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 $\hat{y}$

$$\hat{y} = f(\mathbf{x})$$

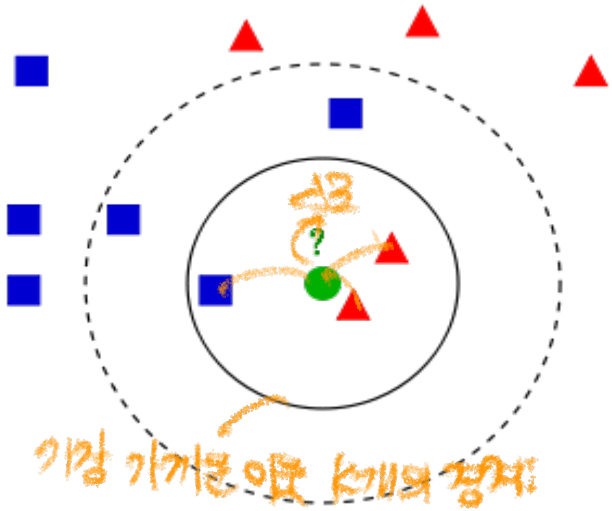
- A probabilistic classifier obtains conditional distributions $\Pr(Y|\mathbf{x})$, meaning that for a given $\mathbf{x} \in \boldsymbol{X}$, they assign probabilities to all $y \in Y$
  - Hard classification

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

Any other classifiers not belonging to a probabilistic approach?

k-Nearest Neighbors(kNN)

- Nonparametric method used for classification and regression
- 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

※ Parametric Methods

- Parametric method
  - 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
  - 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 특정한 확률 분포를 따른다고 가정, 소수의 파라미터로 이를 표현

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.
  - 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

- Distance metrics are mathematical functions that quantify the "distance" or dissimilarity between two points or data objects
- Distance measure should hold the following
  - $d(x, y) \geq 0$
    - Non-negativity
  - $d(x, y) = 0 \Leftrightarrow x = y$
    - Identity of indiscernibles
  - $d(x, y) = d(y, x)$
    - symmetry
  - $d(x, z) \leq d(x, y) + d(y, z)$
    - Subaddivity or triangle inequality

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, $(x_1, y_1), (x_2, y_2)$

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

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

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

- Manhattan distance
  - Calculates the sum of the absolute differences between two points

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

  - Less sensitive to outliers than Euclidean distance

Distance Metrics: Numerical Data

- Minkowski distance
  - Generalizes both Euclidean and Manhattan distances

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

    - p=2→ Euclidean distance
    - p=2→ Manhattan distance

- Mahalanobis distance

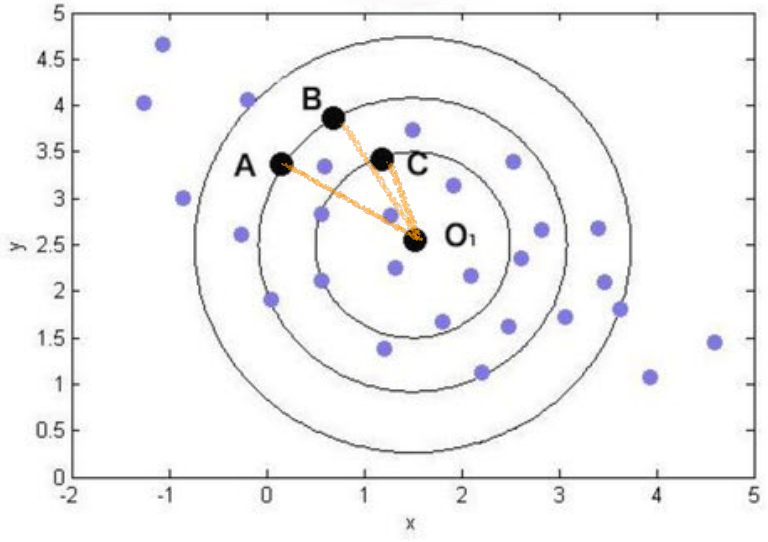
$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T S^{-1}(\mathbf{x}_1 - \mathbf{x}_2)}$$

