

01

# 데이터마이닝의 이해



# 1. 데이터마이닝의 이해



## 가. 데이터마이닝의 개념

대용량의 데이터로부터 자동 또는  
반자동적인 방법을  
통하여 의미 있는 **패턴, 규칙, 관계**를 찾음



데이터를 분석하여 기업에 필요한 자산으로 만드는 정보기술

# 1. 데이터마이닝의 이해



## 나. 데이터마이닝의 특징

- 기업들의 데이터베이스의 필요성이 증가
- 인터넷과 같은 정보기술의 성장 및 기술 발전에 따름
- 영역 전문가가 간과해 버릴 수도 있는 지식과 패턴을 찾음
- 데이터마이닝은 사용자의 경험이나 편견을 배제하고, 전적으로 데이터에 기반하여 지식과 패턴을 추출함

# 1. 데이터마이닝의 이해



## 나. 데이터마이닝의 특징

- 기업들의 데이터베이스의 필요성이 증가
- 인터넷과 같은 정보기술의 성장 및 기술 발전에 따름
- 영역 전문가가 간과해 버릴 수도 있는 지식과 패턴을 찾음
- 데이터마이닝은 사용자의 경험이나 편견을 배제하고, 전적으로 데이터에 기반하여 지식과 패턴을 추출함

### ✓ 데이터마이닝의 다양한 활용분야

- ≫ 카드사의 사기 발견
- ≫ 금융권의 대출 승인
- ≫ 투자 분석
- ≫ 기업의 마케팅 및 판매데이터 분석
- ≫ 생산 프로세스 분석
- ≫ 기타 순수 과학 분야의 자료 분석

# 1. 데이터마이닝의 이해



## 다. 데이터마이닝의 중요성

- 기업은 업무의 효율적인 수행을 위해 데이터베이스를 단순히 활용하는 단계를 벗어남
  - » 데이터 자체의 분석을 통해 고객행동패턴을 추출해내고 그 결과를 업무와 생산의 효율성 증대에 활용

# 1. 데이터마이닝의 이해



## 다. 데이터마이닝의 중요성

- 기업은 업무의 효율적인 수행을 위해 데이터베이스를 단순히 활용하는 단계를 벗어남
  - ≫ 데이터 자체의 분석을 통해 고객행동패턴을 추출해내고 그 결과를 업무와 생산의 효율성 증대에 활용
- 기업은 업무의 효율적인 수행을 위해 데이터베이스를 단순히 활용하는 단계를 벗어남
  - ≫ 정보기술을 기반으로 고객의 다양한 정보를 획득함과 동시에 고객과의 밀접한 관계를 유지함으로써 기업의 수익성을 증대에 기여

# 1. 데이터마이닝의 이해



## 다. 데이터마이닝의 중요성

- 기업은 업무의 효율적인 수행을 위해 데이터베이스를 단순히 활용하는 단계를 벗어남
  - ≫ 데이터 자체의 분석을 통해 고객행동패턴을 추출해내고 그 결과를 업무와 생산의 효율성 증대에 활용
- 기업은 업무의 효율적인 수행을 위해 데이터베이스를 단순히 활용하는 단계를 벗어남
  - ≫ 정보기술을 기반으로 고객의 다양한 정보를 획득함과 동시에 고객과의 밀접한 관계를 유지함으로써 기업의 수익성을 증대에 기여
- 제품을 구매한 기존 고객의 정보를 기반으로 고객에게 맞는 새로운 제품이나 서비스를 제안하기 위함
  - ≫ 데이터마이닝을 이용하여 고객의 구매 패턴을 파악하고 의도를 예측하는 것은 오늘날 실질적인 판매 전략을 수립하는 마케팅 분야에서 상당히 큰 비중을 차지

# 1. 데이터마이닝의 이해



## 라. 데이터마이닝 기법

- 데이터마이닝은 학문적으로 데이터 분석(통계, 전산, 경영 등)과 관련된 다양한 학문이 융합되어 탄생된 융합학문이라고 평가



# 1. 데이터마이닝의 이해



## 라. 데이터마이닝 기법

- 데이터마이닝은 학문적으로 데이터 분석(통계, 전산, 경영 등)과 관련된 다양한 학문이 융합되어 탄생된 **융합학문이라고 평가**

### 정형 데이터 분석

- » 연관관계분석 기법
- » 의사결정나무 기법
- » 인공신경망 기법
- » 사례기반추론
- » 군집분석 기법

### 비정형 데이터 분석

- » 웹 문서
- » 소셜 데이터를 주로 분석하는 텍스트 마이닝
- » 웹 마이닝
- » 오피니언 마이닝
- » 소셜 네트워크 분석

- 데이터를 시각화해서 보여주는 **데이터 시각화 기법** 등이 있음

# 1. 데이터마이닝의 이해



## 마. 데이터마이닝 분석도구/프로그램

- 범용 통계 분석 도구인 R
- SAS사에서 제공하는 Enterprise Miner
- SPSS사에서 제공하는 Clementine
- Weka
- Rapid Miner
- 텍스트 마이닝에 쓰이는 Python
- 다양한 데이터 시각화 프로그램들 (Inforgraphics, Google Chart API, Flot, D3, Processing 등)

02

# 연관관계분석

BIG

DATA

## 2. 연관관계분석



### 가. 연관관계분석 개념

상품 혹은 서비스간의 관계를 살펴보고  
이로부터 유용한 규칙을 찾아내고자 할 때  
이용될 수 있는 기법



동시 구매될 가능성이 큰 상품들을 찾아내는 기법으로  
시장 바구니 분석과 관련된 문제에 많이 적용

## 2. 연관관계분석



### 가. 연관관계분석 개념

- 측정의 기본

- » ‘얼마나 자주 구매되었는가’ 라는 빈도를 기본

- 연관성 규칙의 기본적인 개념

- » 시장바구니 품목들을 식별하는 것에서부터 시작

- 사건 또는 품목 간에 일어나는 연관성을 규명하려는 것이  
연관성 규칙

- » 연관 규칙은 “A라는 어떠한 사건이 일어나면 B라는 다른 사건이  
일어난다.”와 같이 표현함

- » 연관 정도를 정량화 하기 위한 기준 : 지지도, 신뢰도, 향상도를 계산

## 2. 연관관계분석



### 나. 연관관계분석의 특징

- 대량의 데이터로부터 품목간의 어떠한 종속 관계가 존재하는지를 찾아내는 작업
- 연관성 규칙을 통해 요소간의 연관성 패턴 분석
- 연관성 규칙은 데이터 마이닝 기법으로 장바구니 분석을 통한 상품추천이나 상품 진열 등에 사용
- 연관성 규칙은 상품 또는 서비스 간의 관계를 살펴봄으로써 그들 간의 유용한 관계가 존재하는지 파악
- 구체적인 행위를 언급하여 규칙을 도출하기 때문에 이해하기 쉽고 명쾌한 특성을 가짐

