

# NEAREST NEIGHBORS METHODS/ MODEL EVALUATION

Week09



# $k$ -NN

# Review: Types of Classifiers

- A classifier is a function that assigns to a sample,  $\mathbf{x}$  a class label  $\hat{y}$   
$$\hat{y} = f(\mathbf{x})$$
- A probabilistic classifier obtains conditional distributions  $\Pr(Y|\mathbf{x})$ , meaning that for a given  $\mathbf{x} \in \mathcal{X}$ , they assign probabilities to all  $y \in Y$ 
  - Hard classification

$$\hat{y} = \arg \max_y \Pr(Y = y | \mathbf{x})$$

**Any other classifiers not belonging to a probabilistic approach?**

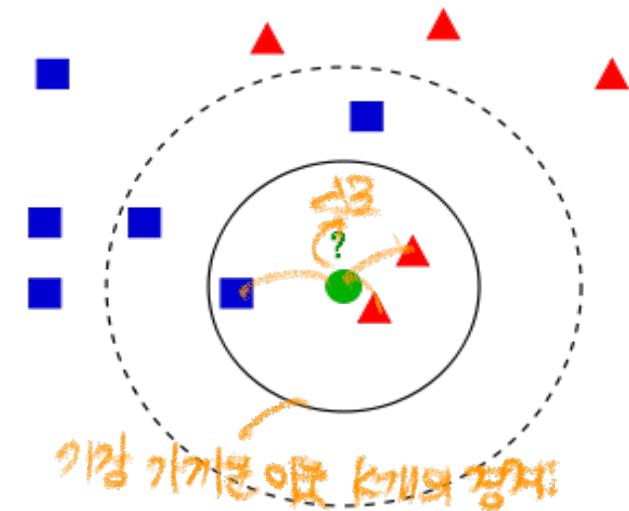
# $k$ -Nearest Neighbors( $k$ NN)

- Nonparametric method used for classification and regression

데이터 분포 모형 X → 새로운 데이터가 들어올 때 주변 이웃 보고판

- For classification

- Output class of data sample is determined by output class of its  $k$ -nearest neighbors
- Majority vote
  - assign the output class to the most common class among  $k$ -nearest neighbors



- For regression

- Output value of data sample is determined by output value of its  $k$ -nearest neighbors of the data sample
- Output value is the average value of  $k$ -nearest neighbors
  - There are several different ways to calculate average

평균은 데이터의 이웃간의 평균으로 정의

# ※ Parametric Methods

- Parametric method **정한 수의 파라미터를 사용하여 모델을 구성**
  - ▣ Parametric methods are statistical and machine learning techniques that assumes a fixed number of parameters for the model (e.g., linear regression, logistic regression)
  - ▣ They requires assumptions about the data distribution
- Key characteristics **데이터가 특정한 확률 분포를 따른다고 가정, 소수의 파라미터로 이를 표현**
  - ▣ Fixed Number of Parameters: Parametric models assume the data follows a specific distribution (e.g., normal distribution) and uses a fixed set of parameters to describe that distribution
  - ▣ Simplicity and Efficiency: The assumption of a fixed form simplifies the learning process, allowing for faster training and predictions
  - ▣ Interpretability: Parametric models are often easier to interpret because their parameters have a clear meaning within the assumed distribution.
  - ▣ Limited Flexibility: This fixed structure and the reliance on specific assumptions mean that parametric models can be less flexible and may not perform well when the data's characteristics don't match the model's assumptions **데이터가 가정한 분포와 맞지 않을 경우 성능이 떨어질 수 있음**

# ※ Nonparametric Methods

- Nonparametric method
  - ▣ Nonparametric methods are statistical and machine learning techniques that do not assume a specific form for the underlying data distribution
  - ▣ These methods are flexible and adapt to the structure of the data without relying on predefined parametric models

데이터 분포에 대해 명시적인 가정을 하지 않는 기법,  
모델 구조나 파라미터 수를 사전에 정의하지 않고 데이터 자체에 기반하여 유연하게 모델링
- Key characteristics
  - ▣ No Fixed Parameters: Unlike parametric models, they do not assume a fixed number of parameters. *특정수의 파라미터 X, 데이터의 양에 따라 보정되는 경우*
  - ▣ Data-Driven Models: The structure of the model is determined by the data rather than predefined equations.
  - ▣ Flexible and Adaptive: Can capture complex relationships without assuming a functional form.
  - ▣ Higher Computational Cost: Since they rely on the data directly, they may require more computation compared to parametric methods

# Distance Metrics

거리

- Distance metrics are mathematical functions that quantify the "distance" or dissimilarity between two points or data objects