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

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

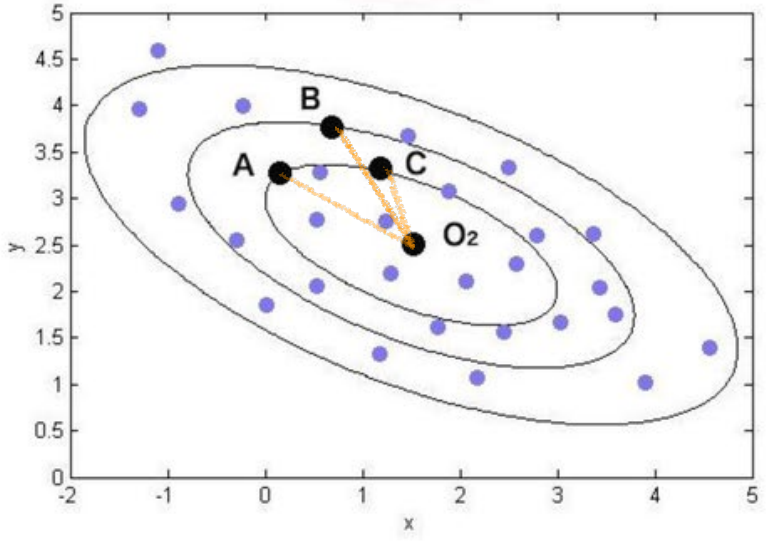
Distance Metrics: Numerical Data

- Comparison between Euclidean distance and Mahalanobis distance



(a)

Euclidean distance



(b)

Mahalanobis distance

Distance Metrics: Categorical Data

- Jaccard distance
  - Used to calculate the distance between binary vectors

y

x		0	1
	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

$$d(x, y) = \frac{b + c}{b + c + d}$$

Distance Metrics: Categorical Data

- Hamming distance

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

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

Question

- Find k-nearest neighbors based on given data points

1) Find k-nearest neighbors of 5th objects when k=3 using Euclidean distance

2) Find k-nearest neighbors of 5th 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

5번째 가까운 (2,4)

① 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
6	(4,0)	$\sqrt{(4-2)^2+(0-4)^2} = \sqrt{4+16} = \sqrt{20} = 2\sqrt{5}$	4.47
7	(1,5)	$\sqrt{(1-2)^2+(5-4)^2} = \sqrt{2+1} = \sqrt{2}$ 6. XX	
8	(6,2)	$\sqrt{(6-2)^2+(2-4)^2} = \sqrt{16+4} = 2\sqrt{5}$ 4XX	

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

② Manhattan distance

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

1	$\|1-2\| + \|1-4\| = 1+3 = 4$	V
2	$\|2-2\| + \|3-4\| = 0+1 = 1$	V
3	$\|4-2\| + \|6-4\| = 2+2 = 4$	V
4	$\|3-2\| + \|1-4\| = 1+3 = 4$	V
6	$\|4-2\| + \|0-4\| = 2+4 = 6$	
7	$\|1-2\| + \|5-4\| = 1+2 = 7$	
8	$\|6-2\| + \|2-4\| = 4+2 = 6$	

Feature Scaling

- Scale of variable affects on determination of nearest neighbors
- 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!

Normalization

- Normalization is to adjust values of variables with different scales to common scale
  - There are several different ways for normalization
- Commonly used normalization method

Z Score normalization $$x \rightarrow \frac{x - \mu}{\sigma}$$

  - μ=mean value of the variable
  - σ=standard deviation of the variable
  - μ and σ 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

