Machine Learning

k-Nearest Neighbors (KNN)

Dept. SW and Communication Engineering

Prof. Giseop Noh (kafa46@hongik.ac.kr)

Contents

Key Concepts of k-Nearest Neighbors

Idea of Nearest Neighbor

k-Nearest Neighbor (k-NN)

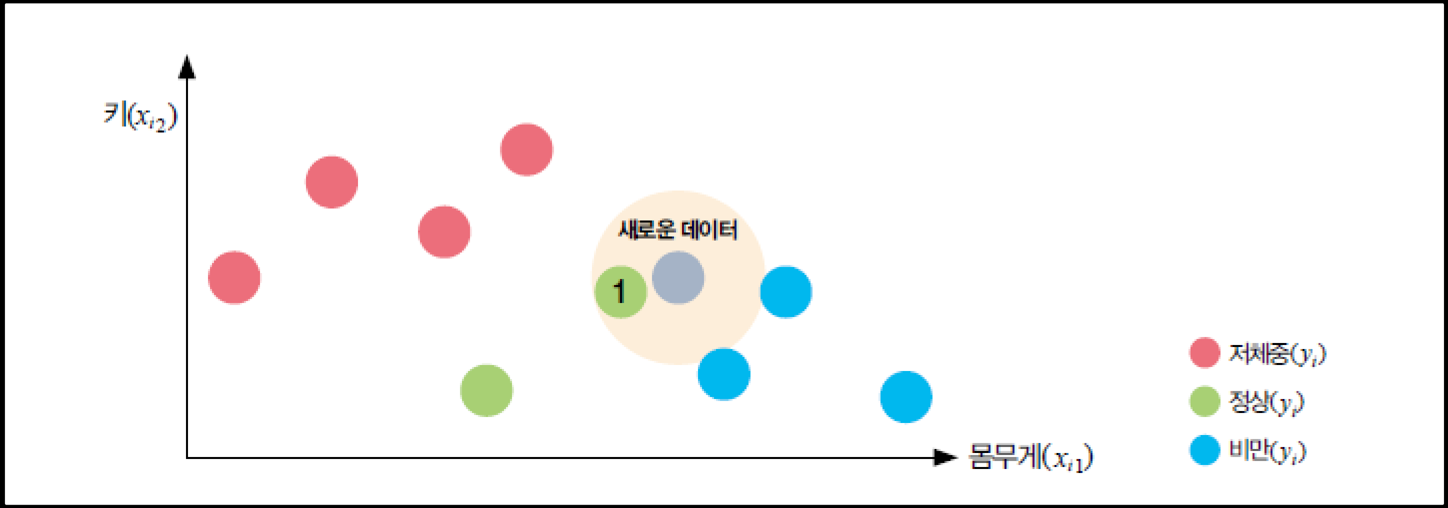
- A type of supervised learning algorithm
- Classifies new data by comparing it with existing data based on similarity

1-Nearest Neighbor (k = 1)

- What is the most similar data point to the new input?

