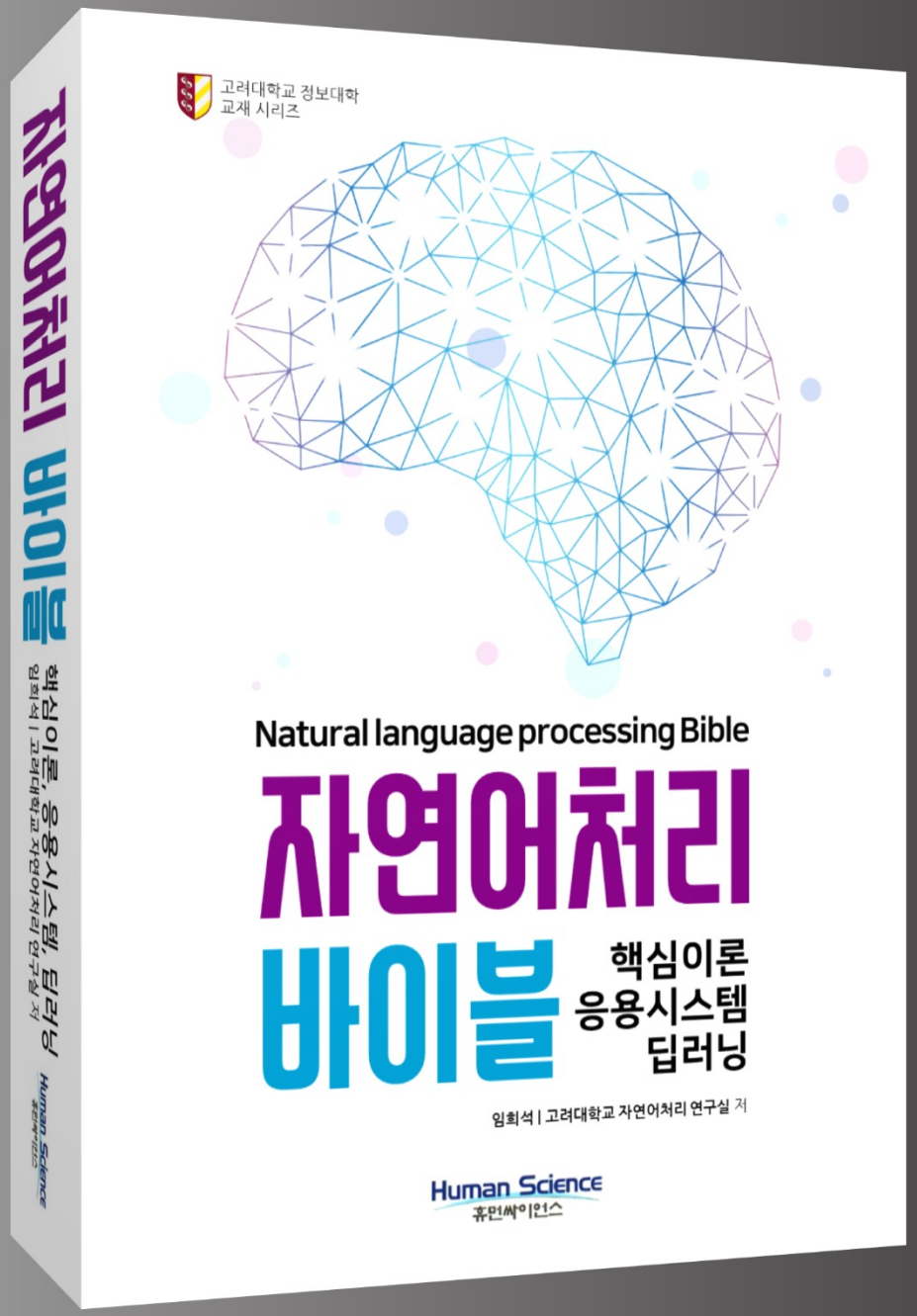


CHAPTER 16
텍스트 분류
(Text Categorization)