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

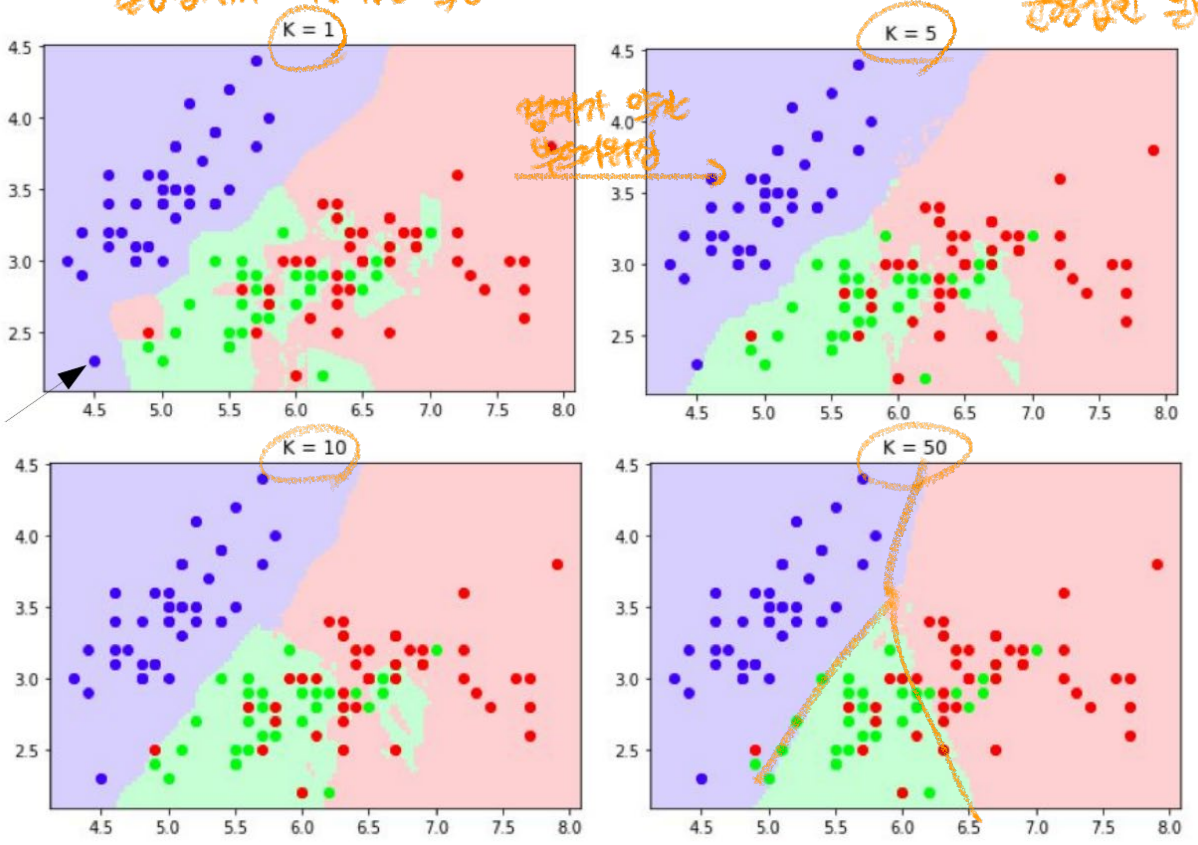
↓

Mahalonobis distance

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

Choosing the Value of k

- Small k leads to high variance (overfitting)
- Large k smooths decision boundaries (underfitting)
- Common approach: use cross-validation to find optimal k



Procedure of kNN

Decide the number of nearest neighbors k and distance measure

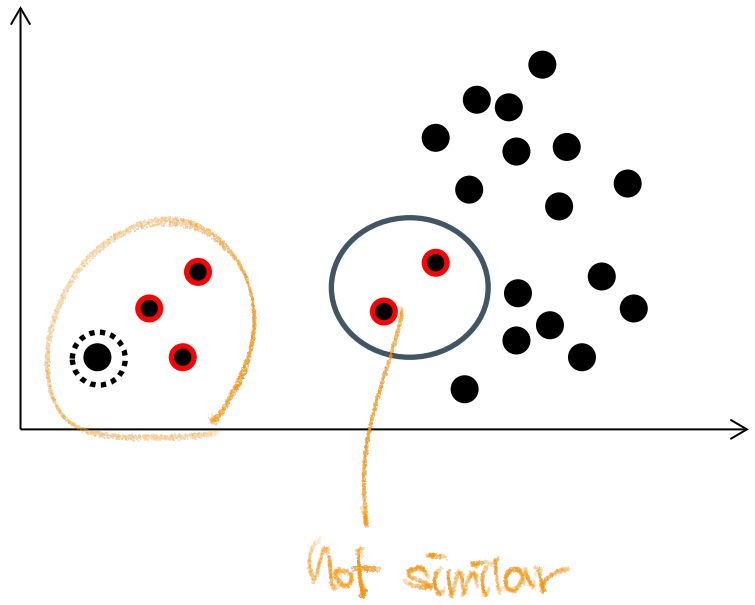
For all data point in test set, find k nearest neighbors