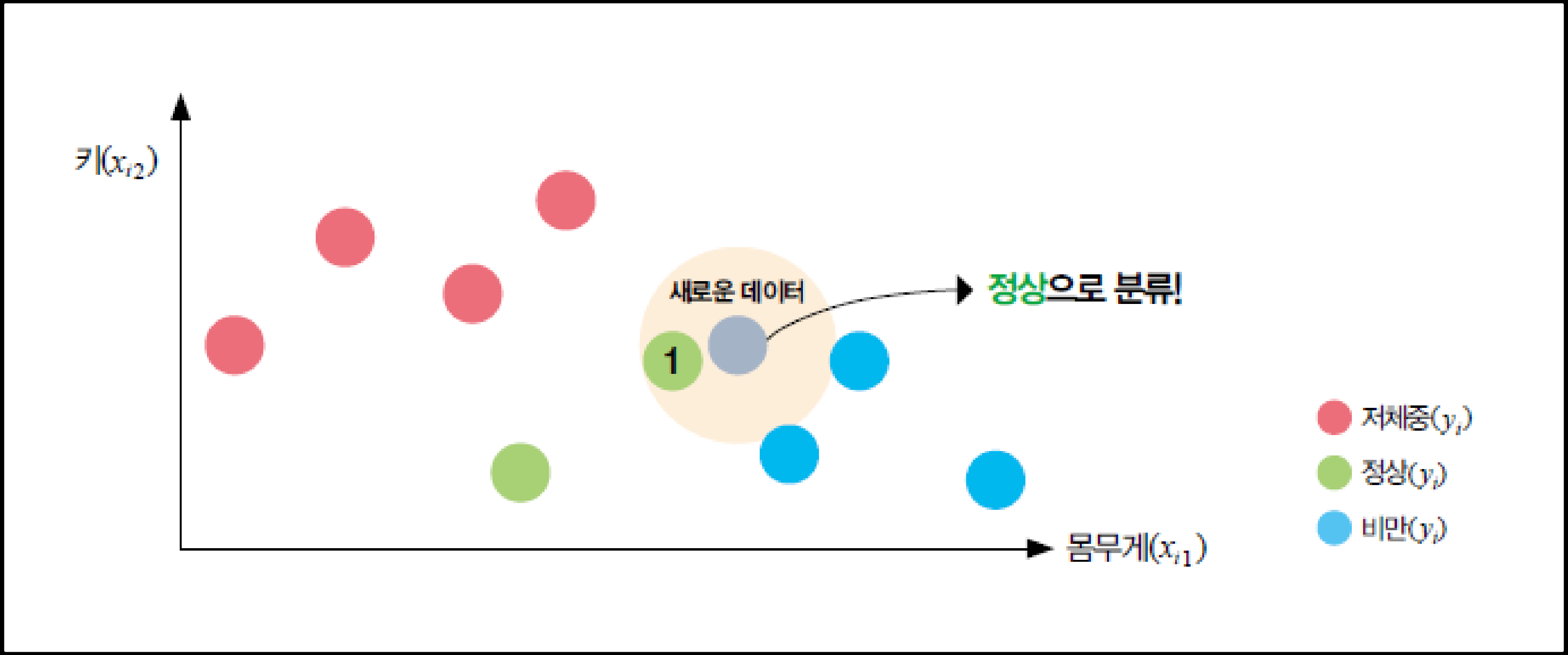
  → Classify it with the same label as the nearest data point

- The value of k indicates how many nearest neighbors to consider for classification