## 2. 연관관계분석



### 다. 연관관계분석 기법

데이터들의 빈도수와 동시 발생 확률을 이용하여 한 항목들의 그룹과 다른 항목들의 그룹 사이에 강한 연관성이 있음을 밝혀주는 기술

#### ④ 연관성 규칙의 예

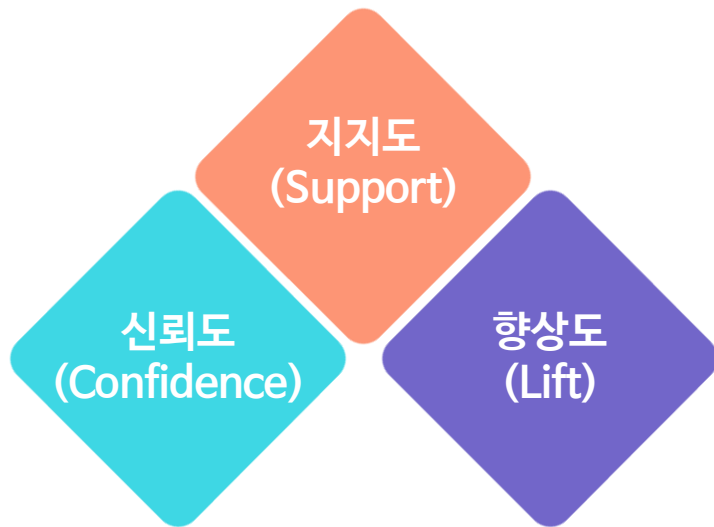
$(\text{Item set } X) \rightarrow (\text{Item set } Y)$   
(if X then Y: 만일 X가 일어나면 Y가 일어난다.)

## 2. 연관관계분석



### 다. 연관관계분석 기법

☑ 연관관계 분석을 이용 연관성을 도출하기 위해서 필요한 기준





## 2. 연관관계분석



### 다. 연관관계분석 기법

#### ④ 연관관계 분석을 이용 연관성을 도출하기 위해서 필요한 기준

##### 기준1 지지도(Support)

- 전체 거래 중에서 어떠한 항목과 다른 항목 사이에 동시에 포함하는 거래의 빈도가 어느 정도 인가를 나타냄



$$\text{Support} = \frac{n(X \cap Y)}{N}$$

## 2. 연관관계분석



### 다. 연관관계분석 기법

#### ✓ 연관관계 분석을 이용 연관성을 도출하기 위해서 필요한 기준

기준2

신뢰도 (Confidence)

- 연관성 규칙의 강도를 나타내고, 이는 다음의 조건부확률로 나타낼 수 있음

공식

$$\text{Confidence} = P(Y|X)$$

## 2. 연관관계분석



### 다. 연관관계분석 기법

#### ④ 연관관계 분석을 이용 연관성을 도출하기 위해서 필요한 기준

기준3

향상도(Lift)

- 어떠한 X상품을 구매한 경우 그 거래가 다른 Y상품을 포함하는 경우와 Y상품이 X와 상관없이 단독으로 구매된 경우의 비율을 제시

- $\text{Support}(X, Y) / (\text{Support}(X) \times \text{Support}(Y))$  로 표현

※P(Y)는 전체 거래 중 Y상품의 거래가 일어나는 확률 제시

공식

$$\text{Lift} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X) \cdot P(Y)}$$

## 2. 연관관계분석



### 다. 연관관계분석 기법

④ 연관관계 분석을 이용 연관성을 도출하기 위해서 필요한 기준

기준3

향상도(Lift)

공식

$$\text{Lift} = \frac{P(Y|X)}{P(Y)} = \frac{P(X \cap Y)}{P(X) \cdot P(Y)}$$

상호 독립적 관계

→ 상품 X와 Y간의 Lift값이 1인 경우

양의 상관관계(보완재)

→ 상품 X와 Y간의 Lift값이 1보다 큰 경우

음의 상관관계(대체재)

→ 상품 X와 Y간의 Lift값이 1보다 작은 경우

## 2. 연관관계분석



### 라. 연관관계분석 활용 분야

- 맥주와 기저귀의 상관관계 발견

» 소비자가 마트에서 물건을 살 때 맥주와 기저귀를 같이 구매 할 확률이 높다는 사실이 연관관계 기법을 통해서 알려짐

장바구니분석  
(market basket  
analysis)



어떤 물건이 어떤 물건과 같이 많이 팔리는가를 알아보는 것으로써 이를 통해, 매장에 진열을 서로 가깝게 하여 상호 판매 촉진 작용이 일어나게 함.

크로스 마케팅  
(cross-marketing)



일명 교차 판매 분석 이라고 하는데, 양복을 산 사람에게 넥타이를, 노트북을 산 사람에게 노트북 가방을, 프린터를 산 사람에게 토너를 추천하는 것

## 2. 연관관계분석



### 라. 연관관계분석 활용 분야

- 맥주와 기저귀의 상관관계 발견

» 소비자가 마트에서 물건을 살 때 맥주와 기저귀를 같이 구매 할 확률이 높다는 사실이 연관관계 기법을 통해서 알려짐

카탈로그 디자인  
(catalog design)

서로 연관이 있는 상품을 같은 카탈로그에  
진열. 책상과 의자를 같은 페이지에  
진열하는 방법

로스-리더  
(loss-leader)

미끼상품, 특가품, 유인상품, 특가상품  
등으로 불림. 소매 기업에서 기회비용을  
고려하여 가격을 낮춰 일반 물건을  
판매하는데, 이를 통해 재고를 낮추고 상점에  
고객을 불러들여 호객행위를 도모함