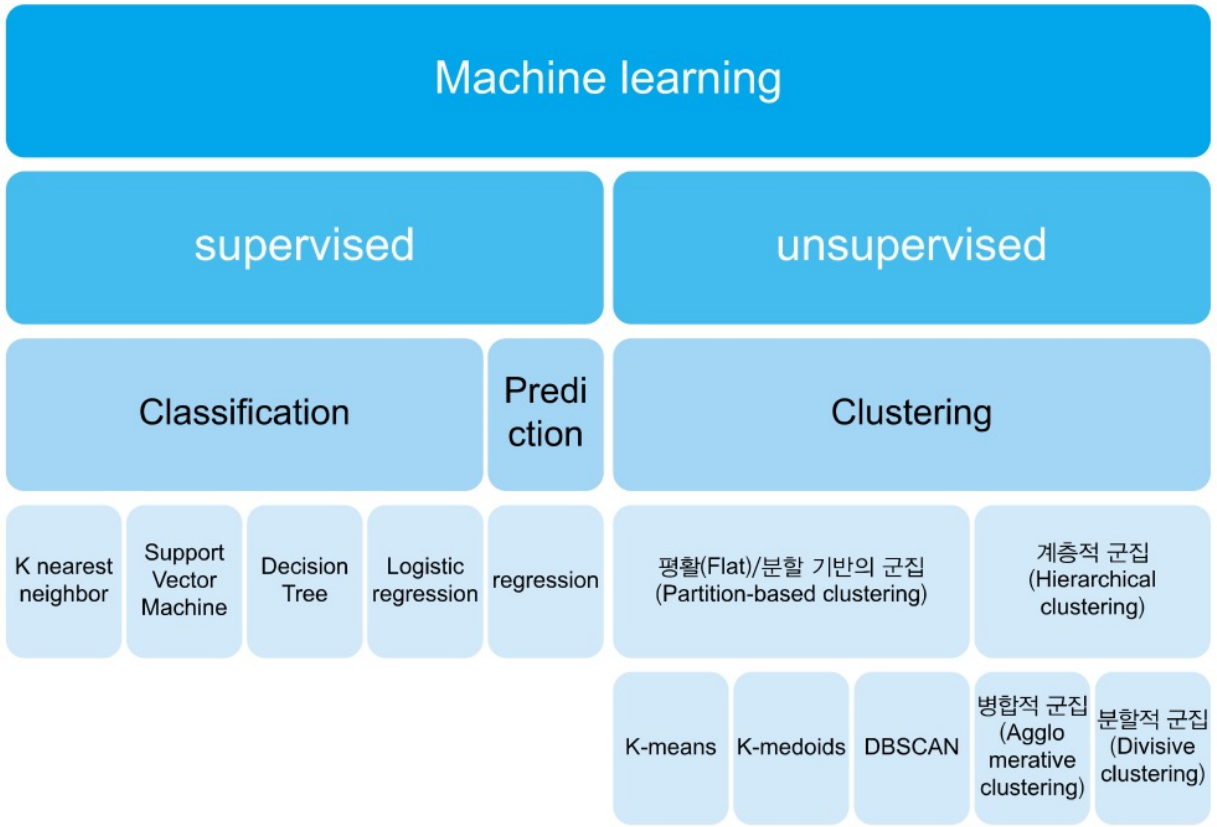
목차

16.1 텍스트 분류란?

- 텍스트 분류
  - 문장 또는 문서를 입력으로 받아 사전에 정의된 클래스 중에 어디에 속하는지 분류(Classification)하거나 각 데이터를 군집화(Clustering)하는 과정

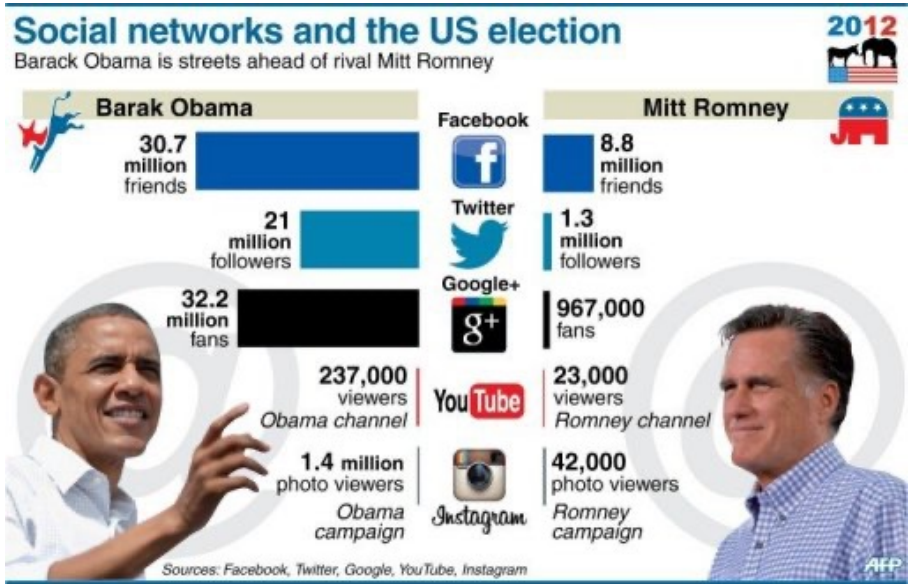
- 분류(Classification)
  - 지도학습에 속하며, 자료를 자동으로 항목에 맞게 범주화하는 작업

- 군집화(Categorization)
  - 비지도학습에 속하며, 정해진 항목이 아닌 항목들 간의 유사 관계에 의하여 스스로 분류



16.2 일상 속 텍스트 분류

- 예1 : 고객의 의견분석; 감성분석(sentiment analysis) : 긍정? 부정?
- 예2 : 고객의 민원의 종류 분류
- 예3: 오바마 대통령 선거에서 빅데이터 감정분석
  - 정치헌금 기부명단, 각종 면허, 신용카드 정보, 소셜 네트워크 서비스(SNS) 등 다양한 빅데이터의 분석을 통해 유권자 개인별 맞춤형 선거 운동을 전개



16.3 감정분석이란 무엇인가

- 감정분석
  - 문장 또는 지문의 감정을 분석하는 것을 의미하며 자연어처리의 하나의 큰 분야
  - 예) 영화리뷰 감정분석
  - 규칙 기반 모델이나, 확률 모델, 딥러닝 모델

- Sentiment Analysis vs Emotion Analysis

WordNet	SentiWordNet	Opinion Lexicon	MPQA
• *Score*: **3.977** • (pos > 0, neg < 0)	• *Positive*: **0.75** • *Negative*: **0.0**	*Polarity*: **positive**	• *Polarity*: **positive** • *Strength*: **weaksubj**

- "awkward"

WordNet	SentiWordNet	Opinion Lexicon	MPQA
• *Score*: **0.0** • (pos > 0, neg < 0)	• *Positive*: **0.125** • *Negative*: **0.5**	*Polarity*: **negative**	• *Polarity*: **negative** • *Strength*: **strongsubj**

그림 16-4 감정분석 사전 종류

- *Movie*: is this review positive or negative?
- *Products*: what do people think about the new iPhone?
- *Public sentiment*: how is consumer confidence? Is despair increasing?
- *Politics*: what do people think about this candidate or issue?
- *Prediction*: predict election outcomes or market trends from sentiment