Key Concepts of k-Nearest Neighbors

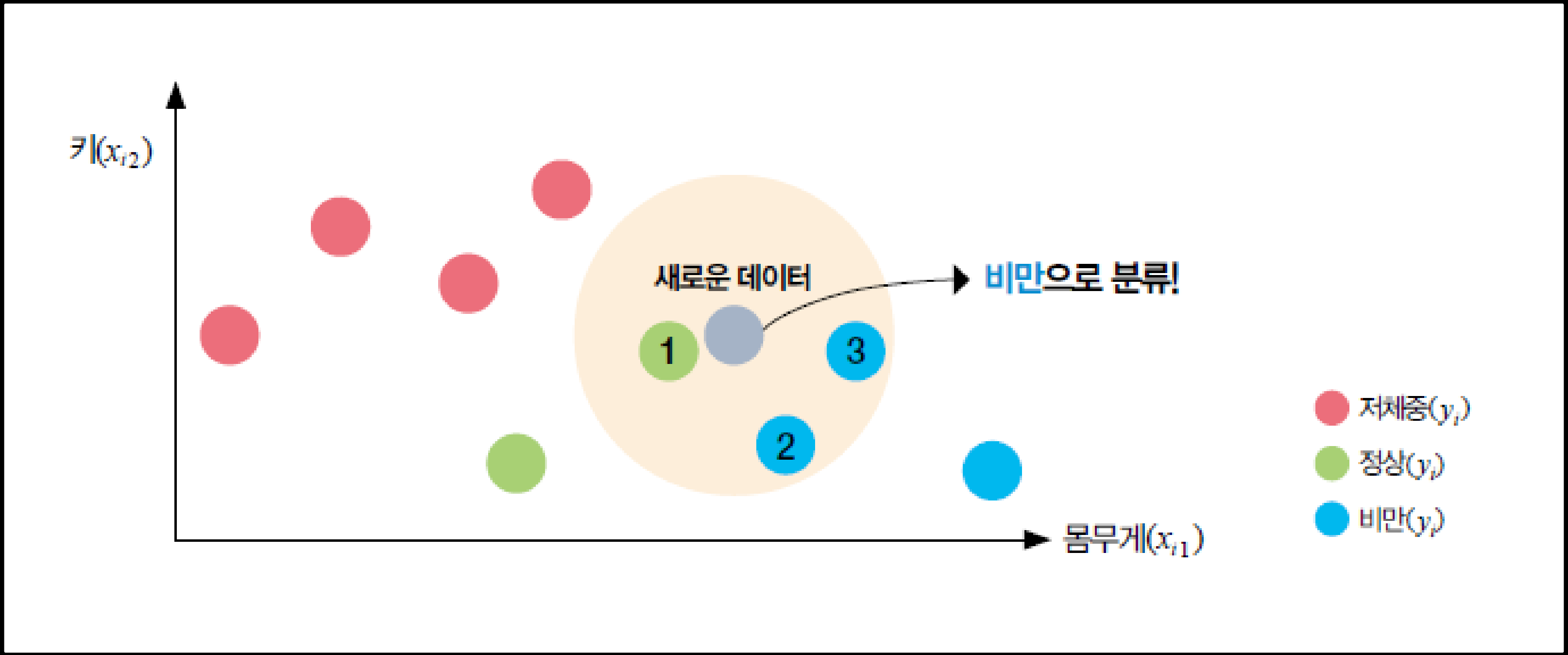
- **Predict which class the new data point will belong to depending on value of k**



3-Nearest Neighbors

3-Nearest Neighbors (k = 3)

- When $k = 3$, how do we determine the label of the new data point?
- The label is determined by majority voting among the labels of the 3 nearest neighbors



Distance Metrics

Distance Metrics

- How do we measure closeness between data points in a table?

  → **Distance metrics**

- In instance-based learning, prediction is made after a new data point arrives
- Explore **Euclidean** and Manhattan distances using the data

	몸무게(x_{i1})	키(x_{i2})	비만 여부(y_i)
생도 1	1	0	0
생도 2	2	1	0
생도 3	3	3	0
생도 4	5	2	1
생도 5	5	4	1
생도 6	6	5	1
생도 7	9	2	1
신입 생도	3	4	?

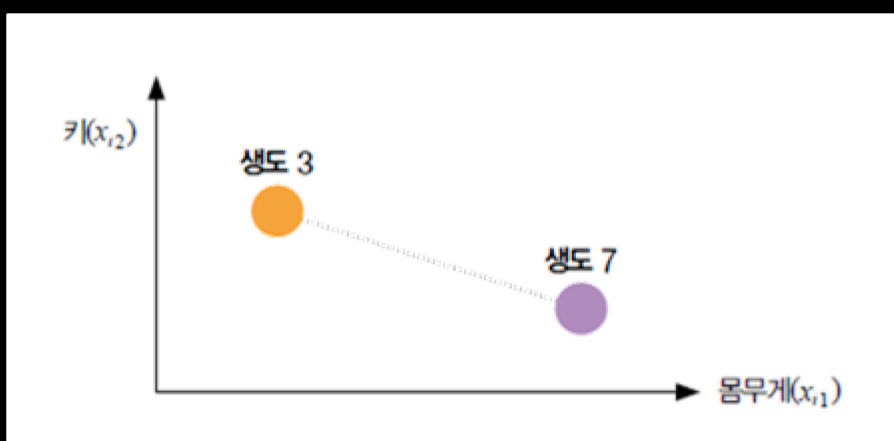
Euclidean Distance

- **The most commonly used distance**

  → The straight-line distance (distance between two points)