군집분류  
(clustering)

상품을 서로 연관성이 있는 것들끼리  
분류하는데 사용됨

## 2. 연관관계분석



### 라. 연관관계분석 활용 분야

- 맥주와 기저귀의 상관관계 발견

» 소비자가 마트에서 물건을 살 때 맥주와 기저귀를 같이 구매 할 확률이 높다는 사실이 연관관계 기법을 통해서 알려짐

어떤 질병이 걸린 사람이 다음에 어떤 특정 질병이 합병증으로 발생  
할 확률이 높은가를 알아보는데 이용

개인 파산이나 기업 파산의 징후를 미리 예측하는데 활용

보험사기 등과 같은 의료사기 적발에 활용

## 2. 연관관계분석



### 마. 연관관계분석 사례

- 토마토, 감자, 당근, 사과 오렌지 등 여러 가지의 야채와 과일을 판매하는 가게에서 수집한 1,000개의 장바구니 데이터를 연관관계로 분석한 사례

#### ④ 연관관계 룰 예 : 토마토와 상추의 연관관계 룰

토마토 → 상추

[Coverage = 0.263, Support = 0.111 (111);  
Strength = 0.422; Lift = 1.94]



## 2. 연관관계분석



### 마. 연관관계분석 사례

- 토마토, 감자, 당근, 사과 오렌지 등 여러 가지의 야채와 과일을 판매하는 가게에서 수집한 1,000개의 장바구니 데이터를 연관관계로 분석한 사례

#### ④ 연관관계 룰 예 : 토마토와 상추의 연관관계 룰

토마토 → 상추

[Coverage = 0.263, Support = 0.111(111);  
Strength = 0.422; Lift = 1.94]

- 연관관계 룰은 토마토와 상추간의 여러 연관관계 정보를 보여줌

Coverage = 0.263



전체 1,000개의 장바구니 중  
263개의 장바구니에서  
토마토가 발견된다는 의미

## 2. 연관관계분석



### 마. 연관관계분석 사례

- 토마토, 감자, 당근, 사과 오렌지 등 여러 가지의 야채와 과일을 판매하는 가게에서 수집한 1,000개의 장바구니 데이터를 연관관계로 분석한 사례

#### ④ 연관관계 룰 예 : 토마토와 상추의 연관관계 룰

토마토 → 상추

[Coverage = 0.263, Support = 0.111(111);  
Strength = 0.422; Lift = 1.94]

- 연관관계 룰은 토마토와 상추간의 여러 연관관계 정보를 보여줌

Support = 0.111



전체 1,000개의 장바구니 중  
111개의 장바구니에 토마토와  
상추가 발견된다는 의미

## 2. 연관관계분석



### 마. 연관관계분석 사례

- 토마토, 감자, 당근, 사과 오렌지 등 여러 가지의 야채와 과일을 판매하는 가게에서 수집한 1,000개의 장바구니 데이터를 연관관계로 분석한 사례

#### ④ 연관관계 룰 예 : 토마토와 상추의 연관관계 룰

토마토 → 상추

[Coverage = 0.263, Support = 0.111(111);  
Strength = 0.422; Lift = 1.94]

- 연관관계 룰은 토마토와 상추간의 여러 연관관계 정보를 보여줌

Strength = 0.422



토마토를 산 고객 중 42.2%가  
상추도 같이 구매한다는 의미

## 2. 연관관계분석



### 마. 연관관계분석 사례

- 토마토, 감자, 당근, 사과 오렌지 등 여러 가지의 야채와 과일을 판매하는 가게에서 수집한 1,000개의 장바구니 데이터를 연관관계로 분석한 사례

#### ④ 연관관계 룰 예 : 토마토와 상추의 연관관계 룰

토마토 → 상추

[Coverage = 0.263, Support = 0.111(111);  
Strength = 0.422; Lift = 1.94]