거리

차이

- Distance measure should hold the following

- $d(x, y) \geq 0$

거리 항상 0 이상

- Non-negativity

- $d(x, y) = 0 \Leftrightarrow x = y$

두 점이 같은 값을 갖거나

- Identity of indiscernibles

- $d(x, y) = d(y, x)$

두 점 사이의 거리는 한 쪽 방향

- symmetry

- $d(x, z) \leq d(x, y) + d(y, z)$

- Subadditivity or triangle inequality

$x \rightarrow z \leq x \rightarrow y \rightarrow z$

# Distance Metrics: Numerical Data

## □ Euclidean distance

- Calculates the straight-line distance between two points in a multi-dimensional space

- Euclidean distance of two dimensional data points,  $(x_1, y_1), (x_2, y_2)$

$$\text{정의} \quad d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

- In general, Euclidean distance of two data points,  $(x_1, x_2, \dots, x_n), (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$

$$\text{정의} \quad d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \dots + (x_n - y_n)^2}$$

## □ Manhattan distance

- Calculates the sum of the absolute differences between two points

$$\sum_i^n |x_i - y_i|$$

각 축 방향으로 점 사이를  
직선으로 이동하여 거리를 측정

- Less sensitive to outliers than Euclidean distance

# Distance Metrics: Numerical Data

## □ Minkowski distance

- Generalizes both Euclidean and Manhattan distances

거리의 제곱근이 되어짐.

일반화된 거리공식

$$\left( \sum_i^n (x_i - y_i)^p \right)^{1/p}$$

■  $p=2 \rightarrow$  Euclidean distance

■  $p=1 \rightarrow$  Manhattan distance

차원마다 거리자이를 거듭제곱한 후 합산하고  
다시  $P$ 제곱근을 취함

## □ Mahalanobis distance

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T S^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

- $S$  is sample covariance matrix

공분산 행렬

법칙인 상관관계 X

- If covariance matrix is diagonal (no correlation), the resulting distance measure is as the same as the normalized distance

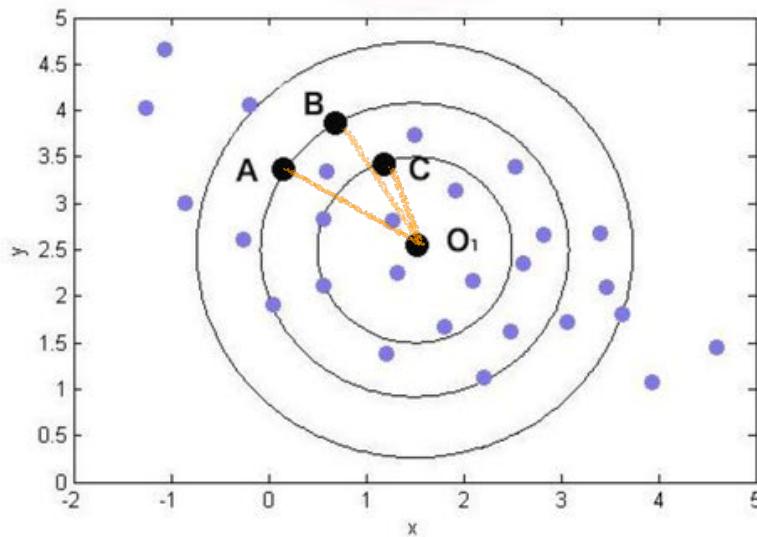
그냥 데일 평균과 같음

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^p \frac{(x_{1i} - x_{2i})^2}{s_i^2}}$$

# Distance Metrics: Numerical Data

- Comparison between Euclidean distance and Mahalanobis distance

동일한 원, 다른 중심점.

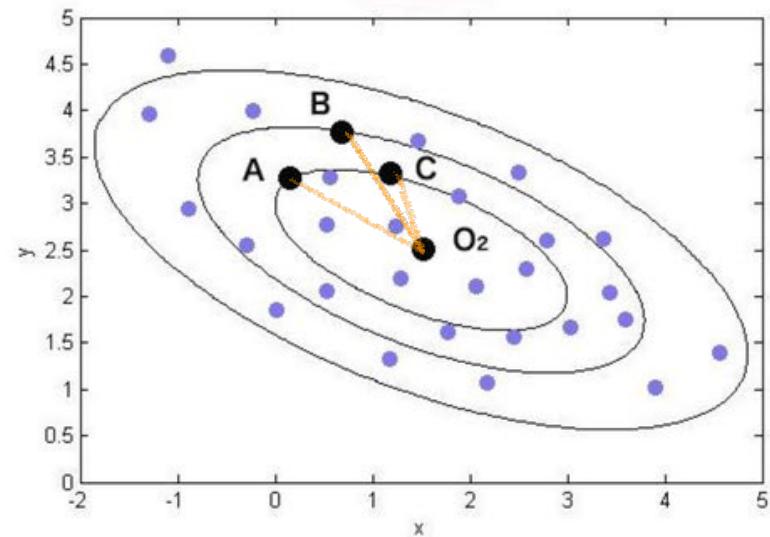


(a)