- **It is calculated as the square root of the sum of squared differences between feature values**

$$Euclidean\ Distance\ (X_1, X_2) = \sqrt{\sum_{j=1}^{p} (x_{1j} - x_{2j})^2}$$



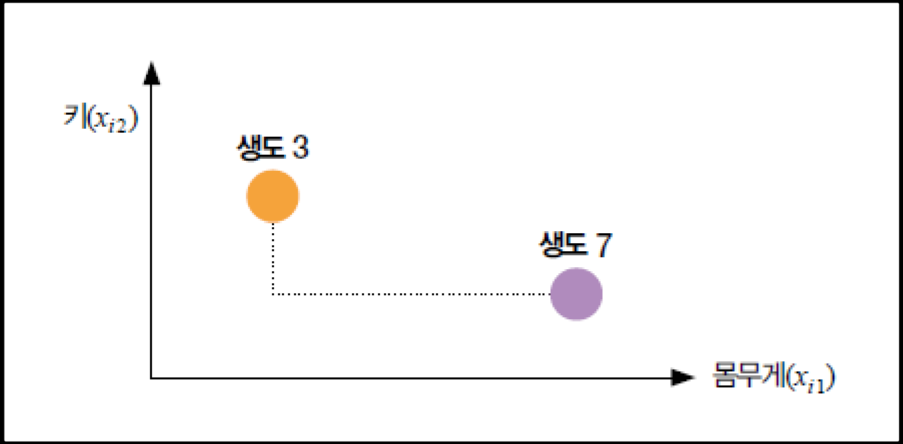
$$\text{Euclidean Distance}(\mathbf{x}_3, \mathbf{x}_7)$$

$$= \sqrt{(3-9)^2 + (3-2)^2} = \sqrt{37} \approx 6.08$$

Manhattan Distance

- **Not commonly used in daily life, but frequently applied in machine learning and programming**
- **Like navigating in a grid-like city (e.g., Manhattan in New York), movement follows the axis-aligned paths**

$$Manhattan\ Distance\ (X_1, X_2) = \sum_{j=1}^{p} |x_{1j} - x_{2j}|$$



$$\text{Manhattan Distacne}(\mathbf{x}_3, \mathbf{x}_7) = |3-9| + |3-2| = 7$$

Classification with k-Nearest Neighbors

Steps in k-NN Operation (1/2)

Goal

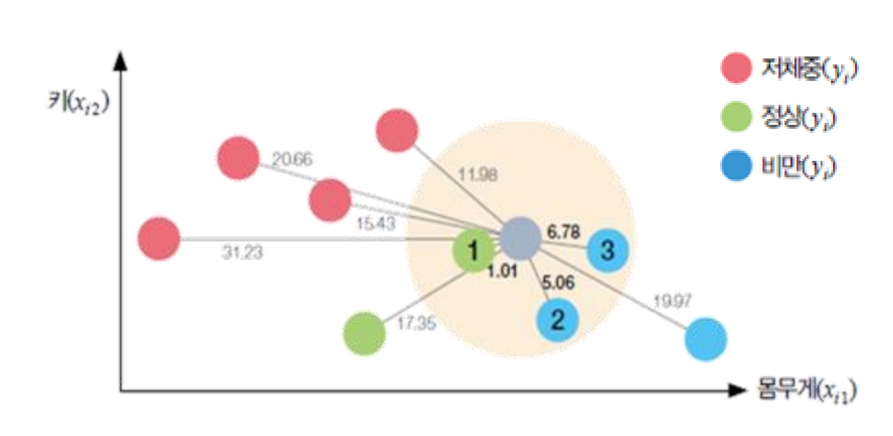
- To predict the label of new data through a 3-step process

