Korea Aerospace University

2AI4203: Machine Learning

Lec07: Dimensionality Reduction & Clustering

Jiyoon Shin

KAU College of AI Convergence

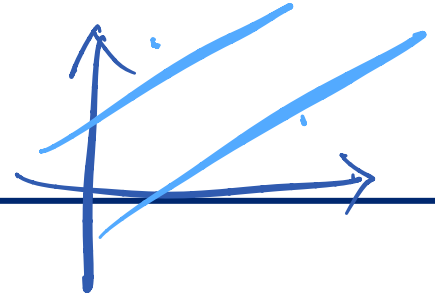
Oct. 14, 2026

Korea Aerospace University

Recap:

Support Vector Machines & Boosting

Feature Scaling





- Why scaling matters
  - Large-scale features can dominate distance calculation
  - SVM, PCA, and K-Means are sensitive to feature scaling
- Standardization

$$x_j' = \frac{x_j - \mu_j}{\sigma_j}$$

- Min-Max Scaling

$$x_j' = \frac{x_j - x_j^{\min}}{x_j^{\max} - x_j^{\min}}$$

- Avoiding data leakage
  - Fit the scaler on training data only and apply it to validation/test data

키, 몸무게, 나이

$X_1 = (160, 50, 20)$

$X_2 = (180, 65, 25)$

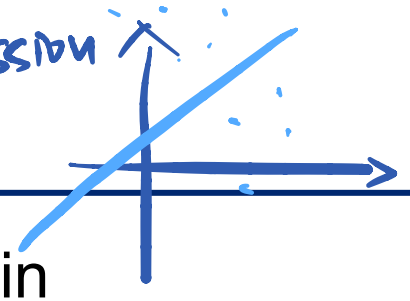
$(160-180)^2 + (\cdot\cdot)^2 + (\cdot\cdot)$

평균 = 0, 표준편차 = 1

[0, 1] 범위로.

train/val/test

Support Vector Machines (SVM)



- Maximum margin: finds the decision boundary with the largest margin
- Support vectors: closest samples that determine the decision boundary

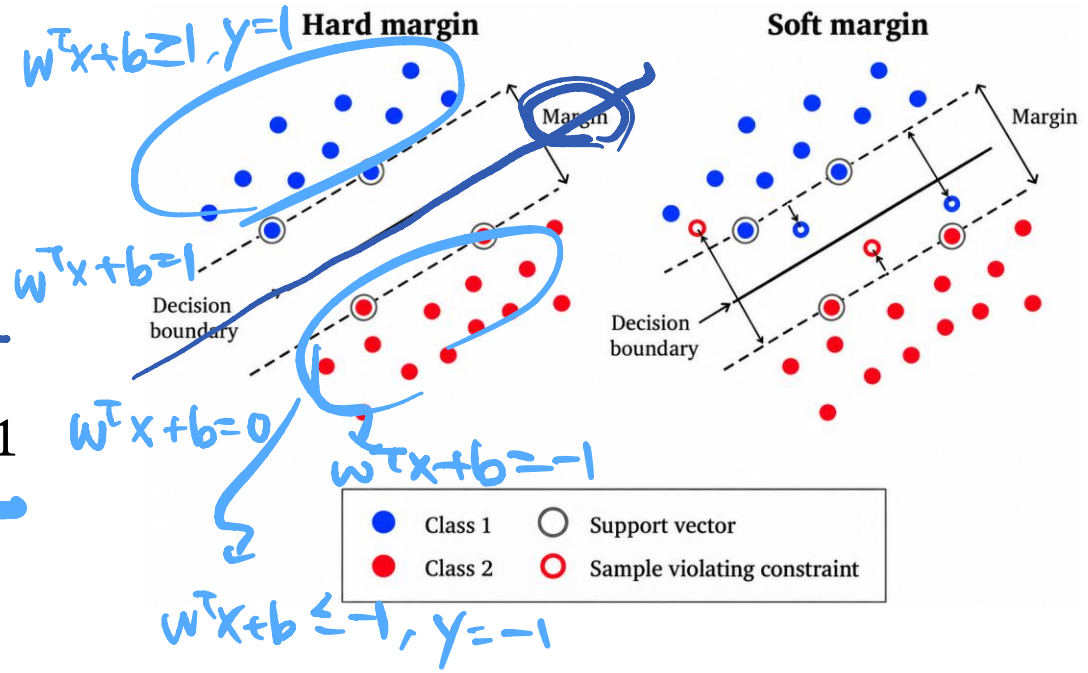
$\frac{2}{||w||}$

- Hard-Margin SVM
  - Requires linearly separable data
  - Maximizes the margin by minimizing

$$\min_{\mathbf{w},b} \frac{1}{2}||\mathbf{w}||^2$$

subject to

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1$$



Support Vector Machines

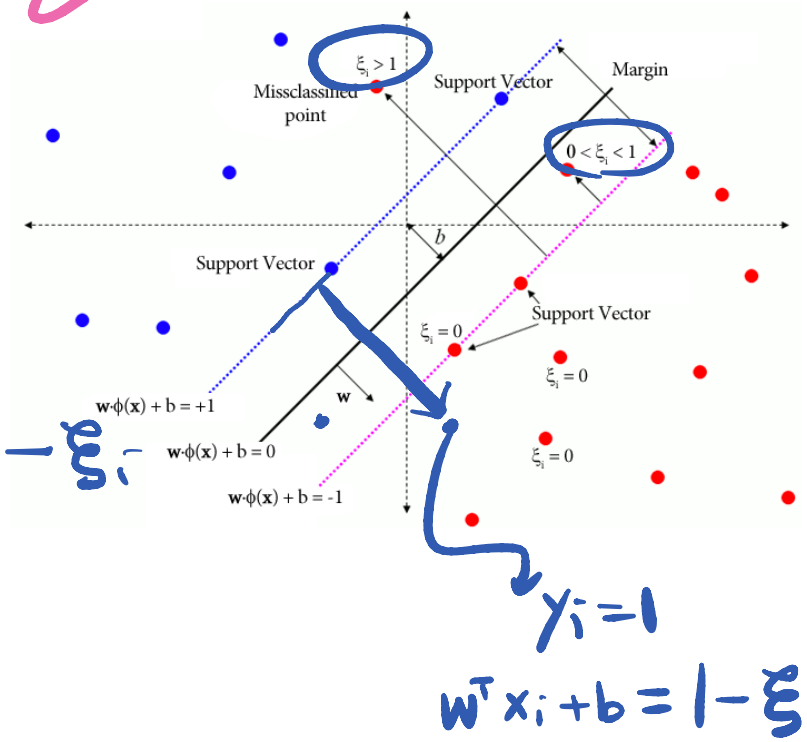
- Soft-Margin SVM
  - Allows margin violations for overlapping or noisy data
  - Balances margin size and violation penalties

$$\min_{\mathbf{w},b,\xi} \frac{1}{2}||\mathbf{w}||^2 + C\sum_{i=1}^{n}\xi_i$$

subject to

$$y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 - \xi_i, \qquad \xi_i \geq 0$$

- ξ_i: slack variable (margin violation)
- C: penalty for violations
- The role of C
  - Small C: more violations allowed, stronger regularization
  - Large C: violations penalized more, weaker regularization



Kernel Methods

SVM → linearly 분류 되어야 함! SVM

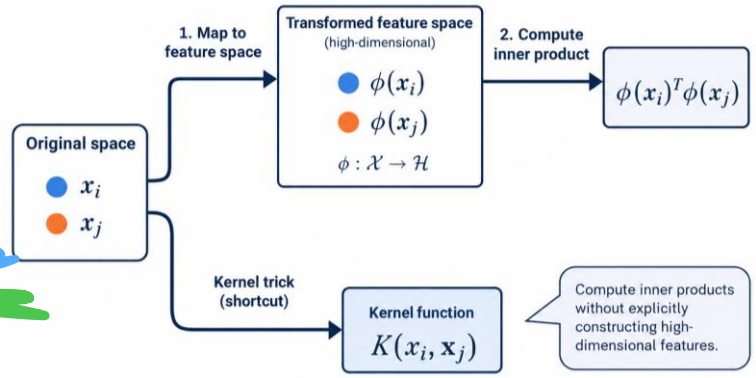
- Feature mapping

$$\mathbf{x} \longrightarrow \phi(\mathbf{x})$$

  - Maps data to a new feature space for linear separation

$x_i^T x_j$ $\phi(x_i)^T \phi(x_j)$

$x = (x_1, x_2)$ $\phi(x) = (x_1^2, x_1 x_2)$ $\phi(x)^T \phi(y) = x_1^2 y_1^2 + x_1 x_2 y_1 y_2$