Obtain output value based on output values of neighbors

Fixed-radius Near Neighbors

Problem of Fixed-Number of Nearest Neighbors

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

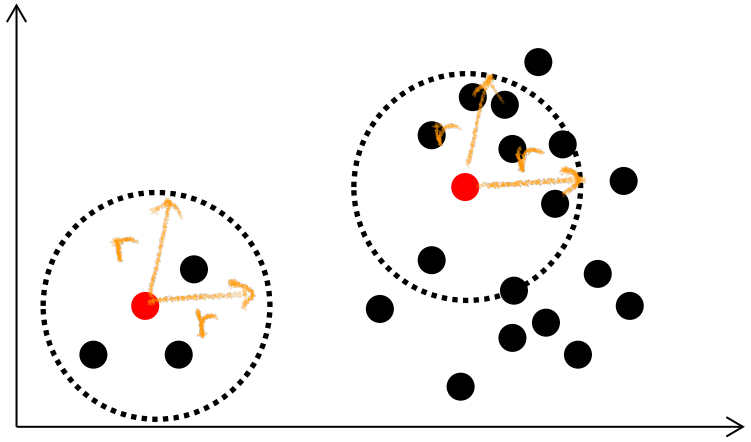


Fixed-Radius Near Neighbors

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 자연적으로 거리기준은 고정



↓ 반경 r은 어떻게 정할까?

Choosing the Value of r

- Small r may lead to insufficient data points
- Large r may include noisy or irrelevant data points
- Adjust based on dataset density

Fixed-Radius Near Neighbors Methods

- The only difference of fixed-radius NN from kNN is the method to find the nearest neighbors
  - Remained steps of classification and regression are the same

Decide radius of range from data point and distance measure

For all data point in test set, find fixed-radius near neighbors

Obtain output value based on output values of neighbors

KNN : 고정 K, 거리 방식 유동적, 밀도 민감도 낮음 (희박 지역도 K개)

Fixed-radius NN : 가변적 K, 고정반경 r 내에서만 탐색, 밀도 민감도 높음 (희박지역에서는 아웃복스 될수도)

Pros and Cons of Nearest Neighbors

- Pros
  - Simple and easy to implement
  - No need for explicit training
  - Works well with well-separated data
- Cons
  - Computationally expensive for large datasets
  - Requires a meaningful distance metric
  - Sensitive to irrelevant or redundant features.

Evaluation of Classifiers

Model Evaluation

- 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?

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)

- Accuracy = $(TP + TN)/(TP + FP + FN + TN)$
- Misclassification rate = $(FP + FN)/(TP + FP + FN + TN)$
- True positive rate = $TP/(TP + FN)$
  - Of all the actual positive instances, how many did the model correctly identify?
  - Also known as sensitivity or recall
- False positive rate = $FP/(TN + FP)$
  - When It's actually negative, how often does it predict true?
- Precision = $TP/(TP + FP)$
  - Of all the items the model labeled as positive, how many were actually positive?
- True negative rate (Specificity) = $TN/(TN + FP)$

Model Evaluation

- F-score
  - 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

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

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

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

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!