Euclidean distance

같이 놓은 두 점과 상관관계를 고려x

다른 원 = 원산 반경 < “+” 가중치



(b)

Mahalanobis distance

데이터 분포가 대칭이거나 축방향 외  
상관관계가 있을 때 유용

# Distance Metrics: Categorical Data

## □ Jaccard distance

- Used to calculate the distance between binary vectors

동일한 특성의 항목 수의 차이를 항목수로 나누어 거리 계산

		0	1
x	0	a	b
	1	c	d

- $a$ : the total number of attributes where  $x$  and  $y$  both have a value of 0
- $b$ : the total number of attributes where the attribute of  $x$  is 0 and the attribute of  $y$  is 1
- $c$ : the total number of attributes where the attribute of  $x$  is 1 and the attribute of  $y$  is 0
- $d$ : the total number of attributes where  $x$  and  $y$  both have a value of 1

$$d(x, y) = \frac{b + c}{b + c + d} \rightarrow \text{부교차항목수} \rightarrow \text{동일한 특성이 있는 항목수}$$

# Distance Metrics: Categorical Data

## □ Hamming distance

$$d(x, y) = \frac{\sum_i I(x_i \neq y_i)}{\dim(x)}$$

| 햄민거리  
vector x의 차이값이

- $I(x_i \neq y_i)$  is 1 if and only if  $x_i \neq y_i$
- $\dim(x)$  is the dimension of  $x$

방법은 binary한 머리를 계산할 때 사용

같은 값이 있는 모든 방수한 vector들에 서로 다른 위치의 개수 기반  
= 자리별로 값이 다른 수의 비율.

# Question

- Find  $k$ -nearest neighbors based on given data points

1) Find  $k$ -nearest neighbors of 5<sup>th</sup> objects when  $k=3$  using Euclidean distance

2) Find  $k$ -nearest neighbors of 5<sup>th</sup> objects when  $k=3$  using Manhattan distance

index	$x$	$y$
1	1	1
2	2	3
3	4	6
4	3	1
5	2	4
6	4	0
7	7	5
8	6	2

5번의 2단계 (2,4)

### III Euclidean distance

$$d = \sqrt{(x_2 - 2)^2 + (y_2 - 4)^2}$$

Index (x,y)

1 (1,1)  $\sqrt{(1-2)^2 + (1-4)^2} = \sqrt{1+9} = \sqrt{10}$  3.16

2 (2,3)  $\sqrt{(2-2)^2 + (3-4)^2} = 1$

3 (4,6)  $\sqrt{(4-2)^2 + (6-4)^2} = \sqrt{4+4} = 2\sqrt{2}$  2.83

4 (3,1)  $\sqrt{(3-2)^2 + (1-4)^2} = \sqrt{1+9} = \sqrt{10}$  3.16

5 (4,0)  $\sqrt{(4-2)^2 + (0-4)^2} = \sqrt{4+16} = \sqrt{20} = 2\sqrt{5}$  4.47

6 (2,5)  $\sqrt{(2-2)^2 + (5-4)^2} = \sqrt{0+1} = \sqrt{1} = 1$  LXX

7 (6,2)  $\sqrt{(6-2)^2 + (2-4)^2} = \sqrt{16+4} = 2\sqrt{5}$  4XX

### ② Manhattan distance

$$d = |x_2 - 2| + |y_2 - 4|$$

1  $|1-2| + |1-4| = 1+3 = 4$  ✓

2  $|2-2| + |3-4| = 0+1 = 1$  ✗

3  $|4-2| + |6-4| = 2+2 = 4$  ✓

4  $|3-2| + |1-4| = 1+3 = 4$  ✓

5  $|4-2| + |0-4| = 2+4 = 6$

6  $|1-2| + |5-4| = 5+2 = 7$

7  $|6-2| + |2-4| = 4+2 = 6$

가장 가까운 3개 : index 2 / index 3 / index 1 or 4

# Feature Scaling

- Scale of variable affects on determination of nearest neighbors

영화. 대체로 편향성이

- Which sample is the nearest neighbor of data sample 1?