- 연관관계 룰은 토마토와 상추간의 여러 연관관계 정보를 보여줌

Lift = 1.94

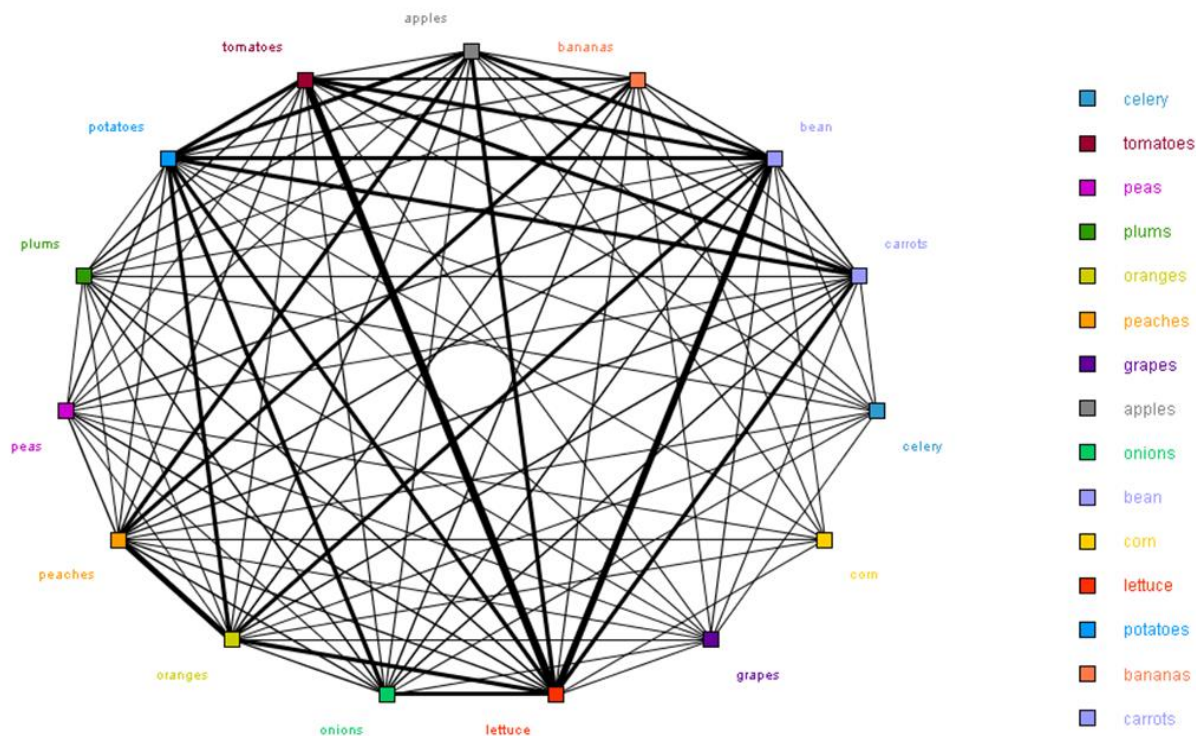


1보다 크기 때문에 토마토와  
상추는 보완재임

## 2. 연관관계분석

### 마. 연관관계분석 사례

#### 👍 그래프로 보여준 각 야채와 과일 간의 연관관계

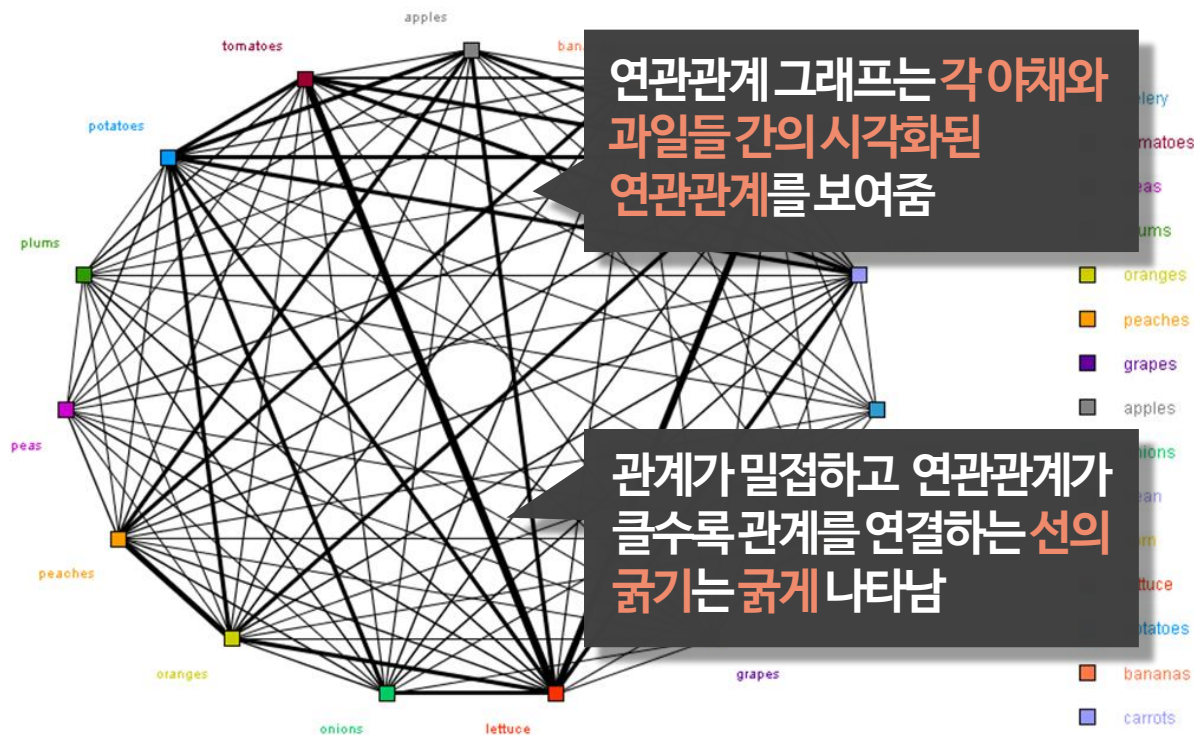


## 2. 연관관계분석



### 마. 연관관계분석 사례

#### 👍 그래프로 보여준 각 야채와 과일 간의 연관관계

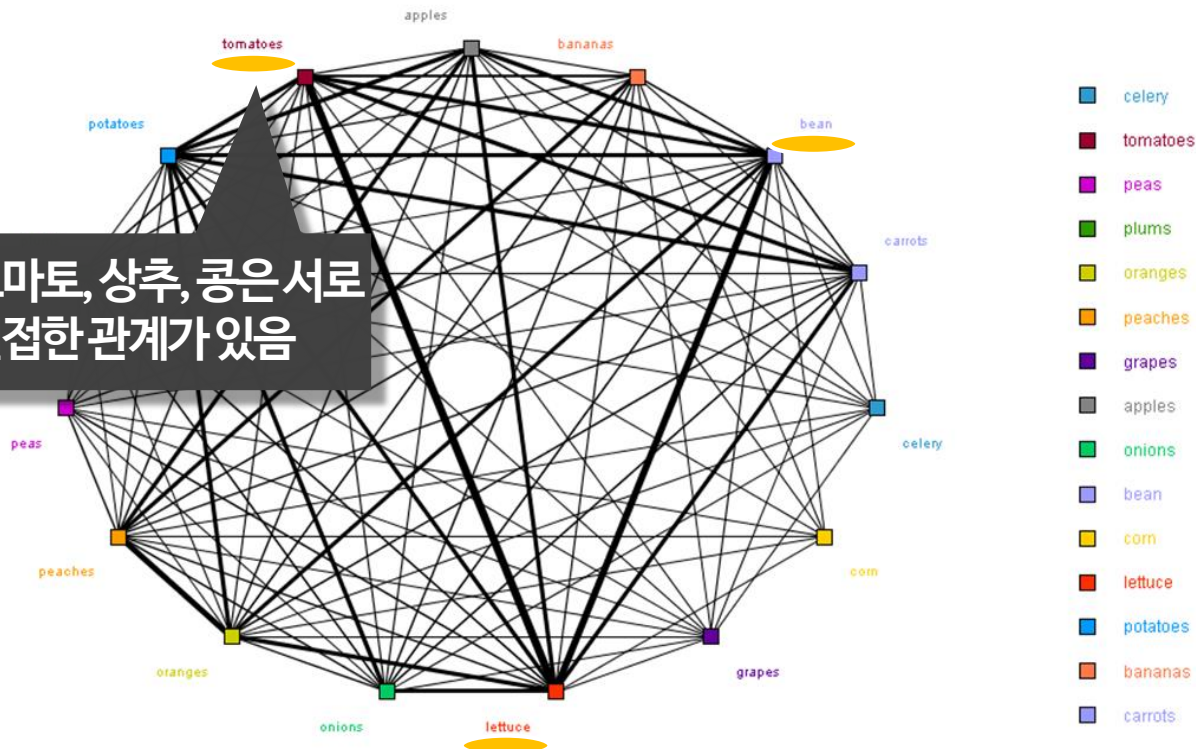


## 2. 연관관계분석

### 마. 연관관계분석 사례

#### ☑ 그래프로 보여준 각 야채와 과일 간의 연관관계

토마토, 상추, 콩은 서로  
밀접한 관계가 있음





## 2. 연관관계분석

### 마. 연관관계분석 사례

#### ✔ 그래프로 보여준 각 야채와 과일 간의 연관관계





03

# 의사결정나무

BIG

DATA

### 3. 의사결정나무



#### 가. 의사결정나무의 개념

##### 의사결정나무

데이터마이닝의 주요 기법 중 하나로서  
분류 및 예측에 주로 사용이 되는 기법

##### ● 형식

- » 데이터를 분석하여 나온 결과물이 의사결정나무라는 그래프 형식으로 표현
- » 규칙 셋이라는 형식으로도 표현

# 3. 의사결정나무



## 가. 의사결정나무의 개념

### ● 특성

- » 다른 통계기반 기법에 비교하여 분석결과 해석이 쉬움
- » 어떠한 변수들이 분류에 중요한 영향을 미치는지 설명이 가능
- » 변수들 간의 상호작용에 대한 해석이 용이

### ● 활용분야

- » 기업의 부도 예측
- » 추가 예측
- » 환율 예측
- » 경제 전망 등

### 3. 의사결정나무



#### 나. 의사결정나무 기법

##### 01 의사결정 나무의 형성

- » 분석의 목적과 자료구조에 따라 적절한 분리 기준과 정지 규칙을 지정하여 의사결정나무를 획득

##### 02 가지치기

- » 분류 오류를 크게 할 위험이 높거나 부적절한 추론 규칙을 가지고 있는 가지를 제거

##### 03 타당성 평가

- » 이익도표나 위험도표 또는 검증용 자료에 의한 교차타당성 등을 이용하여 의사결정나무를 평가

##### 04 해석 및 예측

- » 의사결정나무를 해석하고 예측모형을 설정

### 3. 의사결정나무



#### 나. 의사결정나무 기법

##### ④ 나무의 구조와 의사결정나무 방법 알고리즘

- 나무구조 형성의 형태 중 하나는 **이진트리구조**를 들 수 있음