Step 1. Distance Calculation

- Compute the distance between the new data and all training samples
- Typically use Euclidean distance (default)

Step 2. Find the Nearest Neighbors

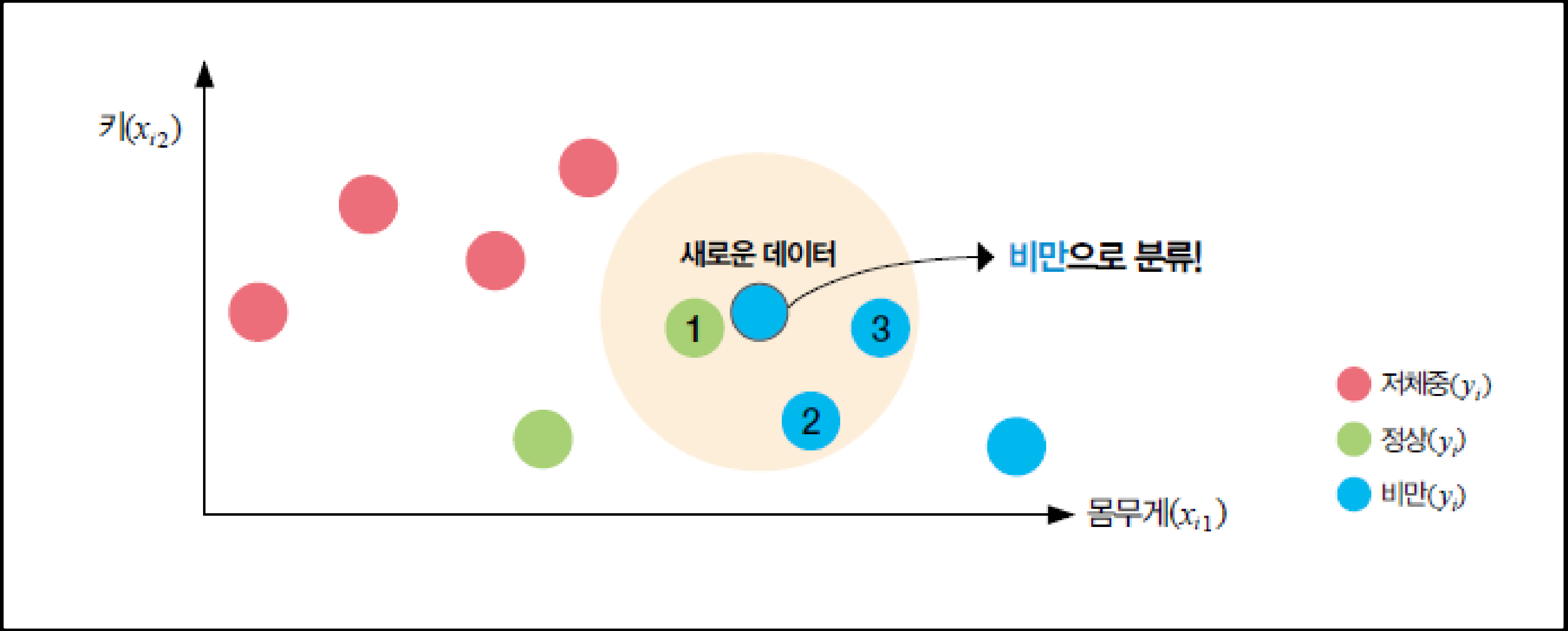
- Select the k closest data points (e.g., k = 3)



Steps in k-NN Operation (2/2)

Step 3. Make Prediction

- Predict the class label of the new data point based on majority voting among the k nearest neighbors
- Choose the class that appears most frequently among the selected neighbors



Example: k-NN Classification (1/3)

- **We are given genetic data of students and whether they were confirmed COVID-19 cases**
- **Let's predict whether a new student is COVID-positive using k-NN (e.g., $k = 1$ or $k = 3$)**

	유전자 1	유전자 2	유전자 3	유전자 4	확진 여부(y_i)
생도 A	2.54	4.33	3.99	2.57	정상
생도 B	3.12	3.87	3.84	3.04	정상
생도 C	2.76	4.17	5.63	3.28	정상
생도 D	3.87	3.56	4.25	3.65	확진
생도 E	3.55	3.91	2.68	4.22	확진
생도 F	4.12	2.86	3.30	3.71	확진
신입 생도	3.24	3.68	3.82	3.77	?