$i$	$x_1$	$x_2$	$x_3$	$x_4$	$y$
1	9	30	100	0.5	1
2	9	25	250	0.1	0
3	9	44	220	0.7	0
4	7.5	75	170	1.2	1
...	...	...	...	...	...



$i$	Distance from $p_1$
1	-
2	150.0838
3	120.8141
4	83.23305
...	...

- Scale of variable  $x_3$  dominates over other variables
- The nearest neighbor is strongly dependent on  $x_3$

It is unfair!

→ feature scaling 이란

Min-Max 표준화

Z-Score Standardization

# Normalization

정규화

- Normalization is to adjust values of variables with different scales to common scale

모든 예제를 위한 변수의 값을 같은 크기로 변환

- There are several different ways for normalization

자기기반 알고리즘에서 특정 범위의 자료  
형태로 바꾸기

- Commonly used normalization method

Z Score normalization

$$x \rightarrow \frac{x - \mu}{\sigma}$$

평균, 표준편차 |

- $\mu$ =mean value of the variable
- $\sigma$ =standard deviation of the variable
- $\mu$  and  $\sigma$  are computed by sample data points

Min-Max normalization

$$x \rightarrow \frac{x - x_{min}}{x_{max} - x_{min}}$$

- $x_{max}$  is the maximum value of variable  $x$  and  $x_{min}$  is the minimum value of variable  $x$
- Normalized value is within [0, 1]

# Normalization

- Normalization based on normal distribution ( $x \rightarrow \frac{x-\mu}{\sigma}$ ) assumes that the sample points are distributed about the center of mass in a spherical manner
- In real data, variables are correlated with other variables

데이터가 평균을 중심으로 형으로 퍼져있다고assumes that the sample points are distributed about the center of mass in a spherical manner

본선이 흥하고 일정적 가정

Need to consider scale (level of spread along axis) and correlation to measure distance

각변수의 차이하는 변수는 상관관계가 있다  
따라 scale이 다르기 때문에 위 가정은 통할 수 있음



Mahalonobis distance

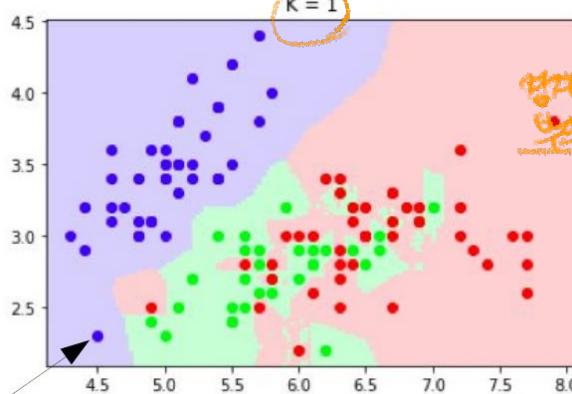
$$d(x,y) = \sqrt{(x-y)^T S^{-1} (x-y)}$$

평균 행렬 covariance matrix를 사용해 각변수간 스케일과 상관관계를 동시에 고려하는 거리 측정

# Choosing the Value of $k$

- Small  $k$  leads to high variance (overfitting) 모델이 너무 민감
- Large  $k$  smooths decision boundaries (underfitting) 가 시나리오에 부적합  
결과를 흔들어 주지 않음
- Common approach: use cross-validation to find optimal  $k$