그림 16-5 일상에서 감정분석

16.4 다양한 텍스트 분류 예시

- 16.4.1 카테고리 및 의도 분류
  - 카테고리 분류: 문장이나 지문을 카테고리로 분류하는 것
  - 의도분류: 의도를 분류하는 것

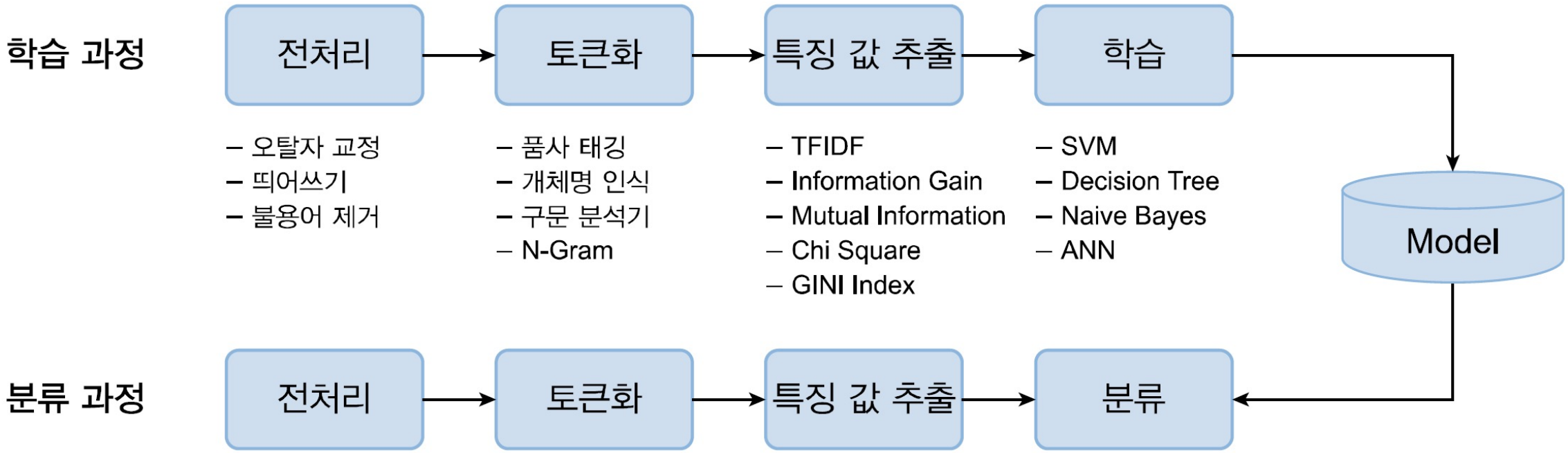
- 16.4.2 스팸 메일 분류
  - 메일이 왔을 때 해당 내용이 스팸 메일인지 햄 분류

| 표 16-1 | 스팸, 햄 메일 분류 예시

텍스트 (메일의 내용)	레이블 (스팸 여부)
어제 보내드린 보고서 확인 부탁드립니다.	정상 메일
마지막 혜택, 놓치지 마세요!	스팸 메일
답장 가능하세요?	정상 메일
(광고) 당신의 스타일을 찾아드립니다.	스팸 메일

16.5 텍스트 분류 프로세스

- 텍스트 분류 프로세스



16.5 텍스트 분류 프로세스

1. 전처리
   - 오탈자 교정, 띄어쓰기 교정, 불용어 제거...

2. 토큰화
   - 품사태깅, 개체명인식, 구문 분석기, N-Gram, Normalization...

3. 특징값 추출: Bag of Words, TF-IDF...
   - Bag of Words:
     - 단어들의 순서는 전혀 고려하지 않고,
       단어들의 출현 빈도(frequency)에만 집중하는 텍스트 데이터의 수치 화 표현 방법
   - TF-IDF(Term Frequency-Inverse Document Frequency)
     - 단어의 빈도와 역문서 빈도를 사용하여 DTM 내의 각 단어들 마다 중요한 정도를 가중치로 주는 방법

4. 학습 진행
   - 다양한 텍스트 분류 알고리즘을 기반으로 학습을 진행. 마지막으로 해당 모델에 대한 평가를 진행

16.6 텍스트 분류

- 대표적인 분류(classification) 알고리즘
  - Probabilistic or Machine Learning
    - Bayesian, SVM, Random Forest, NN
  - Geometric
    - kNN

Naïve Bayes Classifier

Naïve Bayes Classifier

- 문서 d의 분류는 다음 식으로 계산 :

$$\begin{aligned} \text{Class}(d) &= \arg\max_{c \in \mathcal{C}} P(c|d) \\ &= \arg\max_{c \in \mathcal{C}} \frac{P(d|c)P(c)}{\sum_{c \in C} P(d|c)P(c)} \end{aligned}$$

- Must estimate P(*d* | *c*) and P(*c*)
  - P(*c*) : 클래스 *c* 가 관측될 확률
  - P(*d* | *c*) : 클래스 *c*가 주어졌을 때 문서 *d*가 관측될 확률

P(c) 추정하기

- P(*c*) is the probability of observing class *c*
- Estimated as the proportion of training documents in class *c*:

$$P(c) = \frac{N_c}{N}$$

- N_c is the number of training documents in class *c*
- *N* is the total number of training documents

P(*d* | *c*) 추정하기

- P(*d* | *c*) is the probability that document *d* is observed given the class is known to be *c*
- Estimate depends on the *event space* used to represent the documents
- What is an event space?
  - The set of all possible outcomes for a given random variable
  - For a coin toss random variable the event space is *S* = {heads, tails}

Multiple Bernoulli Event Space

- Documents are represented as binary vectors
  - One entry for every word in the vocabulary
  - Entry *i* = 1 if word *i* occurs in the document and is 0 otherwise
- Multiple Bernoulli distribution is a natural way to model distributions over binary vectors
- Same event space as used in the classical probabilistic retrieval model

Multiple Bernoulli Document Representation

document *id*	cheap	buy	banking	dinner	the	*class*
1	0	0	0	0	1	not spam
2	1	0	1	0	1	spam
3	0	0	0	0	1	not spam
4	1	0	1	0	1	spam
5	1	1	0	0	1	spam
6	0	0	1	0	1	not spam
7	0	1	1	0	1	not spam
8	0	0	0	0	1	not spam
9	0	0	0	0	1	not spam
10	1	1	0	1	1	not spam

Multinomial: Estimating P(*d* | *c*)