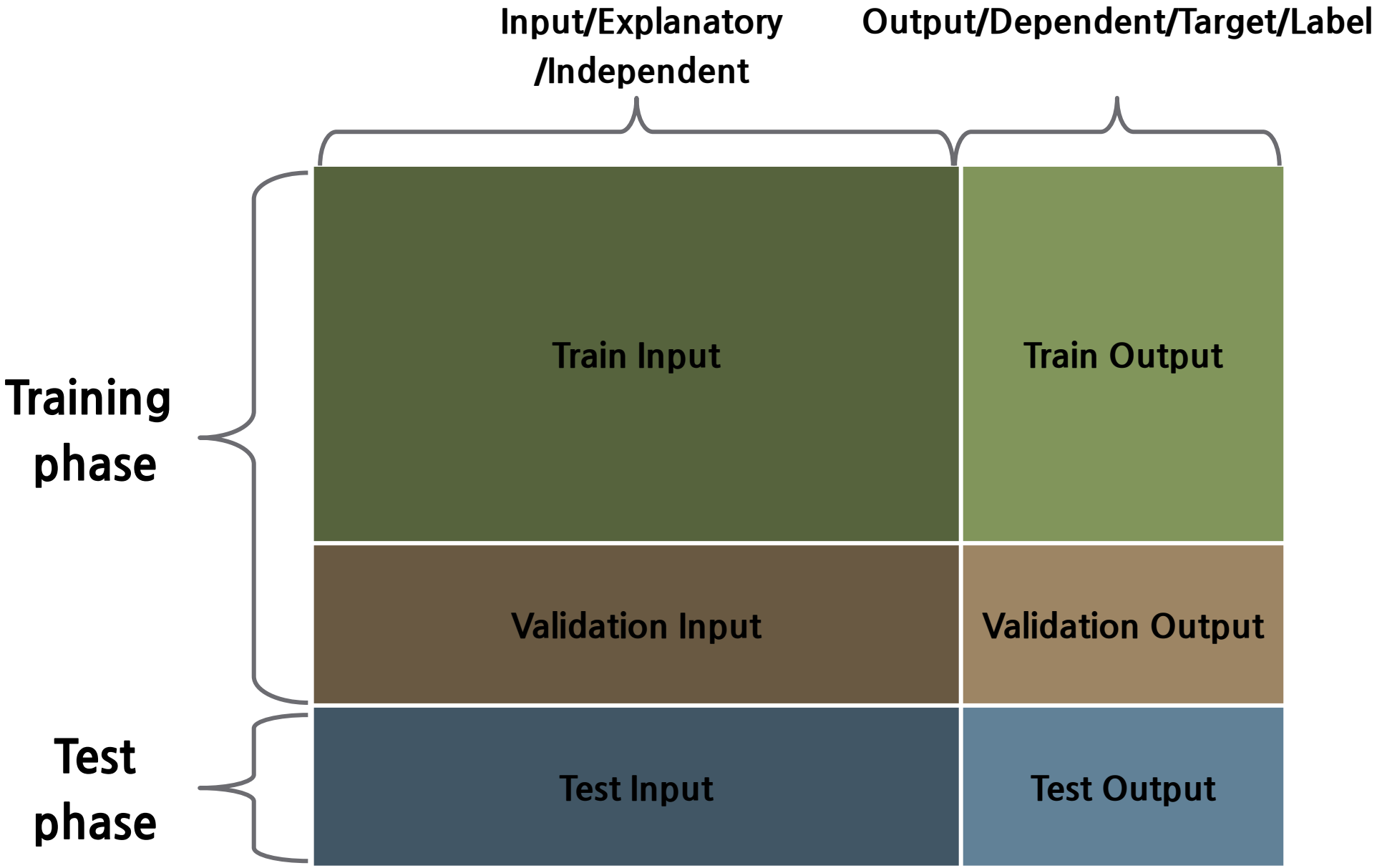
Model Evaluation

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

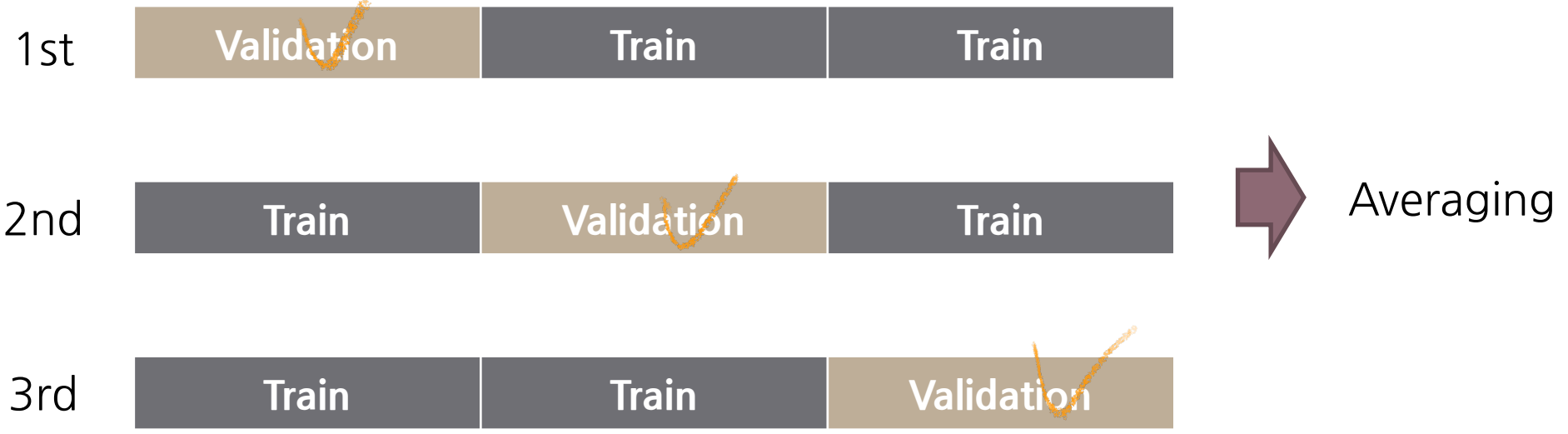


Cross-validation

- 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

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