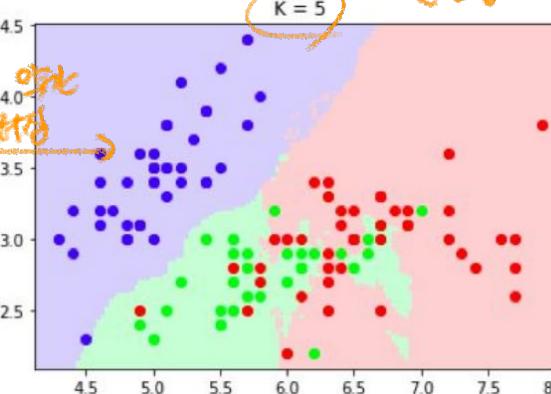
모델 경계가 매우 세밀 - 복잡



결과를 흔들어 주지 않음

결과가 예상  
부적합

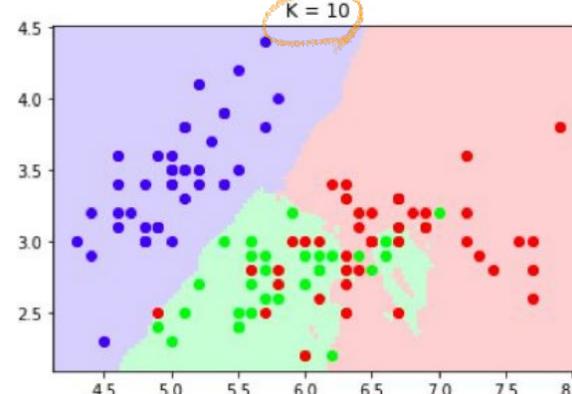
K = 5



도 보는 이웃 고려

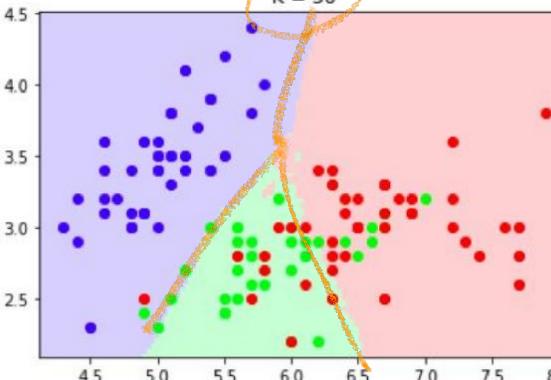
boundary가 부적합

underfitting 가능성↑



K = 10

결과가 예상  
부적합  
클래스간 미세한 차이 빈명X



K = 50

# Procedure of $k$ NN

최적의  $k$  선택 방법 찾기

Decide the number of nearest neighbors  $k$  and distance measure

For all data point in test set, find  $k$  nearest neighbors

Obtain output value (based on output values of neighbors)

classification  
regression

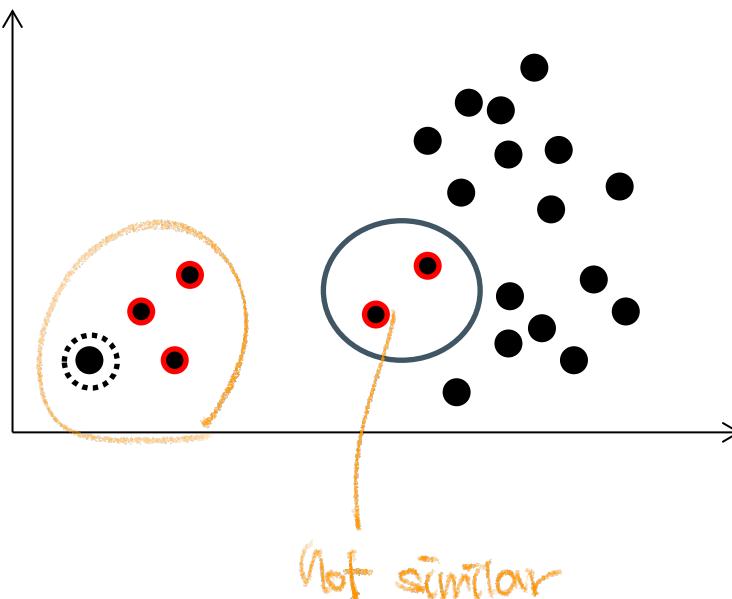
# Fixed-radius Near Neighbors

# Problem of Fixed-Number of Nearest Neighbors

고정된 K 사용시 문제점.

- When distribution of data set is not homogenous, samples not similar to data point  $x$  can be obtained in the nearest neighbors
  - $k = 5$

고정된 K개의 이웃 찾으면 의미없는 이웃 포함될 수도

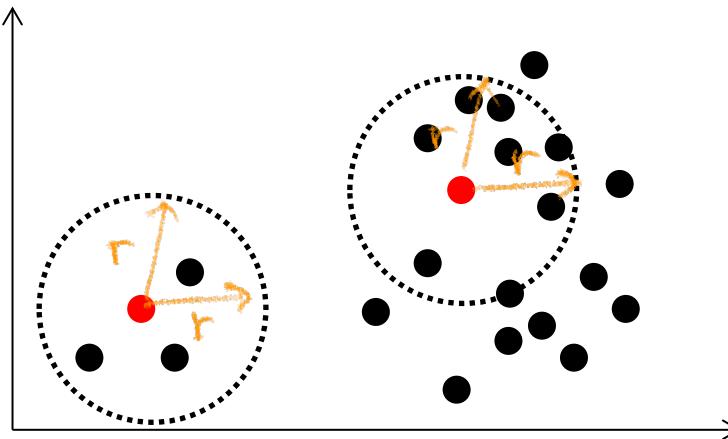


# Fixed-Radius Near Neighbors

고정반경

- Fixed-radius near neighbors are neighbors within fixed range from data point  $x$  고정된 반경 거리에 있는 모든 데이터를 이웃으로 간주
- Because of that, the number of neighbors can be differ from position

$K$  (이웃의 수)는 고정적 ✓  
거리기준은 고정



↓ 반경 r은 어떻게 정할까?