- P(*d* | *c*) is computed as:

$$P(d|c) = \prod_{w \in \mathcal{V}} P(w|c)^{\delta(w,d)} \left(1 - P(w|c)\right)^{1-\delta(w,d)}$$

- Laplacian smoothed estimate:

$$P(w|c) = \frac{tf_{w,c} + 1}{|c| + |\mathcal{V}|}$$

- Collection smoothed estimate:

$$P(w|c) = \frac{tf_{w,c} + \mu \frac{cf_w}{|C|}}{|c| + \mu}$$

Support Vector Machine

16.6 SVM 개요

- 알고리즘은 초평면(hyperplane) 의 법선 벡터(normal vector) ‘w’와 편향 값(bias) ‘b’로 표현되는 분류기(classifier)를 탐색
- 초평면(경계)은 마진을 최대화하는 지점을 찾는 것임

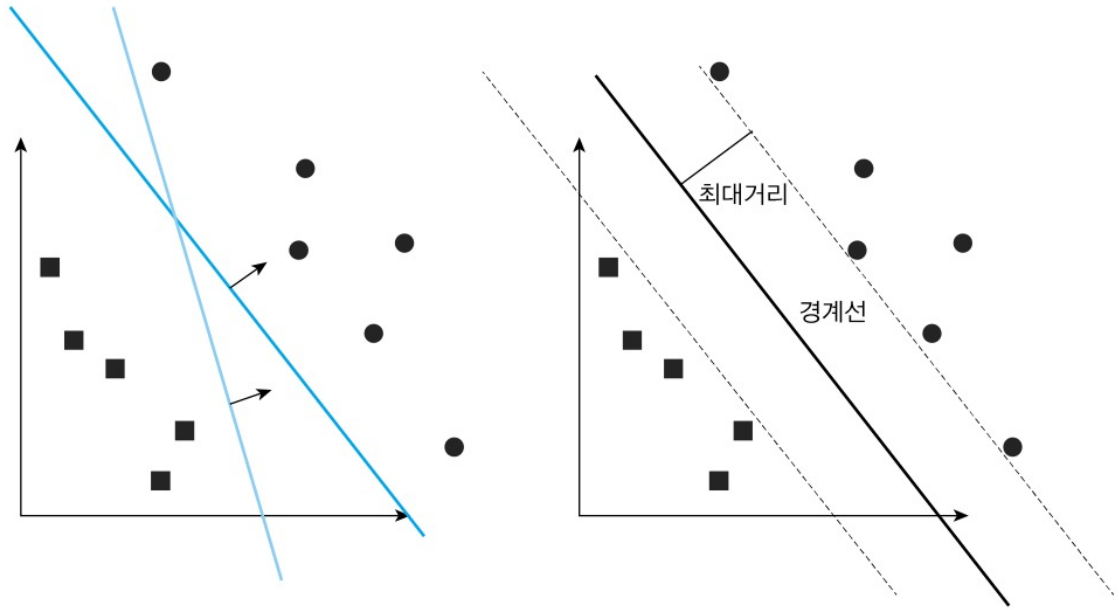
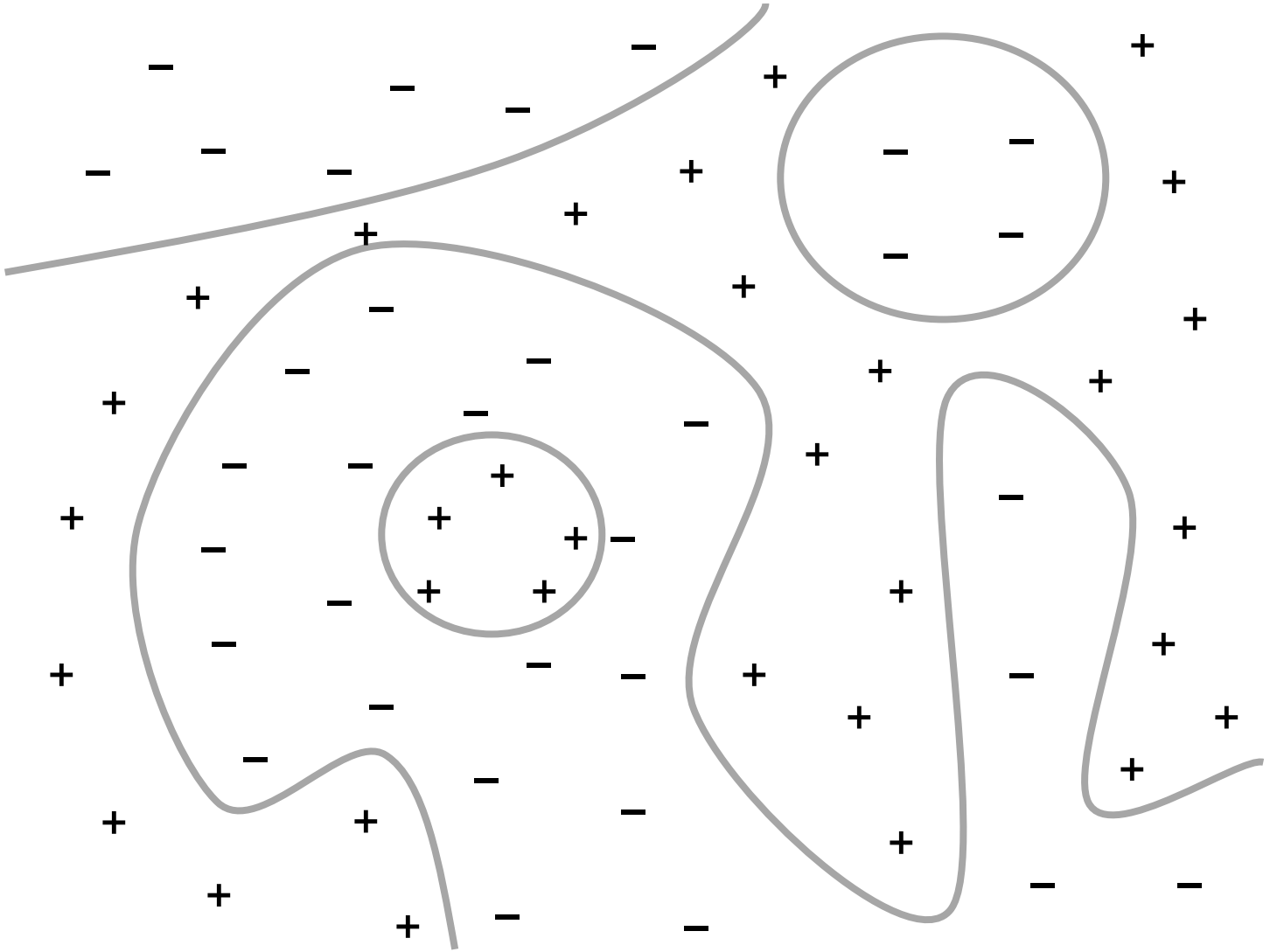