- Kernel Trick

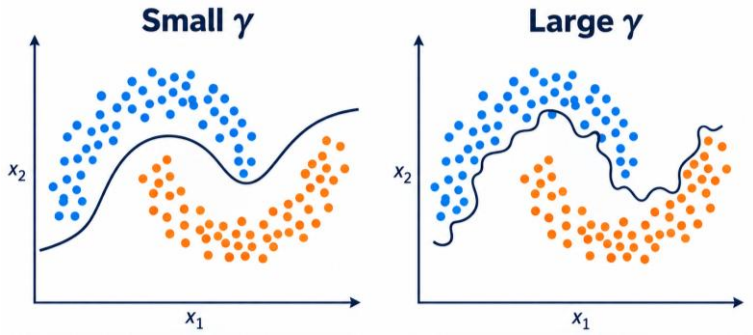
$$K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$$

  - Computes inner products without explicitly constructing transformed features

- RBF Kernel

$$K(\mathbf{x}, \mathbf{z}) = \exp(-\gamma ||\mathbf{x} - \mathbf{z}||^2)$$

  - Small γ: broad influence, smoother boundary
  - Large γ: local influence, more complex boundary

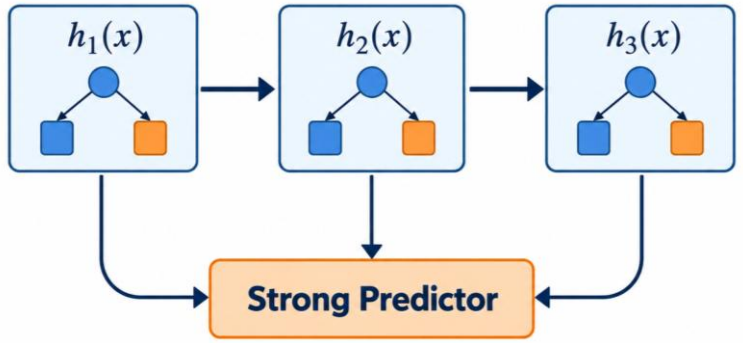


Boosting

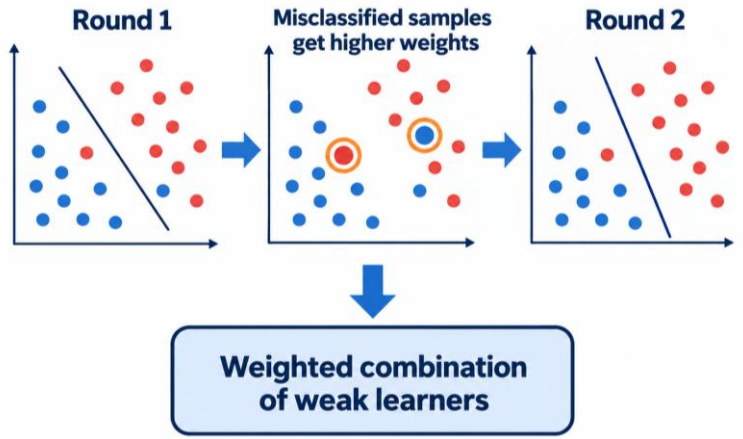
- Combines weak learners sequentially to correct errors

$$F(x) = \sum_{m=1}^{M} \alpha_m h_m(x)$$

  - h_m: weak learner
  - α_m: weight of learner m



- AdaBoost Adaptive Boosting
  - Assigns higher weights to misclassified samples
  - Combines learners through weighted voting



Gradient Boosting

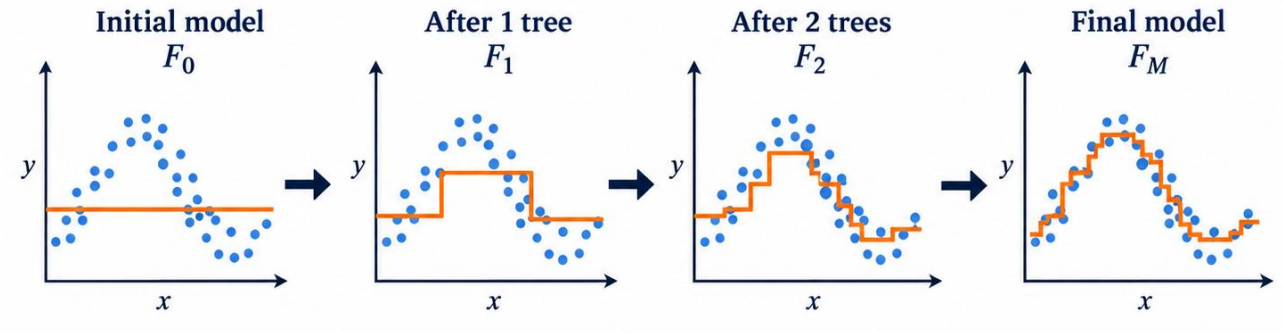
- Builds trees sequentially to correct previous errors
- For squared-error regression, learns residuals

$$r_i = y_i - F_{m-1}(\mathbf{x}_i)$$

- Additive learning

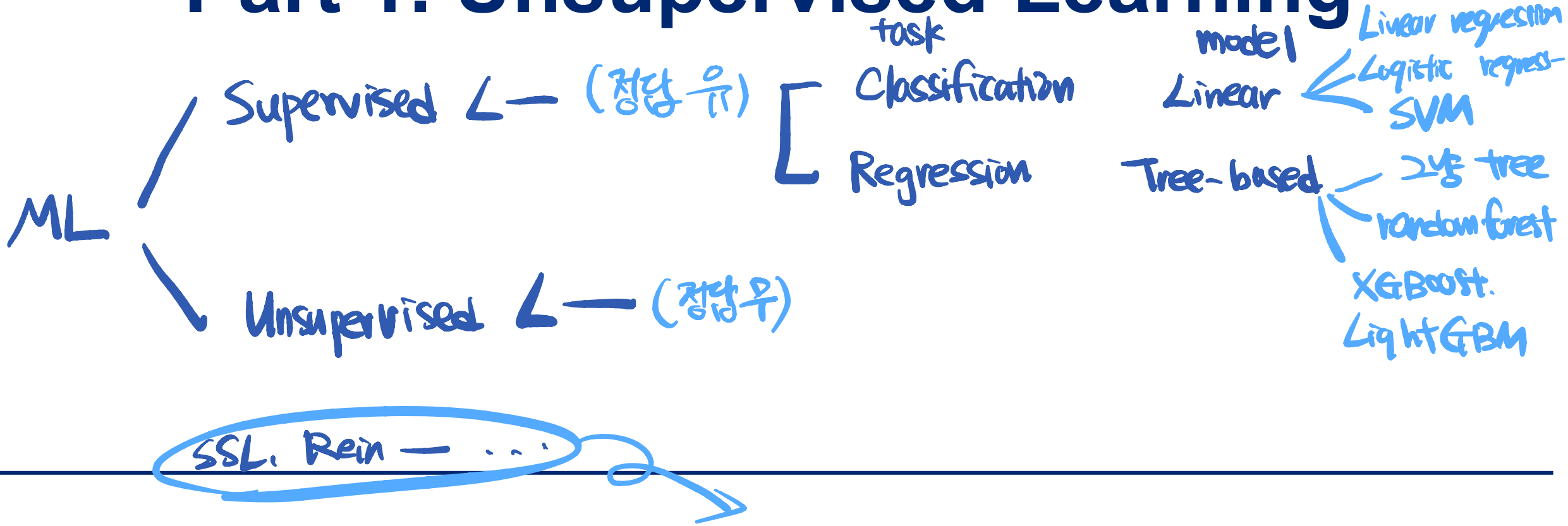
$$F_m(\mathbf{x}) = F_{m-1}(\mathbf{x}) + \eta h_m(\mathbf{x})$$

  - F_{m-1}: current model
  - h_m: new tree
  - η: learning rate



- Methods
  - XGBoost: regularization, efficient tree construction
  - LightGBM: histogram-based splitting, leaf-wise growth