# Choosing the Value of $r$

- Small  $r$  may lead to insufficient data points 부족
  - Large  $r$  may include noisy or irrelevant data points 노이즈 or 헛된 x 데이터
  - Adjust based on dataset density  
데이터 밀도에 따라 조정해야 함
- 문 고정 되면 윤연하지만 잘못 설정하면 성능 차이 가능  
cross-validation 기준으로 성능이 가장 좋은 r 선택,

# Fixed-Radius Near Neighbors Methods

- The only difference of fixed-radius NN from  $k$ NN is the method to find the nearest neighbors 와 차이점은 이웃선택하는 방식
  - Remained steps of classification and regression are the same

Decide radius of range from data point and distance measure  
기준거리에서 이웃을 찾을 고정방법과 평균, 거리평균 선택

For all data point in test set, find fixed-radius near neighbors  
각 데이터 점마다 고정거리 안에 있는 모든 테스트 데이터를 이웃으로 탐색

Obtain output value based on output values of neighbors  
각 이웃의 출력값(예: 예측 확률 예측)

KNN : 고정 K, 거리평균 유통자, 일정 반경을 놓음 (회복치역은 KM)

fixed-radius : 기변적 K, 고정반경 구애내지만 허용, 일정 반경을 놓음  
(회복치역에서는 매우 빠른 처리)

# Pros and Cons of Nearest Neighbors

## □ Pros

- Simple and easy to implement **간단, 구현**
- No need for explicit training **명시적 학습**
- Works well with well-separated data **잘 분리된 데이터에 강함**

## □ Cons

- Computationally expensive for large datasets
- Requires a meaningful distance metric **거리 측정 규칙**
- Sensitive to irrelevant or redundant features.  
**불관련이나 중복된 특성이 많을 때**

# Evaluation of Classifiers

# Model Evaluation

모델과 가설의 타당성을 평가하는 여러 가지 지표가 사용

- In linear regression, several tests are utilized to evaluate regression models and validity of linear regression

## How about classification?

- For classification, model evaluation step is required to check validity of classification algorithms
  - Is the model well trained?
  - Is performance of the model enough?

분류학습을 평가하는  
모든 평가 단계가 필요

# Model Evaluation

## □ Confusion matrix

- A confusion matrix shows the number of correct and incorrect predictions made by the classification model compared to the actual outcomes (target value) in the data

		Real	
		Positive	Negative
Model	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

- Positive can be used as 1
- Negative can be used as 0

# Model Evaluation

- Metrics related with **confusion matrix**

		Real	
		Positive	Negative
Model	Positive	True positive (TP)	False positive (FP)
	Negative	False negative (FN)	True negative (TN)

- $\text{Accuracy} = (TP + TN) / (TP + FP + FN + TN)$  모델이 정확하게 예측한 비율
- $\text{Misclassification rate} = (FP + FN) / (TP + FP + FN + TN)$  오류율
- $\text{True positive rate} = TP / (TP + FN)$  실제 양성 중 모델이 정확하게 정답을 냈을 때 비율
  - Of all the actual positive instances, how many did the model correctly identify?
  - Also known as sensitivity or recall놓치지 않고 끌어온 끝에,
- $\text{False positive rate} = FP / (TN + FP)$  실제는 음성인데 잘못 예측한 비율
  - When it's actually negative, how often does it predict true?
- $\text{Precision} = TP / (TP + FP)$  정밀도
  - Of all the items the model labeled as positive, how many were actually positive?
- $\text{True negative rate (Specificity)} = TN / (TN + FP)$  투표

# Model Evaluation

□ F-score Precision 정밀도와 recall을 동시에 고려하는 지표.

- Consider both the precision and recall

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- Harmonic mean of precision and recall

설명서

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{\beta^2 \cdot \text{precision} + \text{recall}}$$

- The formula in terms of Type I and type II errors

$$F_\beta = \frac{(1 + \beta^2)TP}{(1 + \beta^2)TP + \beta^2FN + FP}$$

precision과 recall이  
반드시 같은  
높은 값.  
특히 database에서 적합

$\beta$  정밀도와 재현율 중 어떤 것을 더 중요하게 몇지 조정하는 parameter

$\beta > 1$  recall 가중치

$\beta < 1$  prec

$\beta = 1$  F-score의 중일

# Model Validation

# Model Evaluation

## **Is it right way to evaluate the model?**

- The same samples used for training are used to calculate accuracy
  - When learning a model, we know correct answers