- 이 구조는 각각의 노드가 두개의 자식노드를 만들어 yes 또는 no 질문에 답함으로써 **터미널노드까지 진행해 나감**
- 단순한 이진트리모양만 있는 것이 아니라 **혼합된 형태의 모형도** 있음

### 3. 의사결정나무



#### 나. 의사결정나무 기법

##### ④ 나무의 구조와 의사결정나무 방법 알고리즘

- 의사결정나무를 형성하는 단계에서 사용되는 대표적인 알고리즘의 종류

CART

지니지수 (Gini Index) 또는 분산의 감소량을  
분리기준으로 활용하고 이진분리를 수행함

C4.5  
알고리즘

엔트로피지수를 분리 기준으로 활용

CHAID

카이제곱 - 검정 또는 F - 검정을 분리 기준으로  
활용하고 다지 분리 수행이 가능

# 3. 의사결정나무



## 나. 의사결정나무 기법

### ✓ 의사결정나무의 장점

- **주요 변수의 선정이 용이**  
중요한 변수만 선별하여 의사결정나무를 구성
- **교호효과의 해석**  
두 개 이상의 변수가 결합하여 목표변수에 어떻게 영향을 주는지 쉽게 파악 가능
- **비모수적 모형**  
선형성, 정규성, 등분산성 등의 가정이 필요 없음
- **해석의 용이성**  
모형의 이해가 쉽고, 새로운 자료의 모형에 적합하며, 어떤 입력변수가 목표변수를 설명하기에 좋은지 쉽게 파악 가능
- **지식의 추출**  
의사결정나무를 룰로 자동 변화가 가능하며, 이 룰은 다양한 활용이 가능

### 3. 의사결정나무



#### 나. 의사결정나무 기법

##### ④ 의사결정나무의 단점

- 비연속성

연속형 변수를 비연속적인 값으로 취급하기 때문에 분리의 **경계점** 근방에서 **예측 오류가 클** 가능성이 있음

- 선형성 또는 주 효과의 결여

선형 또는 주 효과 모형에서와 **같은 결과를 얻을 수 없다는** 한계점

- 불안정성

분석용 자료에만 의존하기 때문에 **새로운 자료의 예측에서는 불안정** 할 가능성이 높음

- 몇몇 의사결정나무 알고리즘이 이진분리를 이용하기 때문에 **분리 가지의 수가 많음**

- 나무형성 시 컴퓨팅 비용이 많이 듦



### 3. 의사결정나무



#### 다. 의사결정나무 활용 분야

- 고객, 상품, 서비스 등의 세분화, 분류, 예측에 주로 응용

##### » 분류(classification)

의사결정나무의 트리(tree) 또는 트리에서 추출된 룰(rule)들을 이용하여 **고객, 상품, 서비스 등의 분류를 하는데 활용**하고, 어떤 기업이 부도가 날 기업인지 아닌지, 어떤 고객이 특정 물건을 살 것인지 아닌지를 알 수 있게 함.