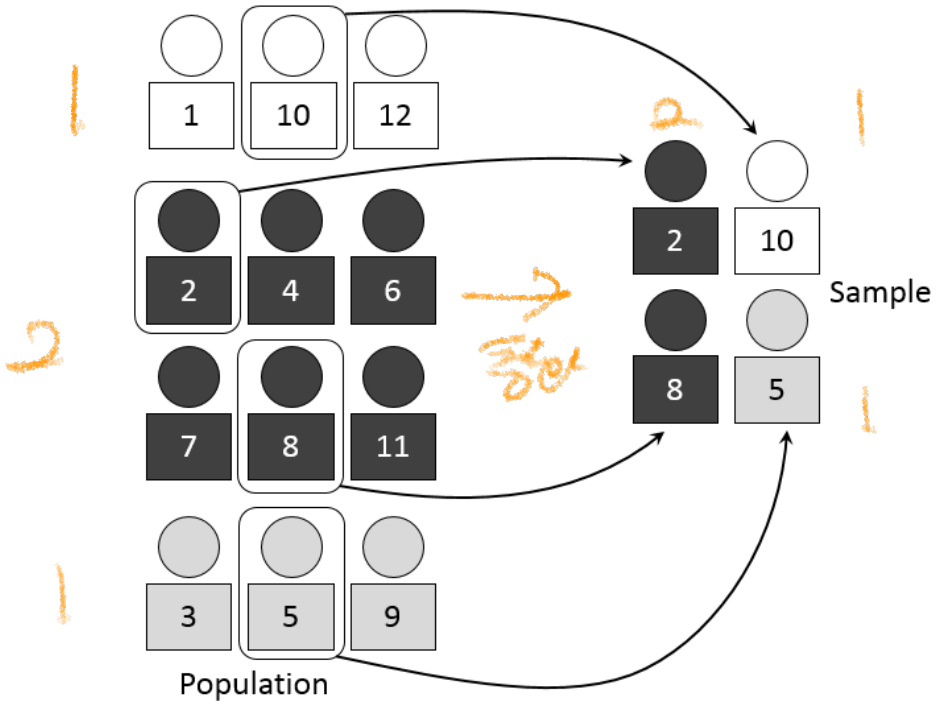
Stratified k-fold Cross-validation

- A variation of k-fold cross validation that preserves the proportion of classes in each fold
  - 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

※ 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