그림 16-9 SVM

16.6 Geometric 분류

- KNN(K-nearest neighbors)
  - 분류 및 회귀 예측 문 제 모두에 사용할 수 있는 supervised 머신러닝 알고리즘 유형 주로 산업의 분류 예측 문제에 사용된다.
  - KNN은 기본 데 이터에 대해 아무 것도 가정하지 않기 때문에 비모수적 학습 알고리즘
- Decision Tree
  - 의사결정 규칙과 그 결과들을 트리구조로 찾는 것이며 데이터 마이닝 분야에서 주로 사용
  - Stochastic Gradient Descent: 데이터에 대한 매개변수를 평가하고 값을 조정하면서 손실함수(Loss Function)를 최소화하는 값을 구하는 방법이다.
- The Random Forest Algorithm
  - 여러 개의 결정트리들을 임의적으로 학습하는 일종의 앙상블 학습방법이며 훈련과정에서 구성한 다수의 결정 트리로부터 분류 또는 평균 예측치(회귀 분석)을 출력함으로써 동작

Nearest Neighbor Classification



Clustering

Clustering

- 군집화
  - 주어진 입력값들을 유사한 특성을 가진 군집들로 생성하는 작업
  - Unsupervised learning
  - 비계층적 군집화와 계층적 군집화
- 군집화 기술의 응용 예
  - 고객 성향 분석 및 맞춤형 추천

- General outline of clustering algorithms
  1. Decide how items will be represented (e.g., feature vectors)
  2. Define similarity measure between pairs or groups of items (e.g., cosine similarity)
  3. Determine what makes a "good" clustering
  4. Iteratively construct clusters that are increasingly "good"
  5. Stop after a local/global optimum clustering is found
- Steps 3 and 4 differ the most across algorithms

16.6 Clustering (군집화)

- k-평균 알고리즘(K-means algorithm)
  - 주어진 데이터를 k 개의 클러스터로 묶는 알고리즘으로, 각 클러스터와 거리 차이의 분산을 최소화하는 방식으로 동작
  - k-평균 클러스터링 알고리즘은 클러스터링 방법 중 분할법에 속함