##### » 타겟 마케팅(target marketing)

고객의 분류를 통하여 **특정 물건을 구매 할 확률이 높은 고객에게만 광고**를 하는 것

##### » 예측(prediction & forecasting)

**미래에 어떤 사건이 일어날 확률을 예측하는 것으로** 기업의 부도 확률 예측, 개인 고객의 개인 파산 확률부터, 날씨예측, 경제 지표예측, 주가 예측 등 다양한 예측에 사용

### 3. 의사결정나무



#### 다. 의사결정나무 활용 분야

##### » 지식의 추출(rule induction)

의사결정나무는 룰(rule) 형태의 지식으로의 변환 가능

예

- 지식 경영 시스템에서 활용
- 전문가 시스템의 룰엔진(rule engine)에서 활용

##### » 주요변수의 추출(variable selection)

복잡한 절차를 걸쳐 주요 변수를 추출하는 과정을 단순화 할 있는데,  
데이터에 나타나는 주요 독립변수 중 중요한 변수만이 의사결정  
나무에 나타남



- 고객관리, 광고 전략, 신상품개발, 품질관리, 신용평가 등  
다양한 분야에서 활용

### 3. 의사결정나무



#### 라. 의사결정나무 사례

##### ✓ 변수의 선정과 데이터의 추출

- 선행연구들에 기초하여 기업 생존/부도에 영향을 미치는 변수로 18개의 재무변수 선정

##### » 예측 입력 변수

유동비율, 당좌비율, 부채비율, 자기자본비율, 총자본순이익율,  
총자본경상이익율, 매출액순이익율, 매출액경상이익율, 매출액증가율,  
순이익증가율, 이자보상비율, 수정이자보상비율, 총자본회전율,  
총부채회전율, 금융비용무담률, 순운전자본비율, 차임금의존도, 현금비율

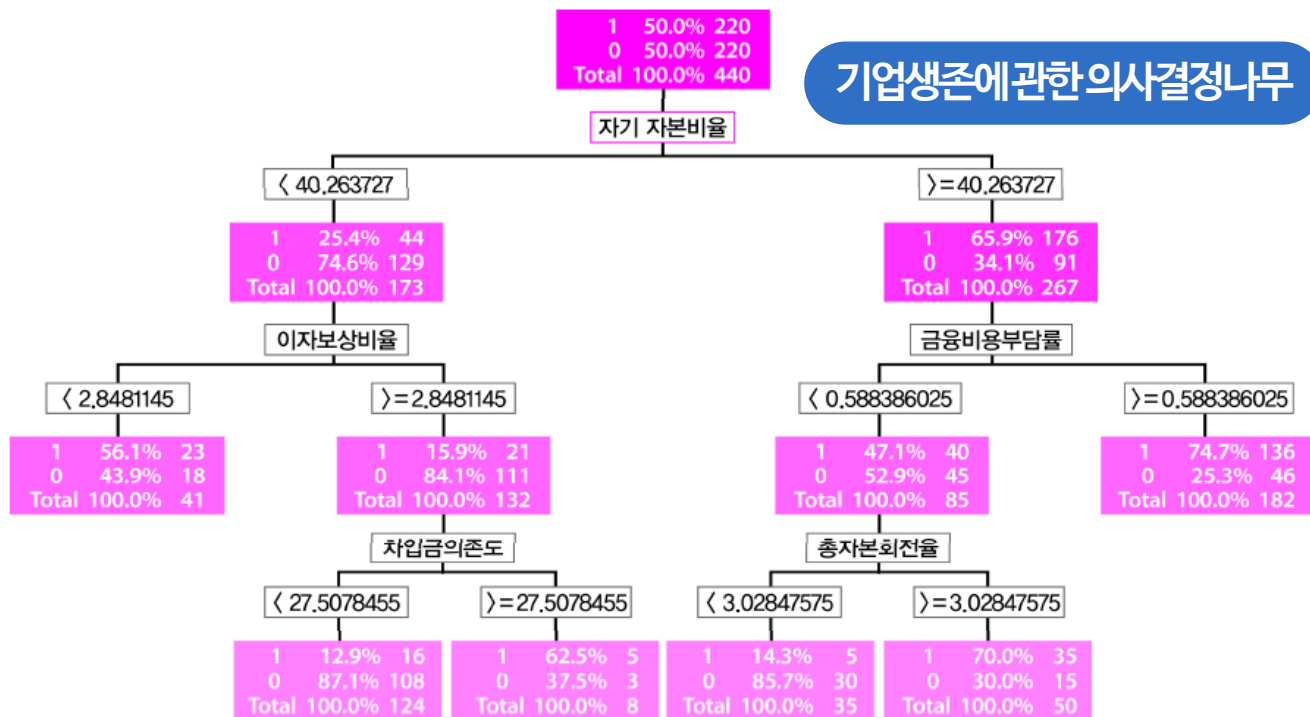
- 표본기업으로 A 업종 중 정상기업 220개사, 부실기업 220개사의 표본 적용