**It is not fair way!**

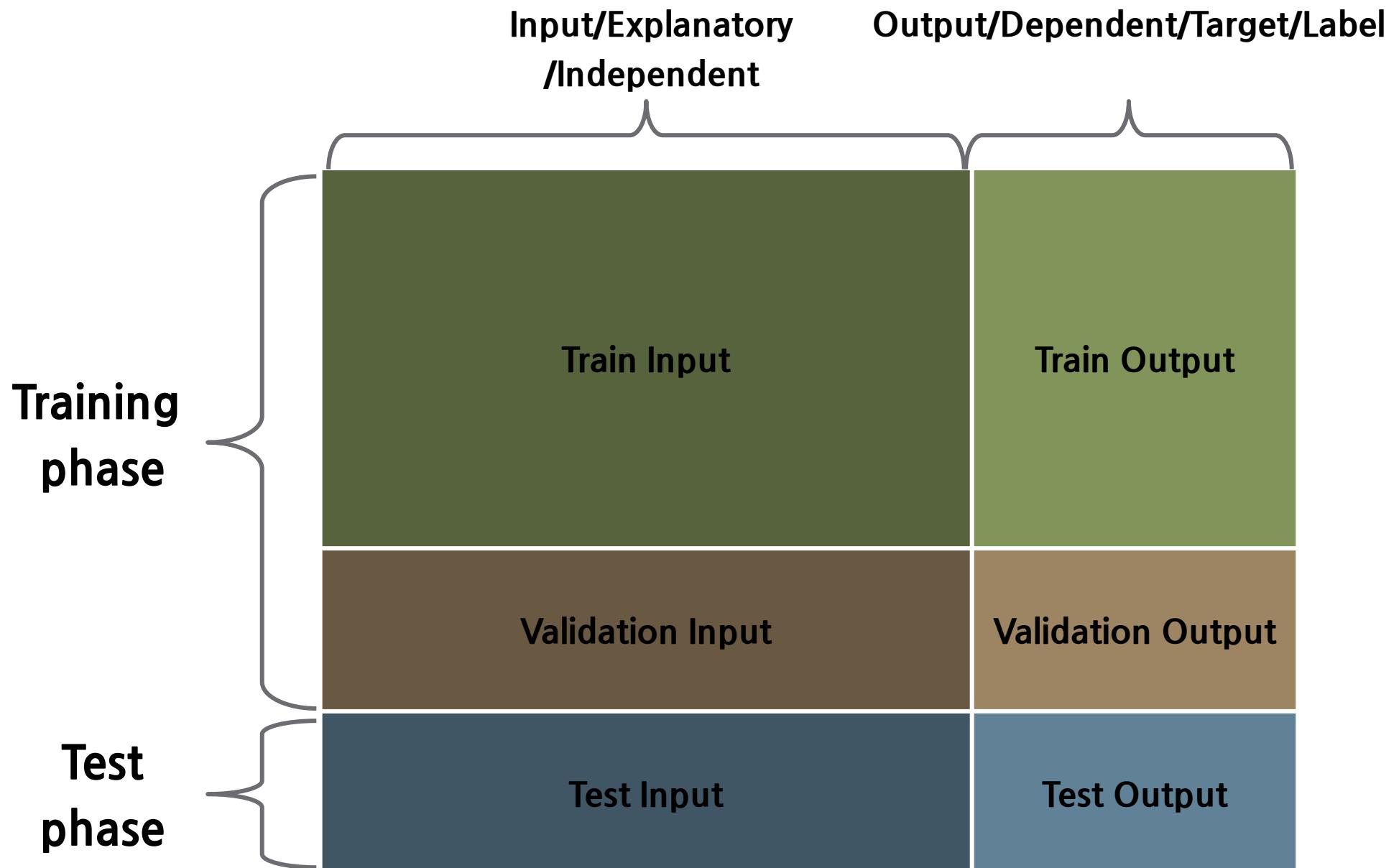
# Model Evaluation

## Is it right way to evaluate the model?

- The same samples used for training are used to calculate accuracy
  - When learning a model, we know correct answers

It is **not** fair way!