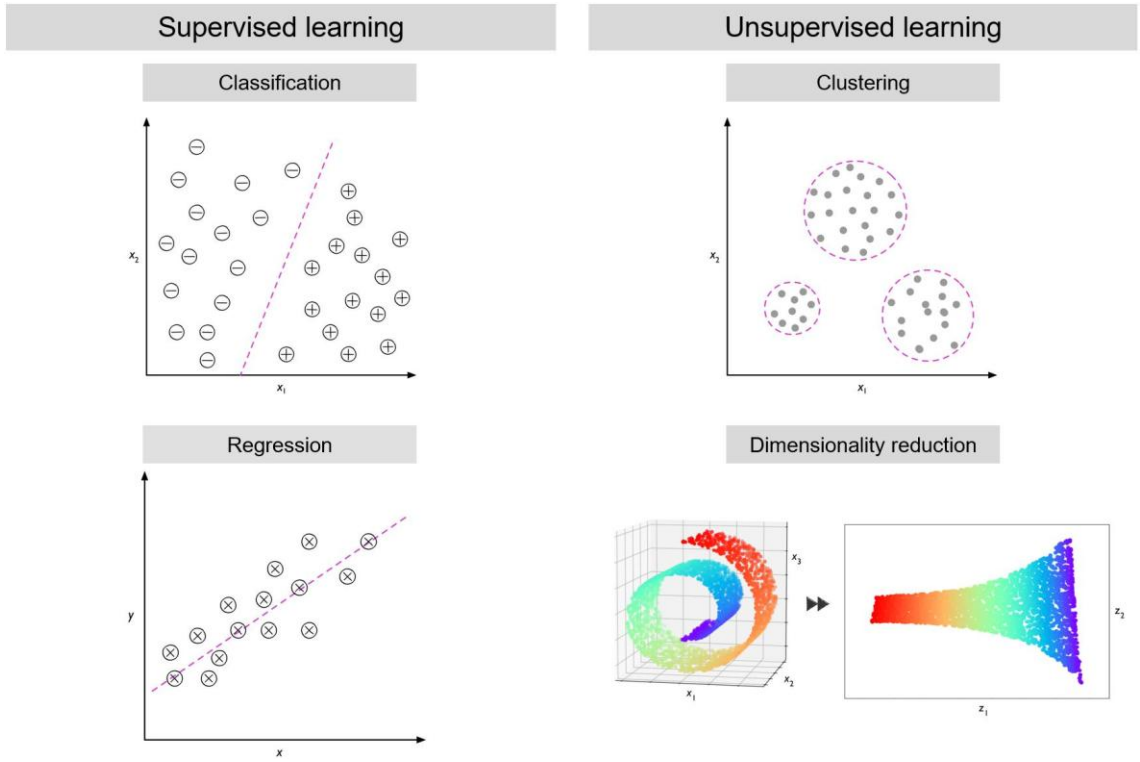
Part 1. Unsupervised Learning



Supervised vs. Unsupervised Learning

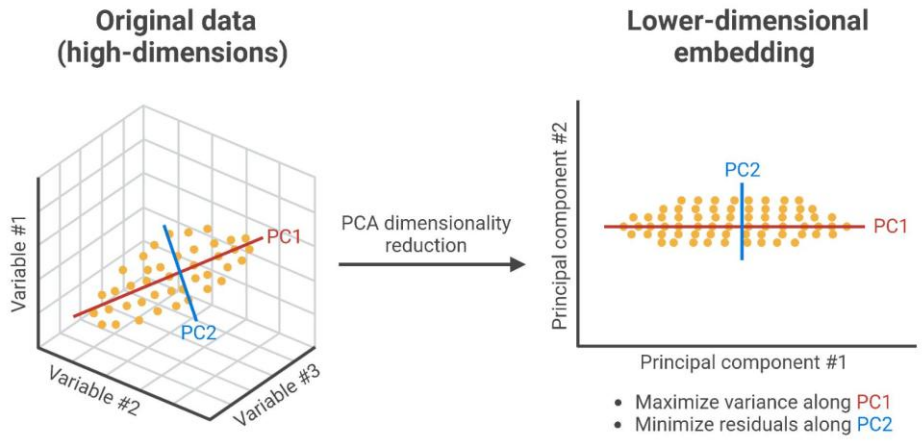
- Supervised Learning
  - Learns from labeled data
  - Predicts target values or labels
  - Examples: classification, regression

- Unsupervised Learning
  - Learns from unlabeled data
  - Discovers patterns or structures
  - Examples: dimensionality reduction, clustering



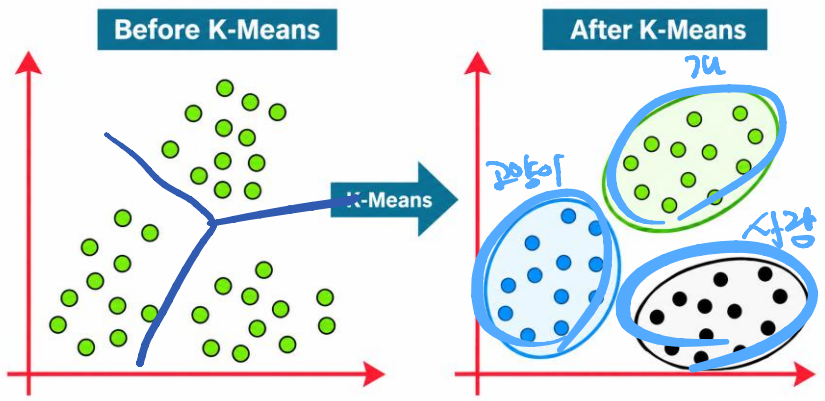
Dimensionality Reduction

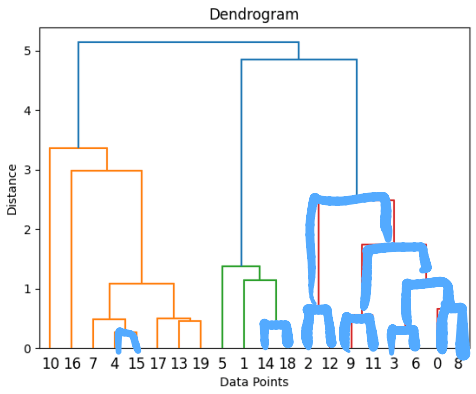
- Why reduce dimensions?
  - Real-world datasets often contain many features
  - High-dimensional data can be difficult to visualize and analyze
- Dimensionality Reduction
  - Transforms data into fewer dimensions
  - Aims to preserve important information or structure
- Example
  - Principal Component Analysis (PCA)



Clustering

- What is clustering?
  - Groups similar data points without labels
- Key idea
  - Within a cluster: data points are similar
  - Between clusters: clusters should be well separated
- Common methods
  - K-means clustering
  - Hierarchical clustering

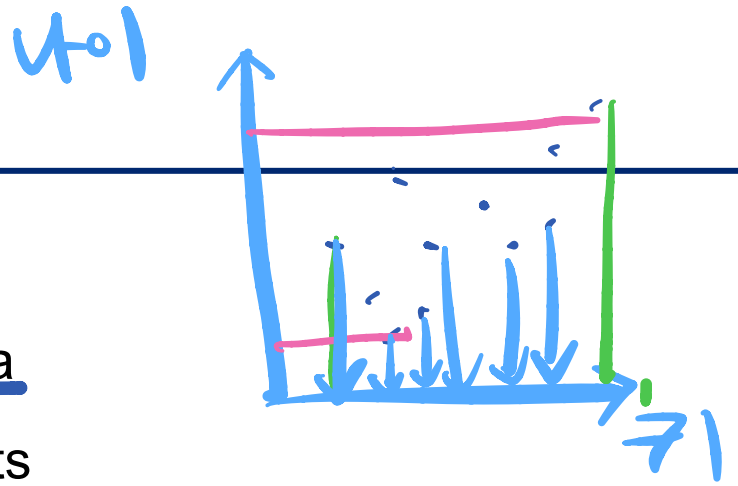




Korea Aerospace University

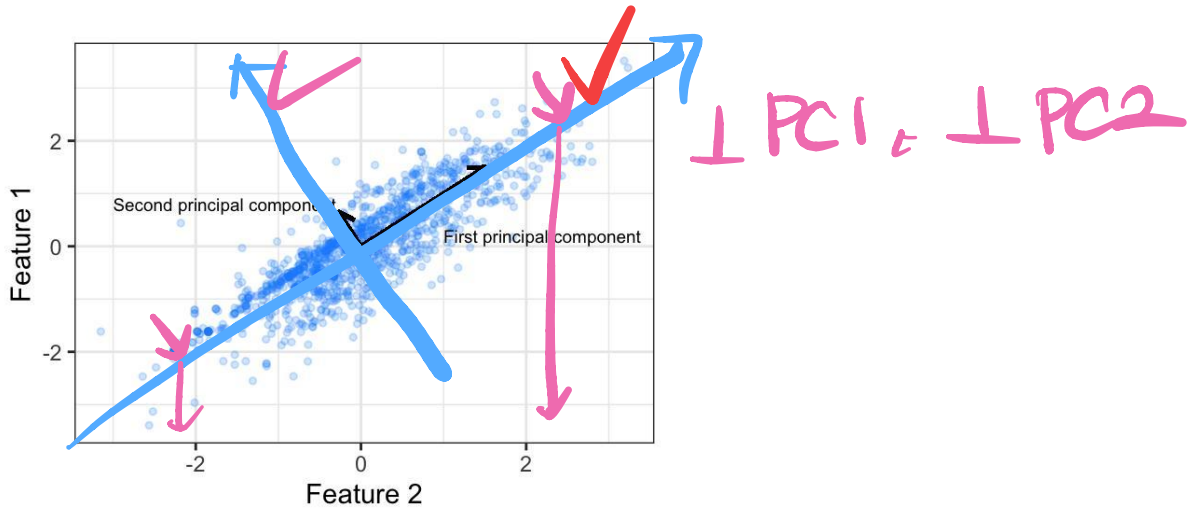
Part 2. Principal Component Analysis (PCA)