Example: k-NN Classification (2/3)

- **Compute the distance between the new student and students A to F using genetic information**
- **Determine the COVID-19 status of the new student based on the label of the nearest neighbor (k = 1)**

	유전자 1	유전자 1	유전자 2	유전자 1	확진 여부	새 관측치와의 거리
생도 A	2.54	4.33	3.99	2.57	정상	1.54
생도 B	3.12	3.87	3.94	3.04	정상	0.76
생도 C	2.76	4.17	5.63	3.28	정상	2.00
생도 D	3.87	3.56	4.25	3.65	확진	0.78
생도 E	3.55	3.91	2.68	4.22	확진	1.28
생도 F	4.12	2.86	3.30	3.71	확진	1.31
신입 생도	3.24	3.68	3.82	3.77	정상	

Example: k-NN Classification (3/3)

- **Determine the COVID-19 status of the new student by using the labels of the 3 nearest neighbors (k = 3)**

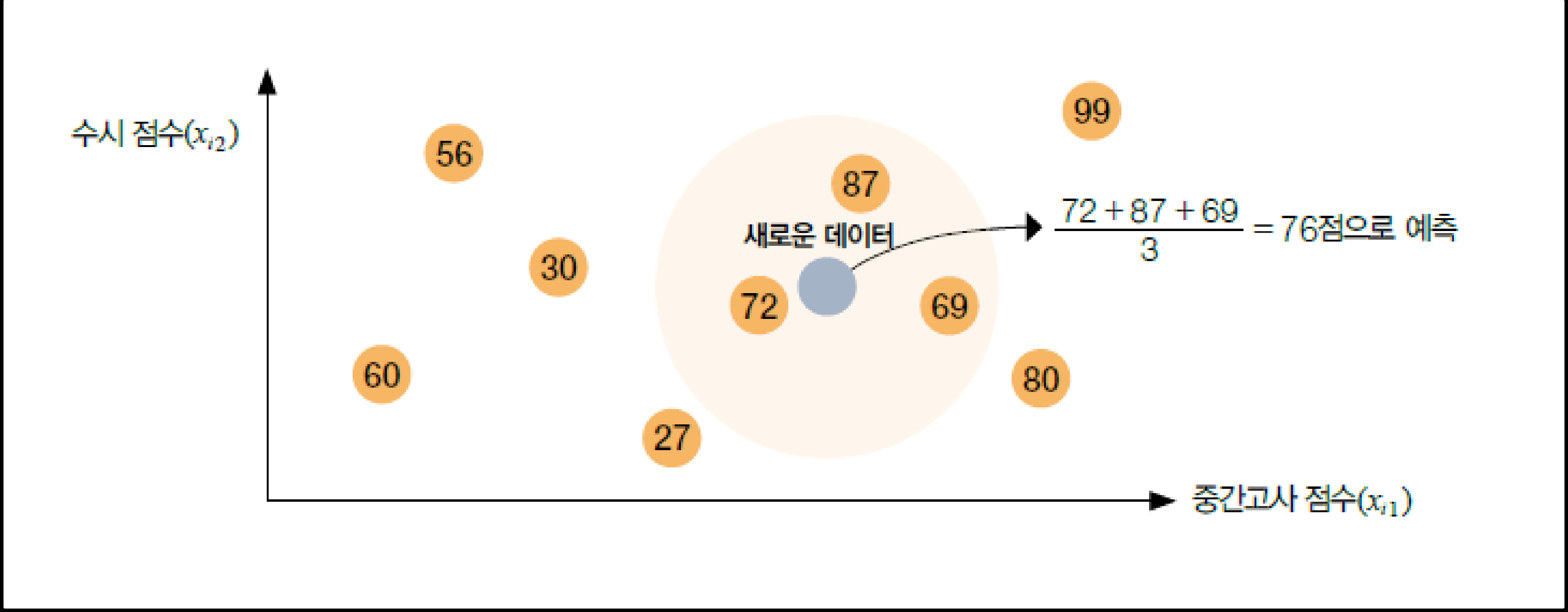
	유전자 1	유전자 2	유전자 3	유전자 4	확진 여부	새 관측치와의 거리
생도 A	2.54	4.33	3.99	2.57	정상	1.54
생도 B	3.12	3.87	3.84	3.04	정상	0.76
생도 C	2.76	4.17	5.63	3.28	정상	2.00
생도 D	3.87	3.56	4.25	3.65	확진	0.78
생도 E	3.55	3.91	2.68	4.22	확진	1.28
생도 F	4.12	2.86	3.30	3.71	확진	1.31
신입 생도	3.24	3.68	3.82	3.77	확진	