# 3. 의사결정나무

## 라. 의사결정나무 사례

### ☑ 데이터 분석을 통한 의사결정나무의 도출

- 의사결정나무를 이용한 데이터분석을 통해 의사결정나무 추출

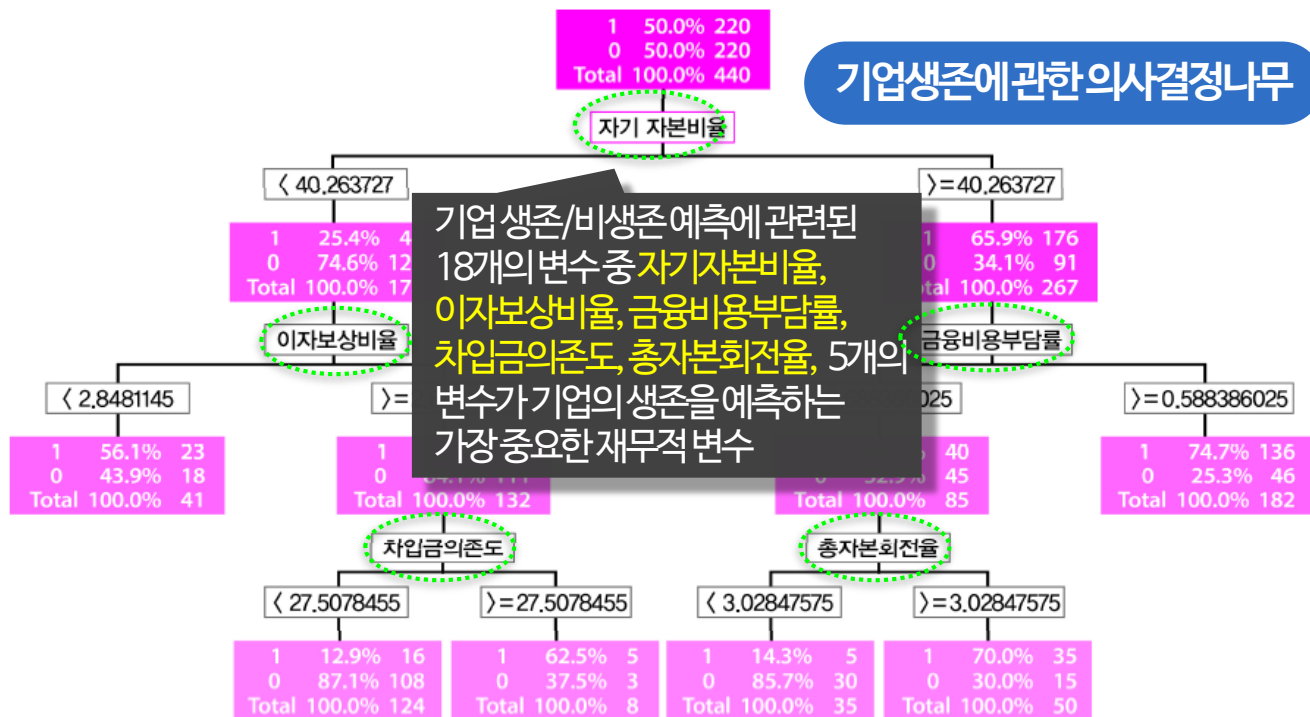


# 3. 의사결정나무

## 라. 의사결정나무 사례

### ✓ 데이터 분석을 통한 의사결정나무의 도출

- 의사결정나무를 이용한 데이터분석을 통해 의사결정나무 추출

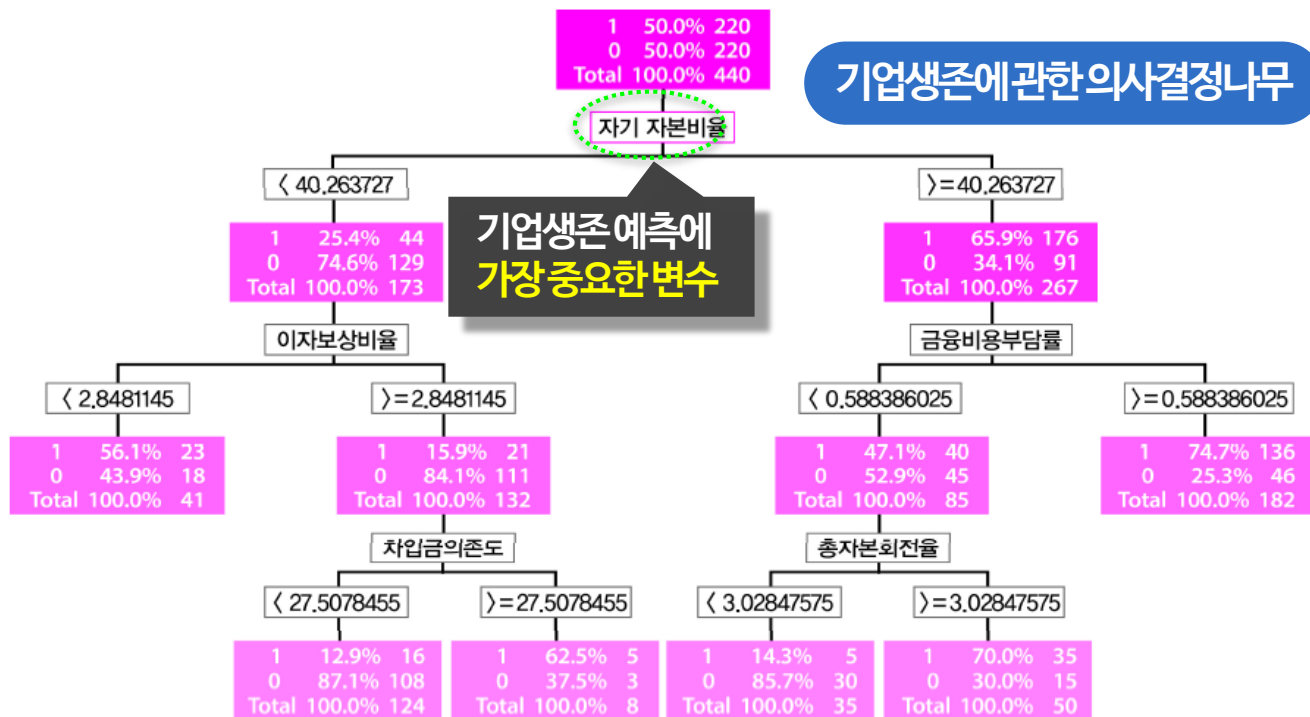


# 3. 의사결정나무

## 라. 의사결정나무 사례

### ✓ 데이터 분석을 통한 의사결정나무의 도출

- 의사결정나무를 이용한 데이터분석을 통해 의사결정나무 추출

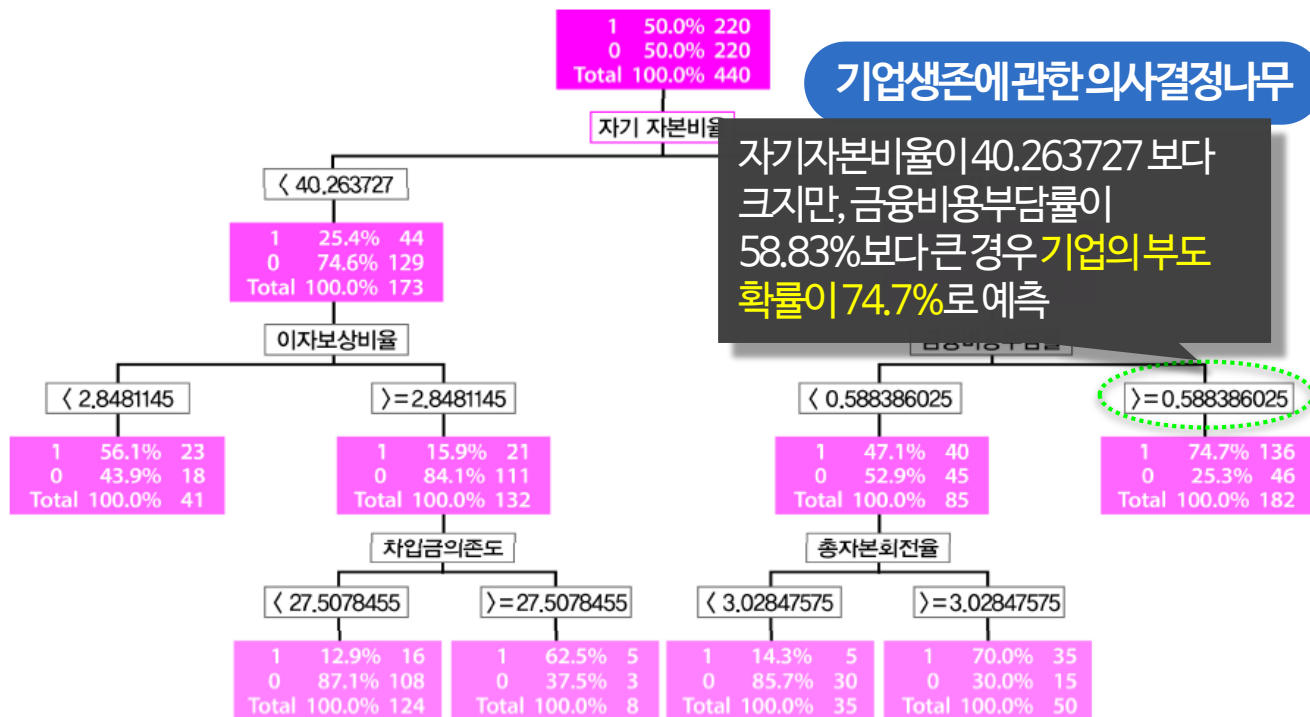


# 3. 의사결정나무

## 라. 의사결정나무 사례

### ☑ 데이터 분석을 통한 의사결정나무의 도출

- 의사결정나무를 이용한 데이터분석을 통해 의사결정나무 추출