Principal Components



- Key idea of PCA
  - Finds new directions that capture maximum variance in data
  - Reduces dimensionality by retaining the leading components
- Principal Components
  - PC1: direction of maximum variance
  - PC2: direction of maximum remaining variance, perpendicular to PC1

PC3 :



Step 1

Data Preprocessing

feature 개수
data 개수

- Data matrix

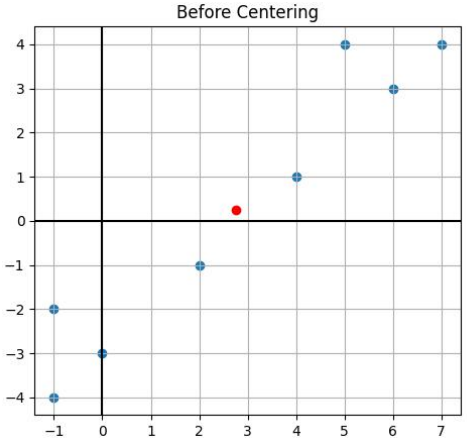
$$X = [\mathbf{x}_1, \ldots, \mathbf{x}_N] \in \mathbb{R}^{D \times N}$$

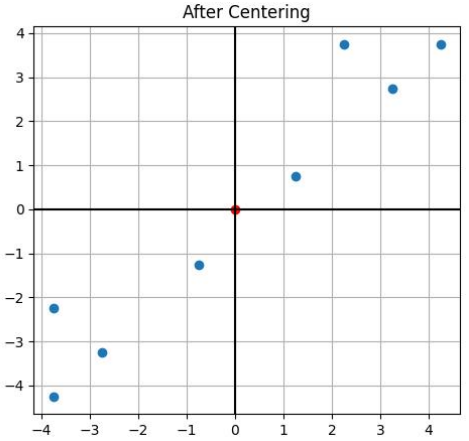
data 개수 = N
feature 개수 = D

$x_1 = \begin{pmatrix} 1 \\ 2 \\ 3 \end{pmatrix}$ $x_2 = \begin{pmatrix} 5 \\ 6 \\ 7 \end{pmatrix}$ $X = \begin{pmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \end{pmatrix}$

- Mean centering
  - Subtract the mean of each feature to center the data around zero

$$X_c = X - \boldsymbol{\mu}\mathbf{1}^T$$





$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$

- $\boldsymbol{\mu} \in \mathbb{R}^D$: feature-wise mean vector
- $\mathbf{1} \in \mathbb{R}^N$: all-ones vector

$\begin{pmatrix} 1 \\ 1 \end{pmatrix}$

$x - \mu$

$\begin{pmatrix} 1 & 5 \\ 2 & 6 \\ 3 & 7 \end{pmatrix} - \begin{pmatrix} \mu_1 & \mu_1 \\ \mu_2 & \mu_2 \\ \mu_3 & \mu_3 \end{pmatrix}$

Data Preprocessing

- (optional) Feature scaling

$$x'_{d,n} = \frac{x_{d,n} - \mu_d}{\sigma_d}$$

  - Standardization is useful when features have different scales
  - Prevents large-scale features from dominating PCA

Step 2.

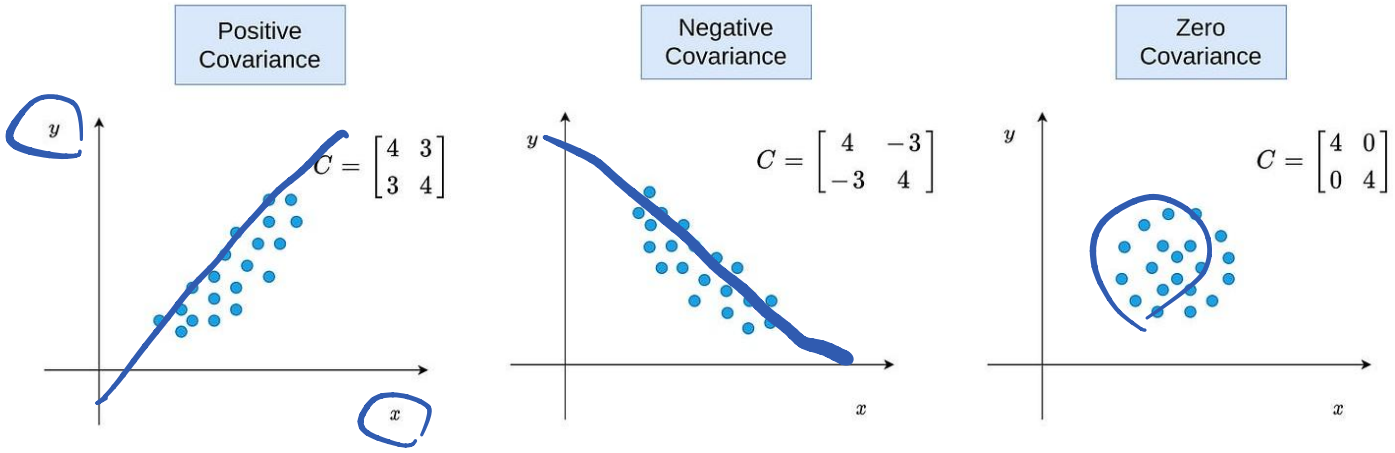
Covariance Matrix

- Variance & Covariance
  - 분산 • Variance: measures the spread of a feature
  - 공분산 • Covariance: measures how two features vary together
- Covariance Matrix

$$S = \begin{bmatrix} \mathrm{Var}(f_1) & \mathrm{Cov}(f_1, f_2) \\ \mathrm{Cov}(f_1, f_2) & \mathrm{Var}(f_2) \end{bmatrix}$$

feature 개수 : data 개수

$$S = \frac{1}{N} X_c X_c^T,$$



$$S = \frac{1}{N} X_c X_c^T$$

$$X_c \in \mathbb{R}^{2\times 2} \quad (N = 2,\ D = 2)$$

$$S = \begin{bmatrix} \text{var}(f_1) & \text{cov}(f_1, f_2) \\ \text{cov}(f_2, f_1) & \text{var}(f_2) \end{bmatrix}$$

$$\text{var}(X) = E\left[(X - \mu_X)^2\right]$$

$$\text{cov}(X, Y) = E\left[(X - \mu_X)(Y - \mu_Y)\right]$$

$$x_1 = \begin{pmatrix} f_1 \\ g_1 \end{pmatrix} \qquad x_2 = \begin{pmatrix} f_2 \\ g_2 \end{pmatrix} \rightarrow X = \begin{pmatrix} f_1 & f_2 \\ g_1 & g_2 \end{pmatrix}$$

$$\rightarrow X_c = \begin{pmatrix} f_1 - \mu_f & f_2 - \mu_f \\ g_1 - \mu_g & g_2 - \mu_g \end{pmatrix}$$

$$S = \frac{1}{2} \begin{pmatrix} f_1 - \mu_f & f_2 - \mu_f \\ g_1 - \mu_g & g_2 - \mu_g \end{pmatrix} \begin{pmatrix} f_1 - \mu_f & g_1 - \mu_g \\ f_2 - \mu_f & g_2 - \mu_g \end{pmatrix} =$$

$$\begin{pmatrix} \dfrac{(f_1 - \mu_f)^2 + (f_2 - \mu_f)^2}{2} & \dfrac{(f_1 - \mu_f)(g_1 - \mu_g) + (f_2 - \mu_f)(g_2 - \mu_g)}{2} \\ \dfrac{(f_1 - \mu_f)(g_1 - \mu_g) + (f_2 - \mu_f)(g_2 - \mu_g)}{2} & \dfrac{(g_1 - \mu_g)^2 + (g_2 - \mu_g)^2}{2} \end{pmatrix}$$

$\text{var}(f)$ $\quad$ $\text{cov}(f, g)$ $\quad$ $\text{cov}(f, g)$ $\quad$ $\text{var}(g)$

Step 3.

Eigenvectors & Eigenvalues

- Finding principal directions
  - PCA finds directions that maximize projected variance

$$\mathrm{Var}(\mathbf{b}^T \mathbf{x}_c) = \mathbf{b}^T S\mathbf{b} \qquad (||\mathbf{b}|| = 1)$$

$$\max \quad \frac{1}{N} b^T x_c (b^T x_c)^T = \frac{1}{N} b^T x_c x_c^T b = b^T \frac{1}{N} x_c x_c^T b = b^T S b$$

$$\mathrm{Cov}(X) = \frac{1}{N} XX^T$$

- Eigenvalue Decomposition

$$S\mathbf{b}_j = \lambda_j \mathbf{b}_j$$

  - Eigenvectors ($\mathbf{b}_j$): principal directions
  - Eigenvalues (λ_j): variances along those directions

$Av = \lambda v$ ⇨ λ가 A의 eigenval
v가 A의 eigen vec

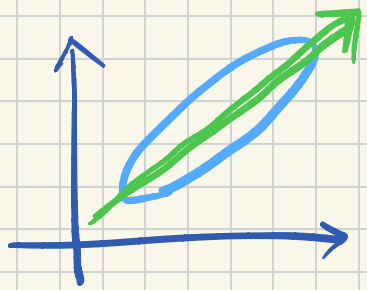
과정 생략

최종목표:

$\mathrm{var}(b^T x_c) = b^T S b$가 최대가 되는 $||b||=1$인 b를 찾는 것!