$$V = \sum_{i=1}^{k} \sum_{j \in S_i} |x_j - \mu_i|^2$$

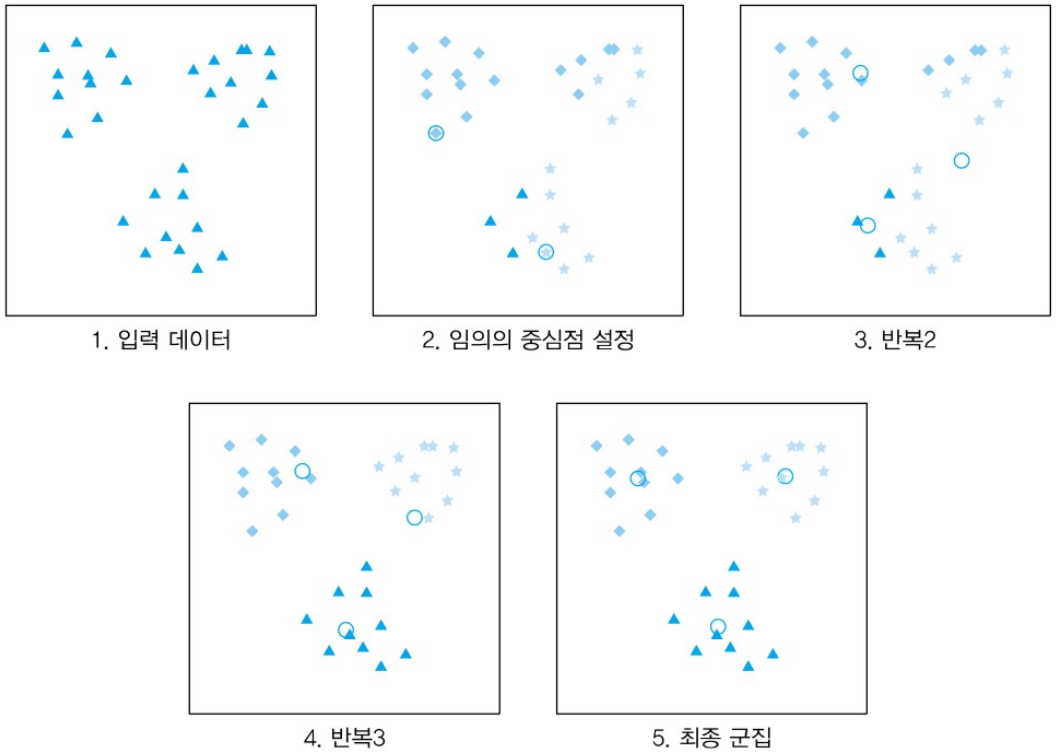
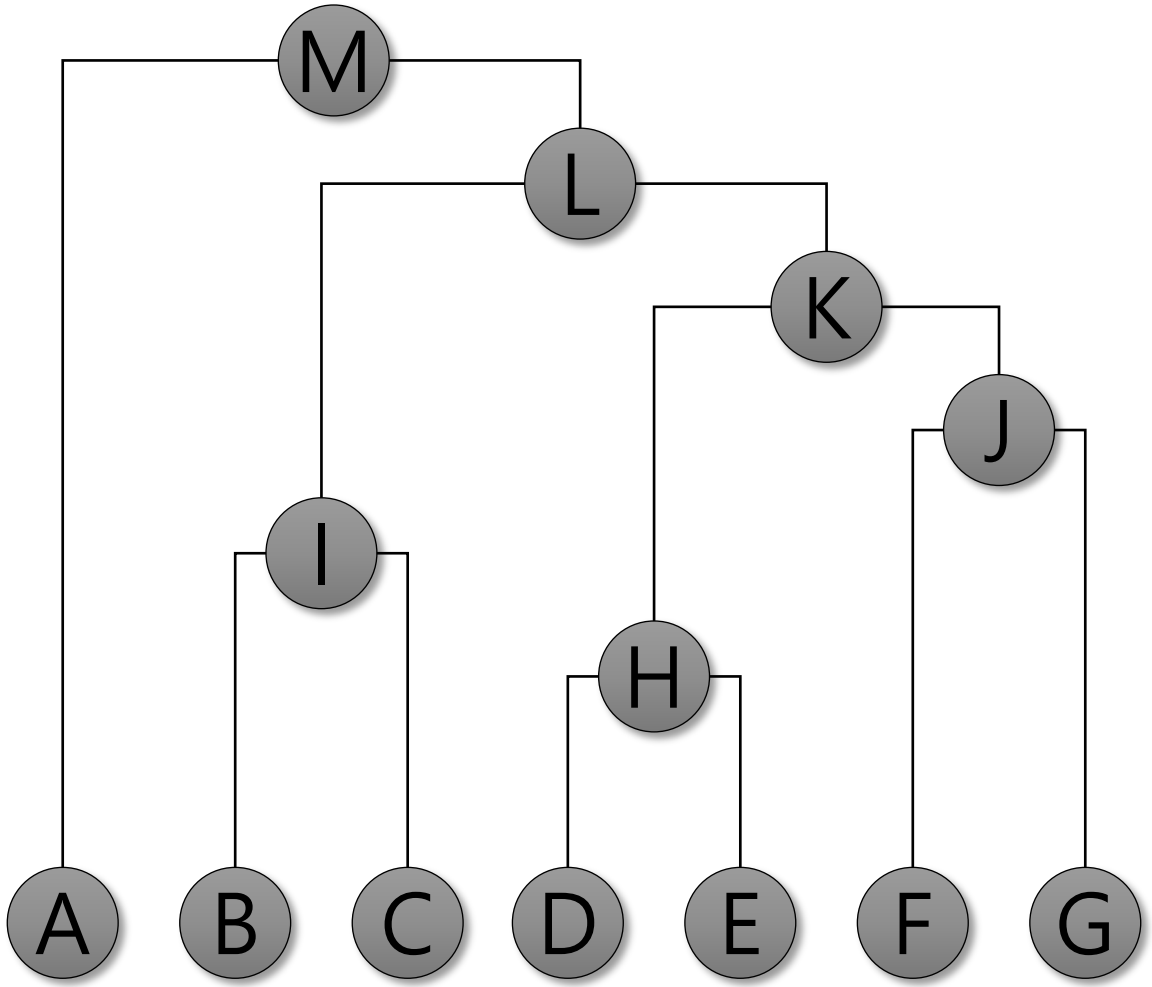


그림 16-10 K-means 알고리즘

계층적 Clustering

- Constructs a hierarchy of clusters
  - The top level of the hierarchy consists of a single cluster with all items in it
  - The bottom level of the hierarchy consists of *N* (# items) singleton clusters
- Two types of hierarchical clustering
  - Divisive (“top down”)
  - Agglomerative (“bottom up”)
- Hierarchy can be visualized as a *dendogram*

