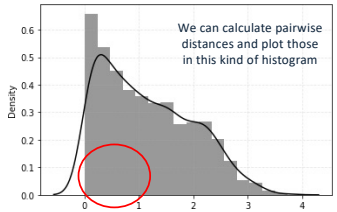
Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

Pros and cons of the K-NN algorithm

15

- The curse of dimensionality
  - The K-NN classification algorithm is typically vulnerable to the curse of dimensionality (i.e., the K-NN classification algorithm is especially sensitive to additional dimensions)
  - Imagine that we are analyzing the following 1-D dataset:





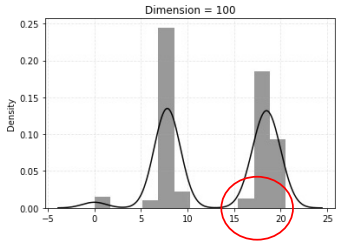
It is important to note that a lot of distances are very small (i.e., there is a high peak around $x = 0$)

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

Pros and cons of the K-NN algorithm

16

- The curse of dimensionality
  - What if we have 100 dimensions?
    - This is where the curse of dimensionality comes in



Here, the idea of the K-NN algorithm starts to break down:

- ❑ Recall that distance information (e.g., which data point is close to me) helps determine class membership for a testing data point
- ❑ As dimensions increase (e.g., 100), distance information become less meaningful
- ❑ Intuitively, they are not really neighbors in a high-dimensional space as every data point is very far from every other data point

It seems that most of the data points are fairly far apart

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

Course summary

- Throughout this module, you have learned:
  - What is the K-NN algorithm?
  - How does the K-NN algorithm work?
  - What are the pros and cons of the K-NN algorithm?

Copyright © by Prof. Junghyun Kim, School of Applied Artificial Intelligence, Handong Global University

THANK YOU

For more information, please reach out to Prof. Junghyun Kim at junghyun.kim@handong.edu