$Sb = \lambda b$

PCA의 목표: data 분포 내에서 가장 큰 variance의 방향을 찾아야함.





$b\ (1,-2)$

$X_c \xrightarrow[\text{왜?}]{} b^T X_c$

$\text{var}(b^T x_c)$가 최대가 되는 b를 찾자!

$$\text{var}(b^T x_c) = \frac{1}{N}(b^T x_c)(b^T x_c)^T = b^T S b \quad \left(S = \frac{1}{N} x_c x_c^T\right)$$

$b^T S b$가 최대가 되는 b를 찾자!

최적화를 하면 ---

$Sb = \lambda b$를 만족하는 b를 찾자!

S의 eigen vector!!

Eigenvectors & Eigenvalues

$$\mathrm{Var}(\mathbf{b}_j^T \mathbf{x}_c) = \mathbf{b}_j^T S \mathbf{b}_j = \lambda_j$$

- Ordering principal components

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_D$$

  - Larger eigenvalues indicate greater variance
  - Keep the eigenvectors with the largest eigenvalues

Explained Variance

- Explained variance ratio

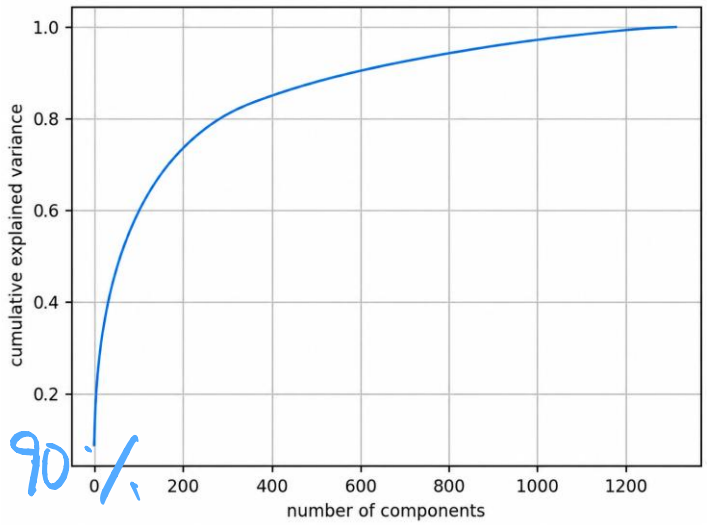
$$\mathrm{EVR}_j = \frac{\lambda_j}{\sum_{i=1}^{D} \lambda_i}$$

  - Proportion of total variance explained by each component

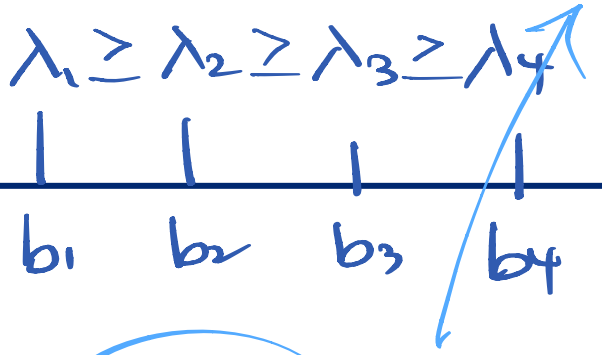
- Choosing the number of components

$$\text{Cumulative EVR}(M) = \frac{\sum_{j=1}^{M} \lambda_j}{\sum_{i=1}^{D} \lambda_i}$$

  - Select the smallest M that preserves sufficient variance
  - Common thresholds: 90% or 95%



Data Projection



- Projection Matrix

$$B = [\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_M] \in \mathbb{R}^{D \times M}$$

  - Choose from the top M eigenvectors

- Projecting Data

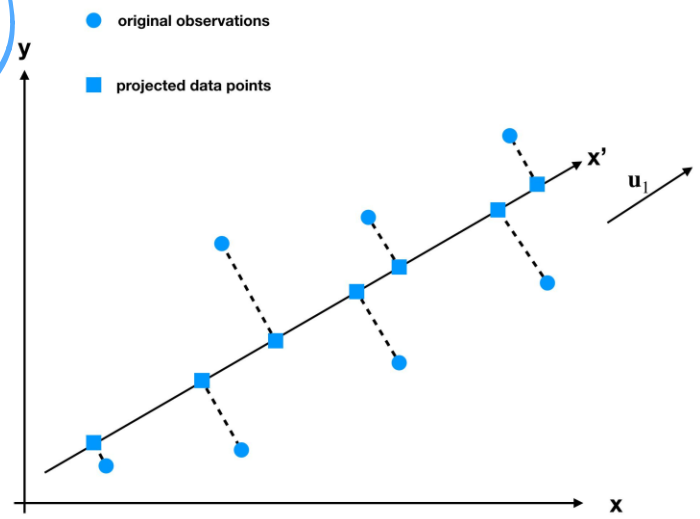
$$Z = B^T X_c$$

  - Projects data onto the selected principal components
  - Reduces dimensionality from D to M

$\mathbb{R}^{D \times N}$

$\mathbb{R}^{M \times D} \cdot \mathbb{R}^{D \times N}$

$\mathbb{R}^{M \times N}$



PCA Example

- Original Data

$$X = \begin{bmatrix} 1 & 2 & 3 \\ 1 & 2 & 3 \end{bmatrix} \in \mathbb{R}^{D \times N}$$

$x_1 = \begin{pmatrix} 1 \\ 1 \end{pmatrix}$ $x_2 = \begin{pmatrix} 2 \\ 2 \end{pmatrix}$ $x_3 = \begin{pmatrix} 3 \\ 3 \end{pmatrix}$

$\mu_1 = 2$

$\mu_2 = 2$

- Step 1. Mean centering

$$X_c = \begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}$$

$$\frac{1}{3}\begin{bmatrix} -1 & 0 & 1 \\ -1 & 0 & 1 \end{bmatrix}\begin{bmatrix} -1 & -1 \\ 0 & 0 \\ 1 & 1 \end{bmatrix}$$

- Step 2. Covariance Matrix

$$S = \frac{1}{3} X_c X_c^T = \frac{1}{3}\begin{bmatrix} 2 & 2 \\ 2 & 2 \end{bmatrix}$$

PCA Example

$Sb = \lambda b$

- Step 3. First principal component

$$\det\begin{pmatrix} \frac{2}{3}-\lambda & \frac{2}{3} \\ \frac{2}{3} & \frac{2}{3}-\lambda \end{pmatrix} = 0$$

$$\left(\frac{2}{3}-\lambda\right)^2 - \left(\frac{2}{3}\right)^2 = 0$$

$$\left(\frac{2}{3}-\lambda-\frac{2}{3}\right)\left(\frac{2}{3}-\lambda+\frac{2}{3}\right) = 0 \Rightarrow \lambda = 0, \frac{4}{3}$$

$$\lambda_1 = \frac{4}{3}, \quad \lambda_2 = 0$$

$$\mathbf{b}_1 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\end{bmatrix}, \quad \mathbf{b}_2 = \frac{1}{\sqrt{2}}\begin{bmatrix}1\\-1\end{bmatrix}$$

$Sb - \lambda b = 0$

$(S-\lambda I)\, b = 0 \quad (b \neq 0)$

$\det(S-\lambda I) = 0$

$\det(S-\lambda I) \neq 0$

$\Leftrightarrow$ $S-\lambda I$의 역행렬 존재

$(S-\lambda I)^{-1}(S-\lambda I)b = 0$

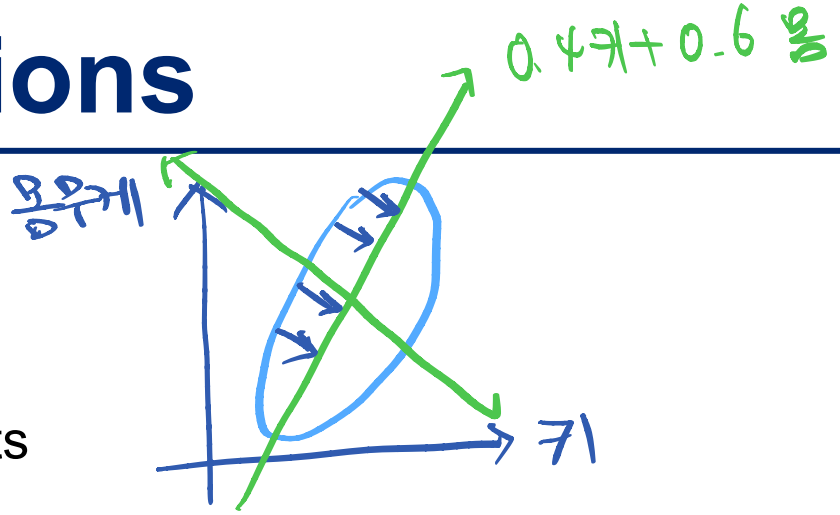
$b = 0$

- Step 4. Projection

$$Z = \mathbf{b}_1^T X_c = [-\sqrt{2} \quad 0 \quad \sqrt{2}]$$

$$\frac{1}{\sqrt{2}}\begin{pmatrix}1 & 1\end{pmatrix}\begin{pmatrix}-1 & 0 & 1\\ -1 & 0 & 1\end{pmatrix}$$