# Data Partition



# I. Cross-validation

- A model validation technique for assessing how the results of a statistical analysis will generalize to an independent data set
  - ▣ Estimate how accurately a predictive model will perform in practice
  - ▣ It is also used to determine the best set of parameters

통계학적 결과가 특정한 data set에 일반화될 수 있는지 평가하기 위한 모듈  
예를 들어 얼마나 잘 작동할지.  
최적의 파라미터 설정을 찾는 데 사용

Partitioning a sample  
of data

표본 데이터를 나눔

Performing the  
analysis on one subset

하나의 서브셋에 대해 분석

Validating the analysis  
on the other subset

다른 서브셋에 대해 모델 평가

# $k$ -fold Cross-validation

- In  $k$ -fold cross-validation, the original sample is randomly partitioned into  $k$  equal sized subsamples  
    ■ Of the  $k$  subsamples, a single subsample is retained as the validation data  
    ■ The remaining  $k - 1$  subsamples are used as training data  
    ■ The cross-validation process is then repeated  $k$  times with each of the  $k$  samples used exactly once as the validation data

한 번씩 훈련에 한 번씩 훈련에 사용



# Stratified $k$ -fold Cross-validation

- A variation of  $k$ -fold cross validation that preserves the proportion of classes in each fold  
각 폴더에서 클래스 비율이 유보가 보장된다.
  - ▣ Ensures that class distribution is approximately the same across all folds
  - ▣ Useful for imbalanced classification problems
- Advantages
  - ▣ Better handling of class imbalances
  - ▣ More representative test sets
  - ▣ Suitable for classification tasks with skewed data distributions

클래스 분포를 잘 대응

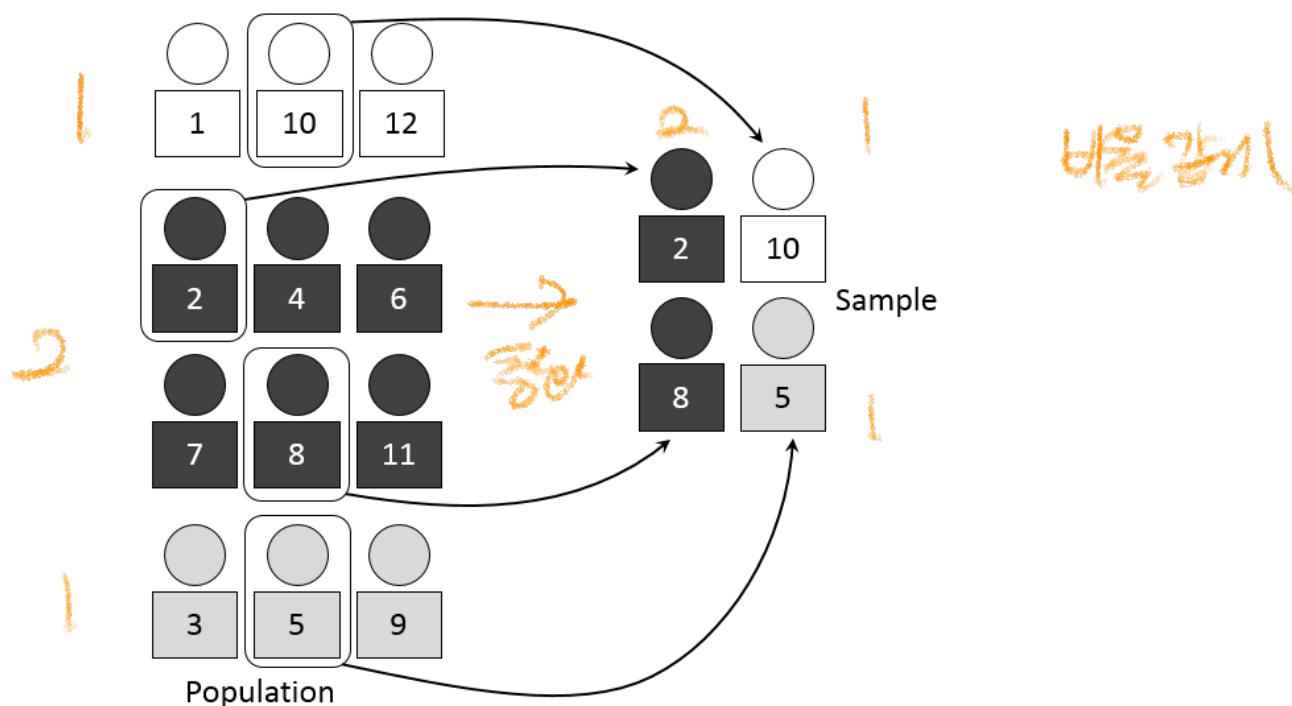
클래스 분포를 잘 대응

클래스 분포가 적합

# \* Stratified Sampling



- Stratified sampling is a method of sampling from a population which can be partitioned into subpopulations
  - ▣ For classification analysis, stratified sampling aims at splitting one data set so that each split are similar with respect to class distribution
    - To ensure that the train and test sets have approximately the same percentage of samples of each target class as the complete set



# Hyperparameter Tuning & Model Selection