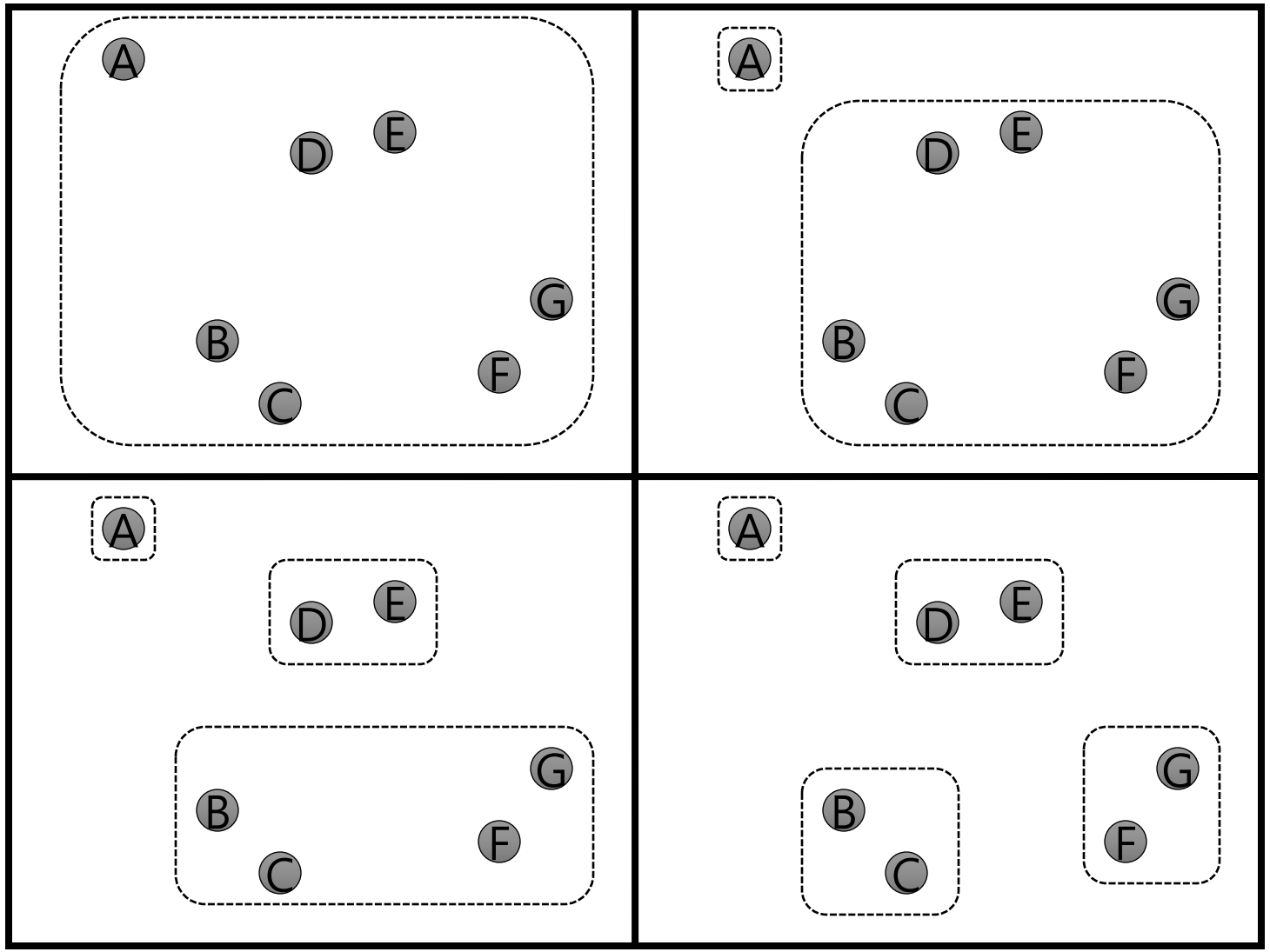
Dendrogram 예



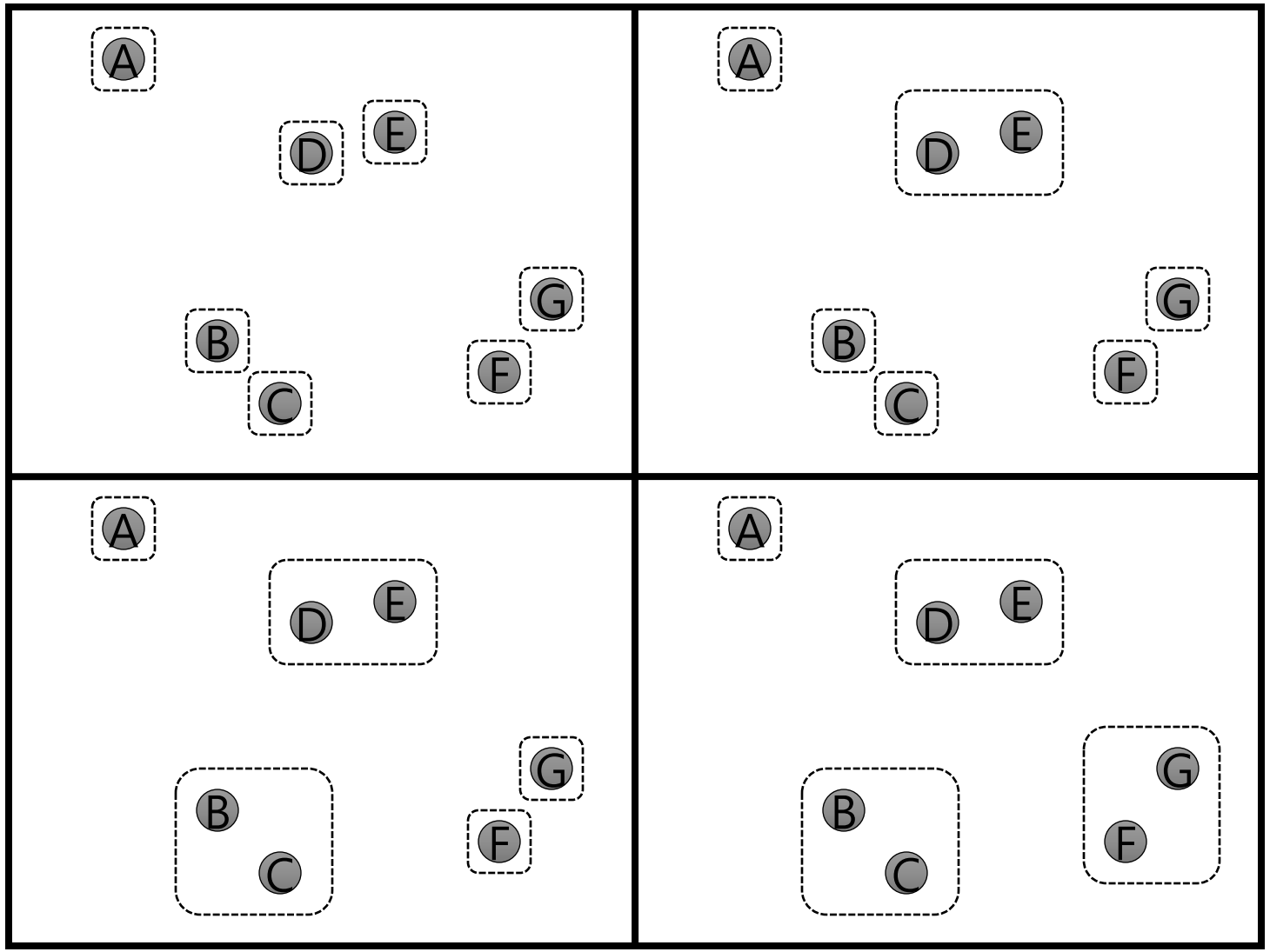
Divisive and Agglomerative Hierarchical Clustering