Regression with k-Nearest Neighbors

k-NN Algorithm for Regression

- If the target variable *yy* is continuous (not categorical), we apply regression
- The algorithm works similarly to classification

  → Predict the value as the average of the k nearest neighbors' values
- Optionally, you can weight neighbors inversely proportional to distance



Example: k-NN Regression (1/2)

- **We are given students' final subject scores and their final admission scores for an AI program**
- **Let's predict the admission score for a new student using k-NN (e.g., k = 1 or k = 3)**

	미적분학	기초물리학	프로그래밍	통계의 이해	군사영어	인공지능 입문
생도 A	7.5	7.5	7.0	9.5	8.5	5.0
생도 B	7.5	7.0	7.5	8.0	8.0	6.0
생도 C	8.0	7.0	8.0	8.0	8.5	8.5
생도 D	8.5	8.0	9.5	7.5	6.0	7.0
생도 E	10.0	9.5	9.0	7.5	7.5	10.0
생도 F	9.0	9.0	8.0	8.0	8.0	9.0
신입 생도	9.0	8.5	8.0	7.0	8.0	?

Example: k-NN Regression (2/2)

	미적분학	기초물리학	프로그래밍	통계의 이해	군사영어	인공지능 입문	새 관측치와의 거리
생도 A	7.5	7.5	7.0	9.5	8.5	5.0	3.28
생도 B	7.5	7.0	7.5	8.0	8.0	6.0	2.40
생도 C	8.0	7.0	8.0	8.0	8.5	8.5	2.12
생도 D	8.5	8.0	9.5	7.5	6.0	7.0	2.65
생도 E	10.0	9.5	9.0	7.5	7.5	10.0	1.87
생도 F	9.0	9.0	8.0	8.0	8.0	9.0	1.12
신입 생도	9.0	8.5	8.0	7.0	8.0	?	

순위	생도	거리	인공지능 입문 점수
①	생도 F	**1.12**	**8.0**
②	생도 C	**2.20**	**6.0**
③	생도 B	**2.40**	**8.5**

$$Prediction = \frac{8.0 + 6.0 + 8.5}{3} = 7.5$$

Pros & Cons

- **Strengths of k-Nearest Neighbors**
  - Robust to outliers
  - Can consider data distribution
  - Effective with large datasets

- **Limitations of k-Nearest Neighbors**
  - Difficulty in selecting the optimal k
  - Must choose the right distance metric
  - Higher computational cost (especially with large datasets)

Hyperparameters of k-Nearest Neighbors

Hyperparameters in k-Nearest Neighbors (1/2)

What is a Hyperparameter?