Strengths & Limitations



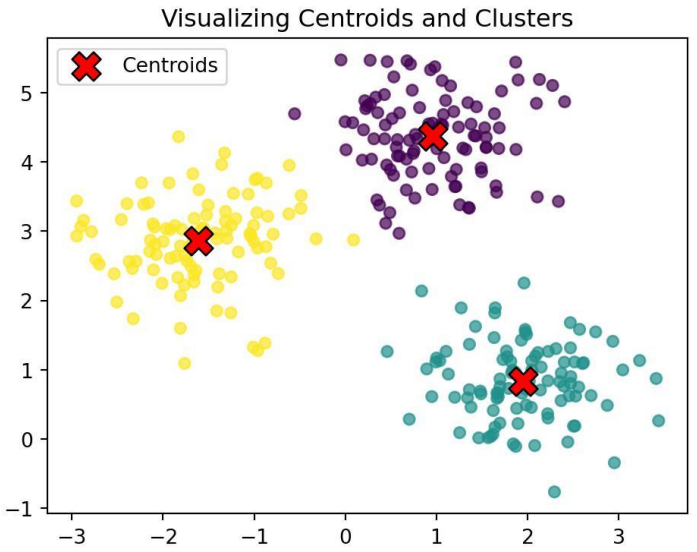
- Strengths
  - Reduces dimensionality
  - Produces uncorrelated components
  - Enables visualization
  - May reduce computational cost
- Limitations
  - Uses linear projections
  - Sensitive to feature scaling
  - Components can be hard to interpret
  - High variance does not guarantee useful information

Korea Aerospace University

Part 3. K-Means Clustering

K-Means Clustering

- Key idea
  - Divides data into K clusters
  - Each cluster is represented by a centroid (mean of its points)
  - Assigns each point to the nearest centroid
- Initialization
  - Choose the number of clusters K
  - Initialize K centroids (e.g., randomly selected data points)



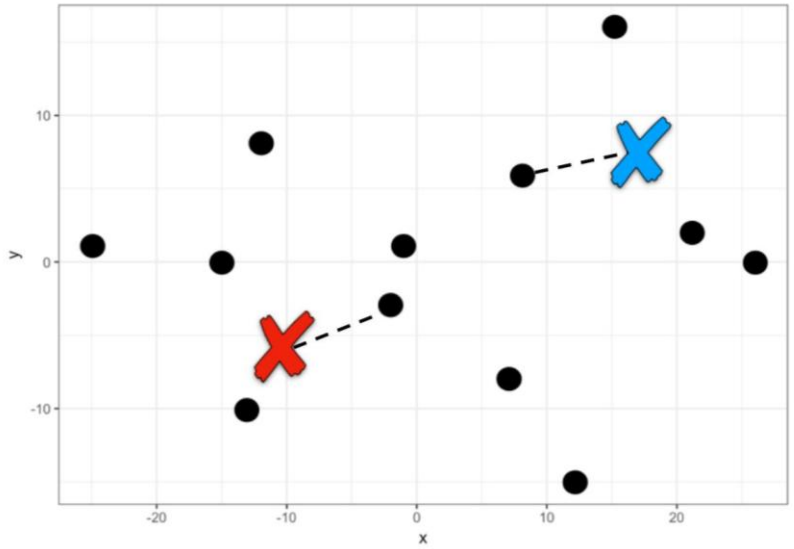
K-Means Objective

- Objective
  - Keep data points close to their assigned centroids
  - Minimize the within-cluster sum of squares (WCSS)

$$J = \sum_{k=1}^{K} \sum_{x_i \in C_k} ||x_i - \mu_k||_2^2$$

$$\sqrt{(5-4)^2+(3-4)^2}^2$$

- C_k: set of points in cluster k
- μ_k: centroid of cluster k

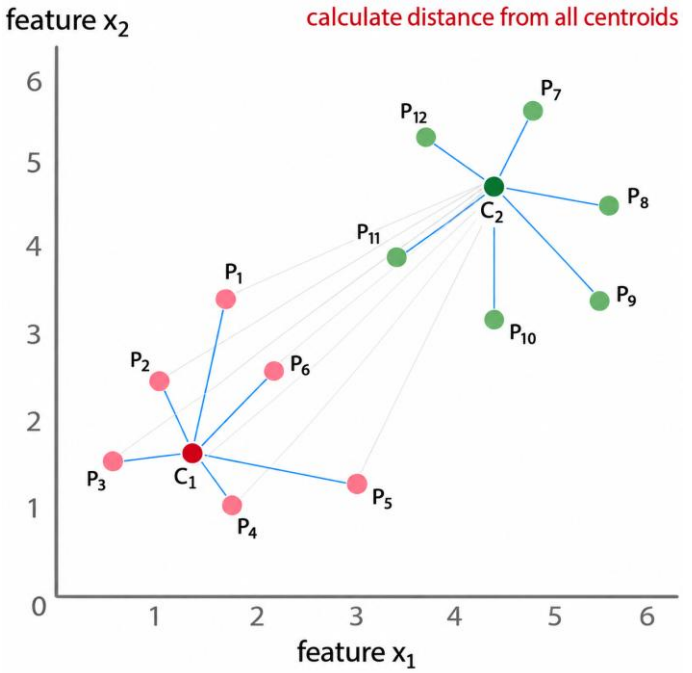


Cluster Assignment

- Step 1: Assign data points
  - Compute the distance from each point to every centroid
  - Assign each point to its nearest centroid

$$c_i = \arg\min_k ||x_i - \mu_k||_2^2$$

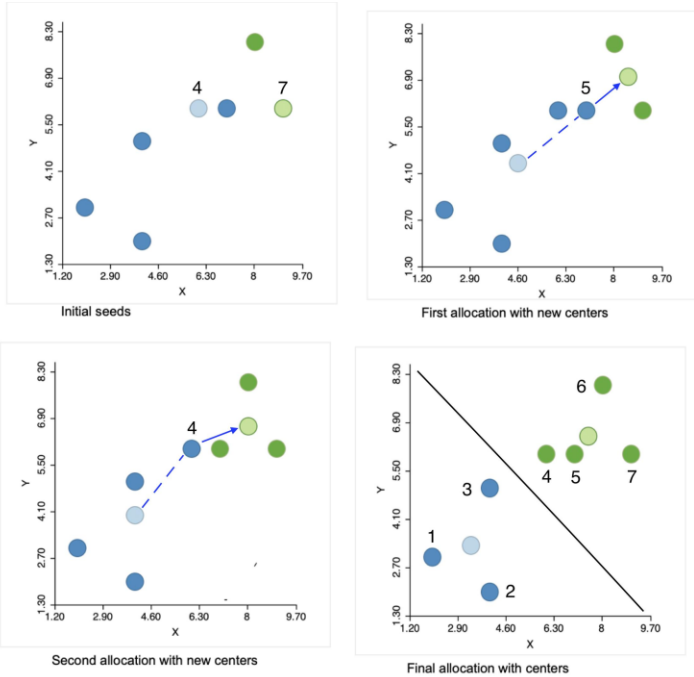
  - c_i: cluster assignment of point x_i