- Divisive
  - Start with a single cluster consisting of all of the items
  - Until only singleton clusters exist…
    - **Divide** an existing cluster into two new clusters
- Agglomerative
  - Start with N (# items) singleton clusters
  - Until a single cluster exists…
    - **Combine** two existing cluster into a new cluster
- How do we know how to divide or combined clusters?
  - Define a division or combination cost
  - Perform the division or combination with the lowest cost

Divisive Hierarchical Clustering



Agglomerative Hierarchical Clustering



Clustering Costs

- Single linkage

$$COST(C_i, C_j) = \min\{dist(X_i, X_j) | X_i \in C_i, X_j \in C_j\}$$

- Complete linkage

$$COST(C_i, C_j) = \max\{dist(X_i, X_j) | X_i \in C_i, X_j \in C_j\}$$

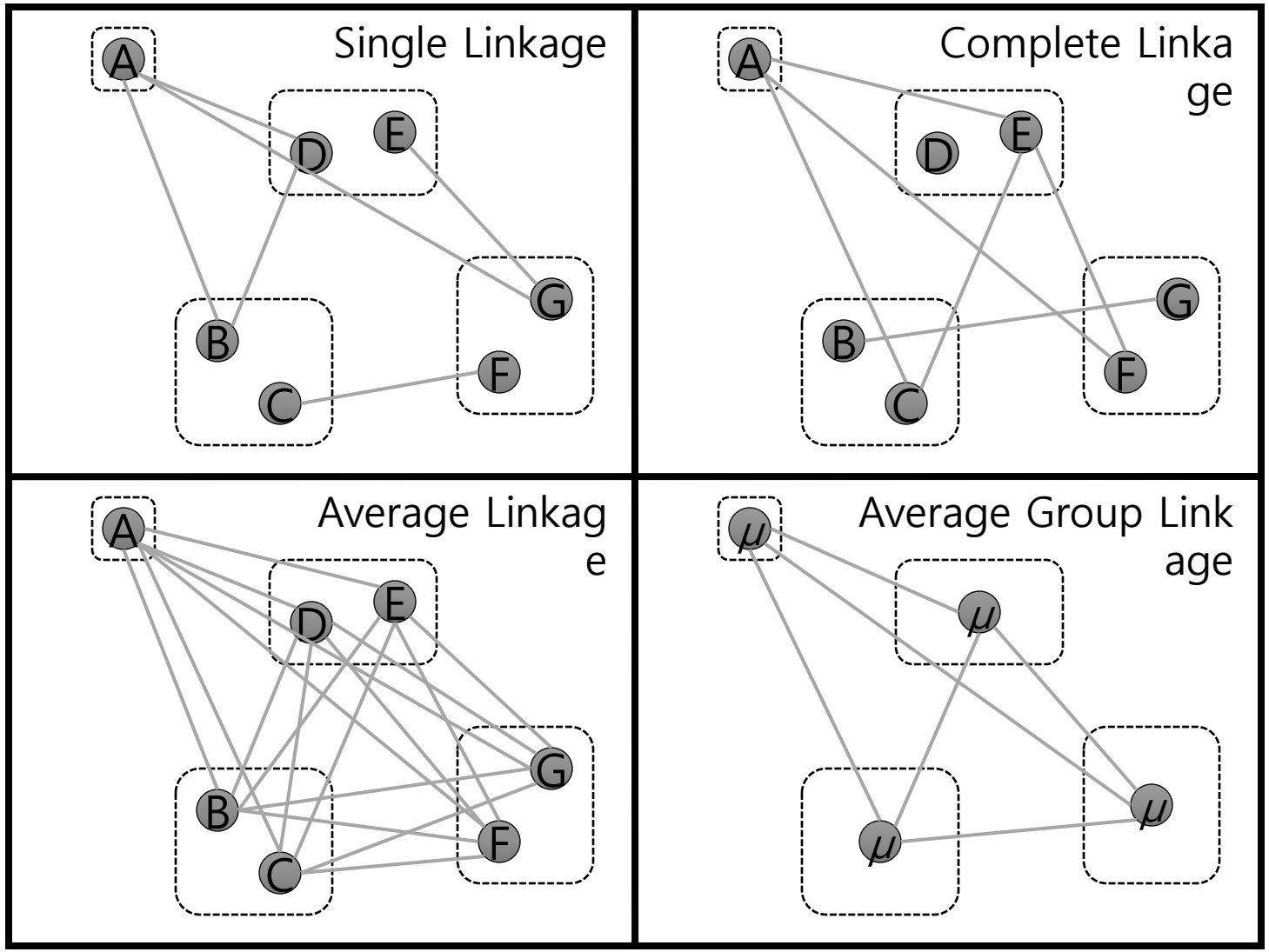
- Average linkage

$$COST(C_i, C_j) = \frac{\sum_{X_i \in C_i, X_j \in C_j} dist(X_i, X_j)}{|C_i||C_j|}$$

- Average group linkage

$$COST(C_i, C_j) = dist(\mu_{C_i}, \mu_{C_j})$$

Clustering Strategies



Evaluation : Micro- vs. Macro-Averaging

- If we have more than one class, how do we combine multiple performance measures into one quantity?
- Macroaveraging: Compute performance for each class, then average.
- Microaveraging: Collect decisions for all classes, compute contingency table, evaluate.

Micro- vs. Macro-Averaging: Example

Class 1

	Truth: yes	Truth: no
Classifier : yes	10	10
Classifier : no	10	970

Class 2

	Truth: yes	Truth: no
Classifier: yes	90	10
Classifier: no	10	890

Micro Ave. Table

	Truth: yes	Truth: no
Classifier: yes	100	20
Classifier: no	20	1860

- Macroaveraged precision: (0.5 + 0.9)/2 = 0.7
- Microaveraged precision: 100/120 = .83

- Microaveraged score is dominated by score on common classes