- A parameter set manually by the user, not learned through training
- Proper tuning of hyperparameters can significantly impact model performance

Role of k

- k refers to the number of nearest neighbors
- Choosing the right k is critical for performance

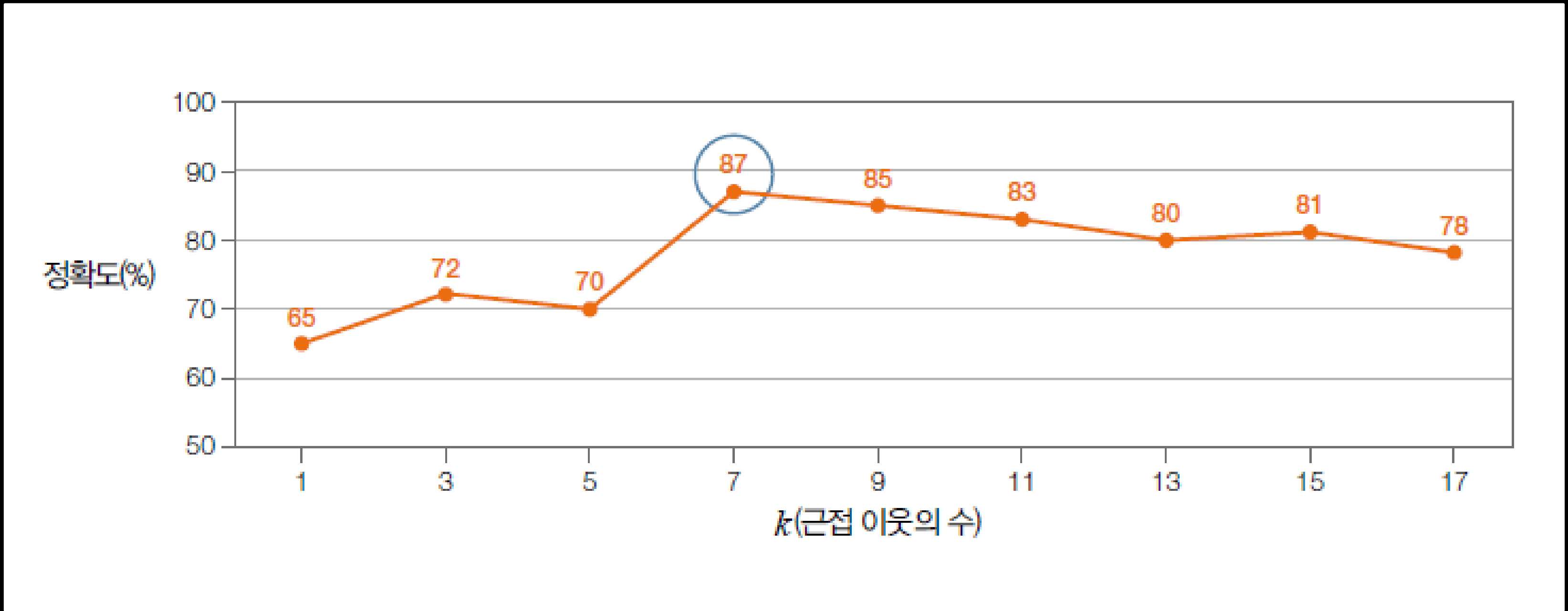
Issues with Choosing k

- Small k
  - May overfit to local noise
  - High variance, prone to outliers
- Large k
  - May include distant/irrelevant neighbors
  - Risk of underfitting and class confusion

Hyperparameters in k-Nearest Neighbors (2/2)

How to Find the Best k

- Change the value of k and evaluate model performance
- Choose the k that yields the best validation result



Implementing k-Nearest Neighbors

kNN Practice

- **GitHub repository and is linked to the textbook content**
  - Implementing k-NN using scikit-learn library from Textbook
    - https://github.com/KMA-AIData/ML/tree/main/CH07
  - Implementing by Professor
    - Use codes provided by Prof.

수고하셨습니다 ..^^..

Thank you!