Centroid Update

- Step 2: Update centroids
  - Calculate the mean of all points assigned to each cluster
  - Move each centroid to the mean of its assigned points

$$\mu_k = \frac{1}{|C_k|} \sum_{x_i \in C_k} x_i$$



- Example

$$C_1 = \{2, 4, 6\} \implies \mu_1 = \frac{2 + 4 + 6}{3} = 4$$

Iteration & Convergence

ex. [1, 2, 4, 8, 9] K=2
init. A=1, B=4

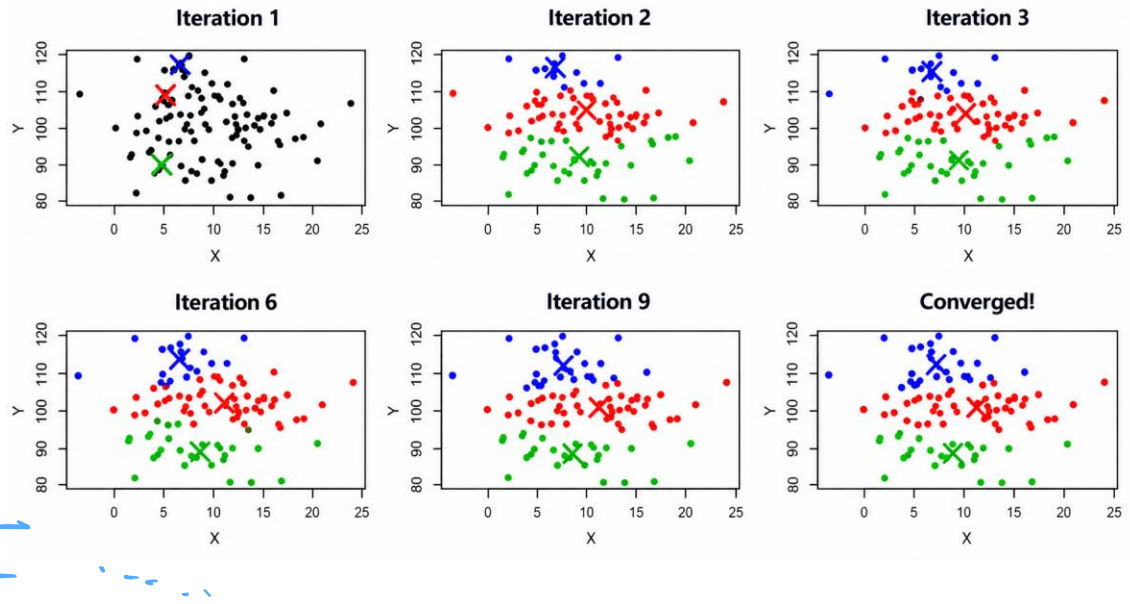
- Iteration
  - Repeat cluster assignment and centroid update
  - WCSS does not increase across iterations

Step 1. cluster A (1) 2
cluster B (4) 8, 9

step 2. $A = \frac{1+2}{2} = 1.5$

$B = \frac{4+8+9}{3} = 7$

- Convergence
  - Stop when cluster assignments stabilize
  - K-Means may converge to a local optimum

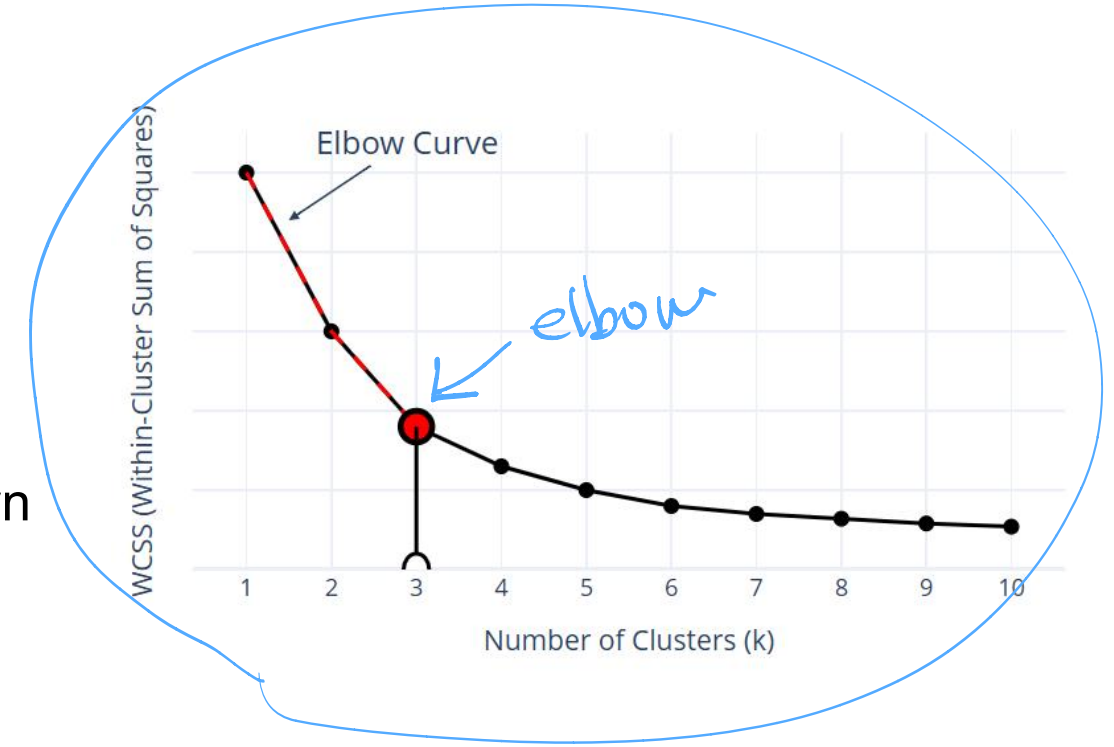


Step 1. cluster A (1.5) 1, 2, 4
cluster B (7) 8, 9

Step 2. $A = \frac{1+2+4}{3} = \frac{7}{3}$ $B = \frac{8+9}{2} = \frac{17}{2}$ ...

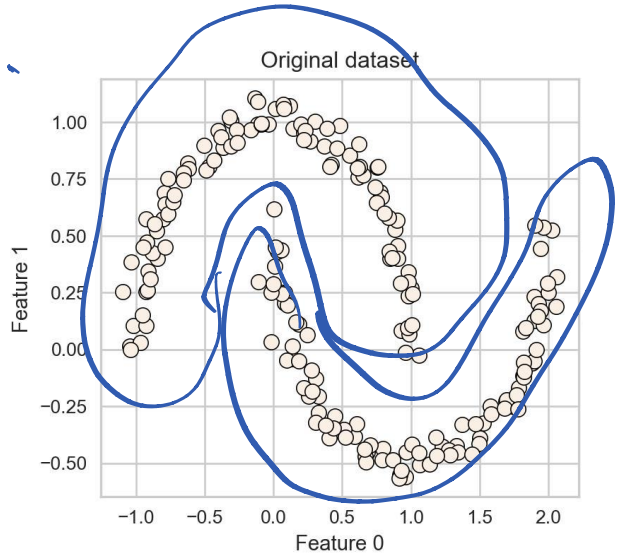
Elbow Method

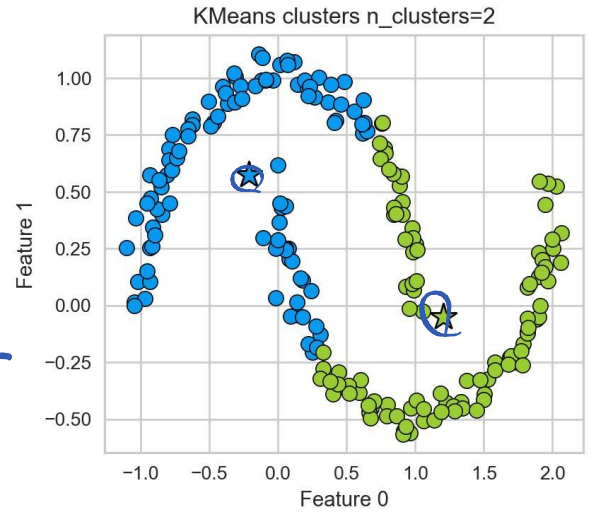
- Choosing K
  - WCSS generally decreases as K increases
- Elbow method
  - Run K-Means for different values of K
  - Plot WCSS against K
  - Find the elbow where improvement slows down
- Limitation
  - The elbow point is not always clear



Strengths & Limitations

- Strengths
  - Simple and computationally efficient
  - Scales well to large datasets
  - Works well for compact, roughly spherical clusters
- Limitations
  - Requires specifying K
  - Sensitive to initialization, outliers, and feature scaling
  - Struggles with complex cluster shapes
- Practical tips
  - Scale features when appropriate
  - Use K-Means++ and multiple initializations





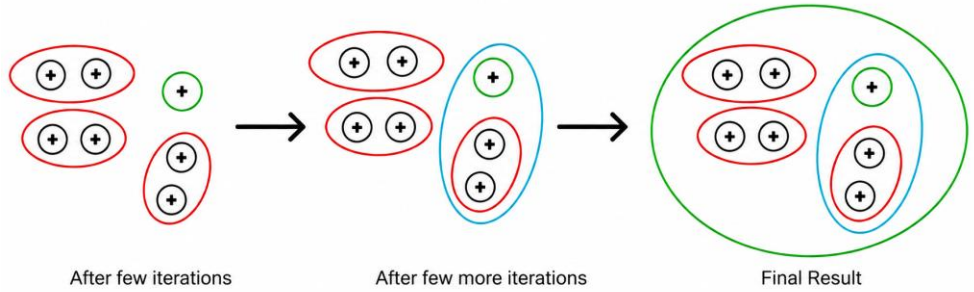
Korea Aerospace University

Part 4. Hierarchical Clustering & Evaluation

Hierarchical Clustering

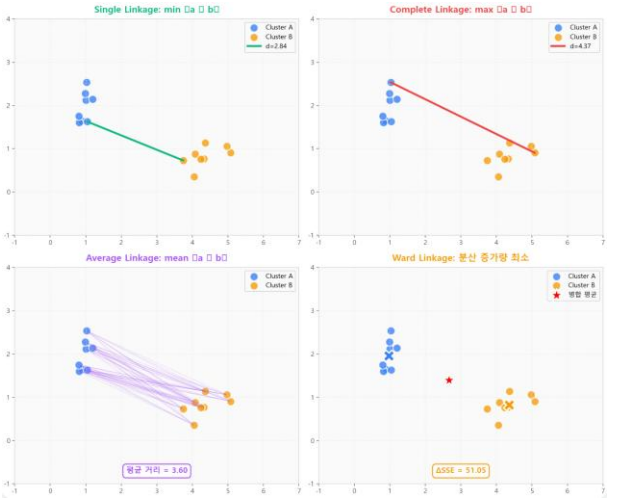
- Key Idea
  - Builds a hierarchy of clusters based on similarity
  - Represents the hierarchy as a tree structure

- Agglomerative Clustering (Bottom-Up)
  - Start with each data point as a separate cluster
  - Merge the two closest clusters
  - Repeat until all points belong to one cluster



Linkage Methods (cluster 끼리의 거리 측정법)

- How do we measure distance between clusters?
  - Single linkage: minimum pairwise distance
  - Complete linkage: maximum pairwise distance
  - Average linkage: average pairwise distance
  - Ward linkage: minimum increase in within-cluster sum of squares



$C_1 = \{1, 3\}$ $C_2 = \{5, 7\}$

$\mu_1 = 2$ $\mu_2 = 6$

WCSS

$WCSS(C_1) = (1-2)^2 + (3-2)^2 = 2$

$WCSS(C_2) = (5-6)^2 + (7-6)^2 = 2$

$WCSS_{bef} = 4$

A, B, C

(A, B), (B, C), (A, C)

- Key point
  - Different linkage methods can produce different clustering results

$C_{12} = \{1, 3, 5, 7\}$

$\mu_{12} = 4$

$WCSS(C_{12})$

$= (1-4)^2 + (3-4)^2 + (5-4)^2 + (7-4)^2 = 20$

$WCSS_{aft} = 20$

WCSS의 차이

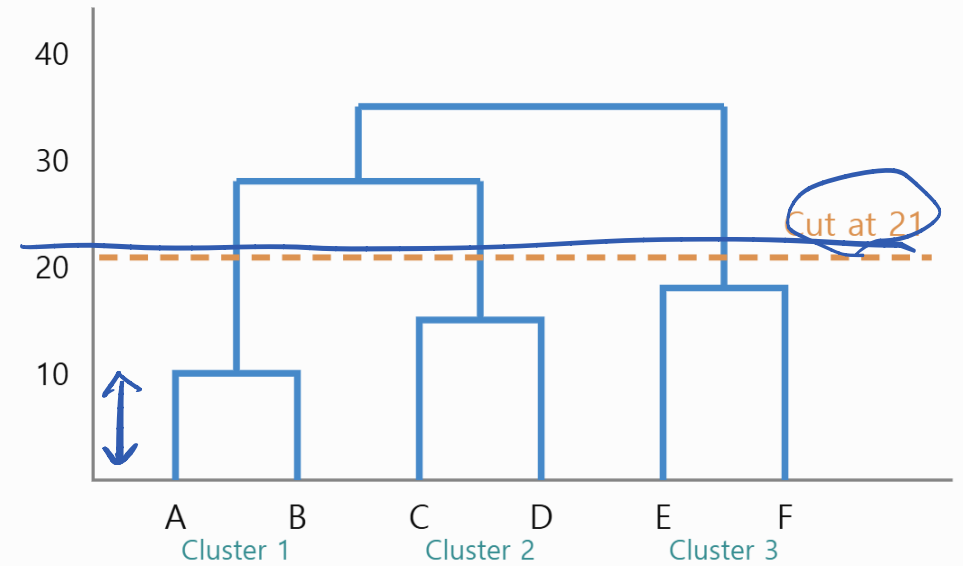
$= 16$

Dendrogram

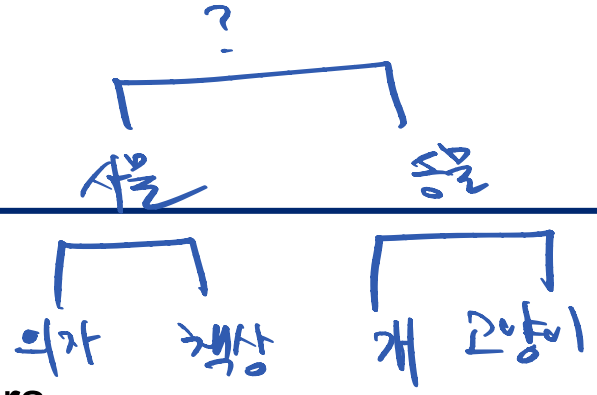
- What is a dendrogram?
  - A tree diagram showing how clusters are merged
  - The height of each merge represents its dissimilarity level

- Choosing the number of clusters
  - Cut the dendrogram at a selected height
  - The resulting branches determine the clusters

- Key idea
  - A lower cut generally produces more clusters
  - A higher cut generally produces fewer clusters



K-Means vs. Hierarchical



- When to use each?
  - K-Means: large datasets with compact, well-separated clusters
  - Hierarchical: smaller datasets where hierarchical structure is useful

	K-Means	Hierarchical
Approach	Centroid-based	Hierarchy-based
Number of Clusters	Specify K in advance	Select from dendrogram
Output	Flat clusters	Cluster hierarchy
Efficiency	Generally faster	Often more expensive
Main limitation	Initialization & cluster shape	Linkage choice & scalability

Silhouette Score

- Evaluating clustering
  - Good clustering aims for high cohesion and separation
- Silhouette Score

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

  - $a(i)$: average distance to points in the same cluster
  - $b(i)$: minimum average distance to points in another cluster
  - Overall score: average of $s(i)$ across all data points



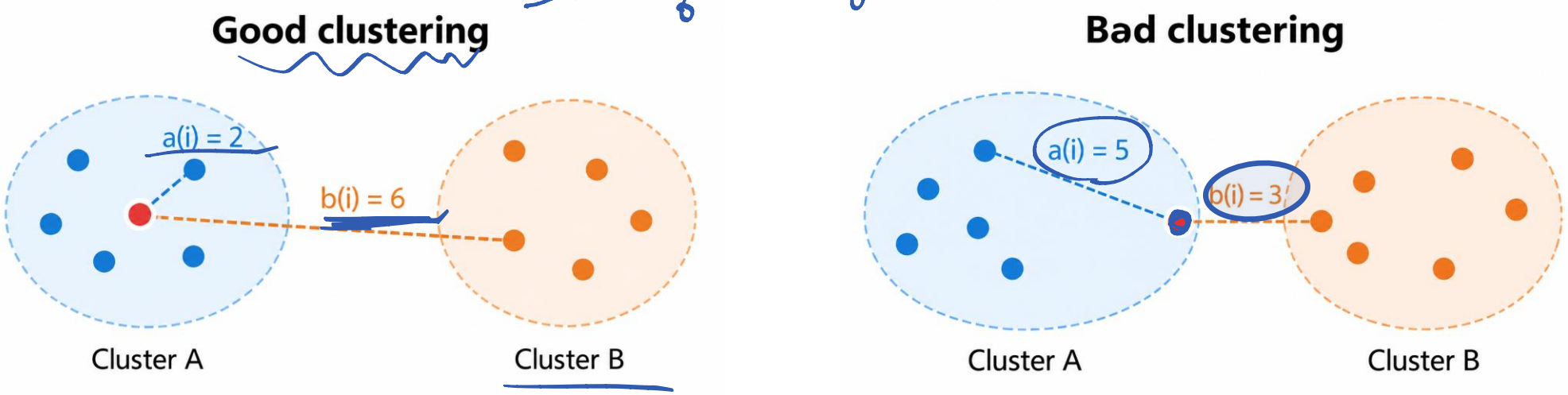
Silhouette Score

- Interpretation
  - Close to +1: well clustered
  - Around 0: near a cluster boundary
  - Close to -1: closer to another cluster

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))}$$

$\frac{6-2}{6} = \frac{4}{6}$

$\frac{3-5}{5} = -\frac{2}{5}$



Next Lecture

- Midterm Exam
  - When: Oct. 21, 2026 (Wed), 14:30−16:30
    - No entry after 14:30
    - 14:30−14:40: submit all belongings except writing utensils
    - 14:40−16:30: exam
  - Where: Electronics Building, Room 422
  - Coverage: Lec01−Lec07

Korea Aerospace University

Questions?

jiyoonshin@kau.ac.kr