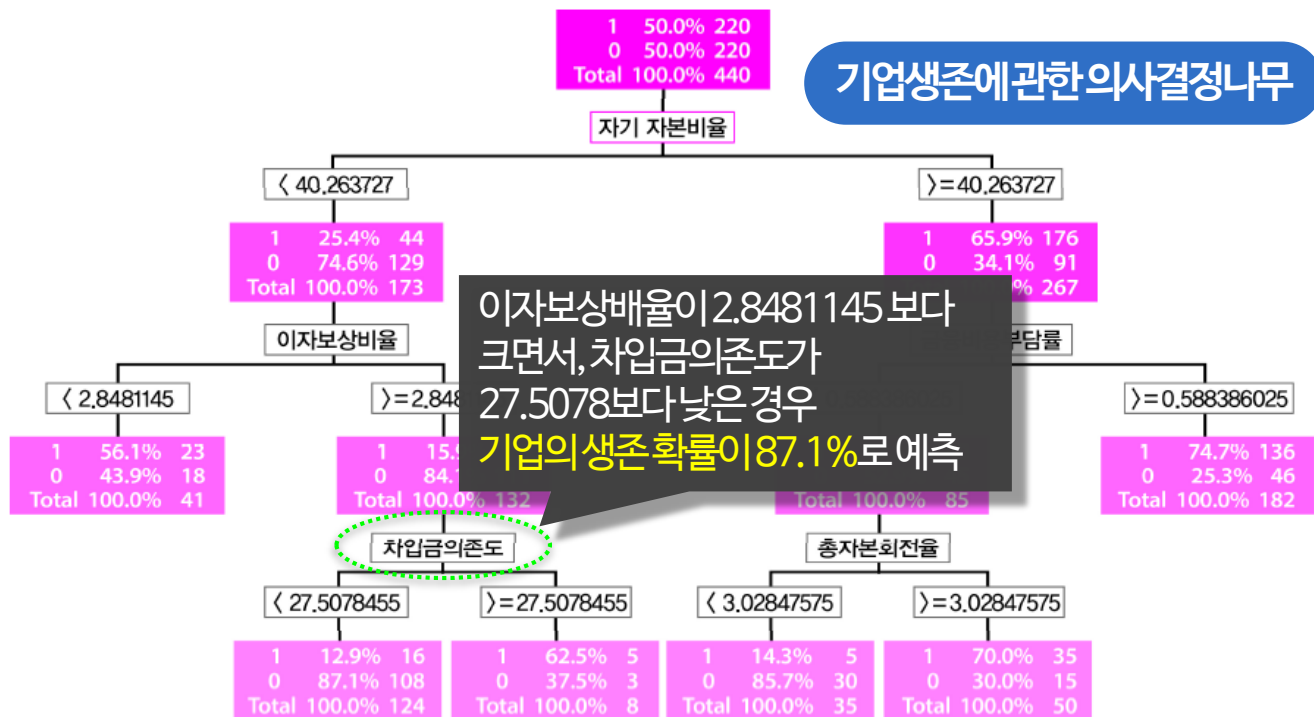


# 3. 의사결정나무

## 라. 의사결정나무 사례

### ☑ 데이터 분석을 통한 의사결정나무의 도출

- 의사결정나무를 이용한 데이터분석을 통해 의사결정나무 추출





04

# 인공신경망

BIG

DATA

## 4. 인공지능경망



### 가. 인공지능경망의 개념

생물학적 뇌의 작동 원리를 그대로 모방하는  
방법으로, 데이터 안의 독특한 패턴이나 구조를  
인지하는데 필요한 모델을 구축하는 기법



간단한 계산능력을 가진 처리 단위,  
뉴런 또는 노드들이 서로 복잡하게 연결된 컴퓨터 시스템으로서  
외부에서 주어진 입력에 대하여 반응을 함

## 4. 인공지능경망



### 가. 인공지능경망의 개념

- 구성하고 있는 다수의 뉴런끼리의 상호연결성에 기인함
  - » 뉴런  
생체내의 신경세포와 비슷한 것으로써 가중치화 된 상호연결성으로 서로 연결됨

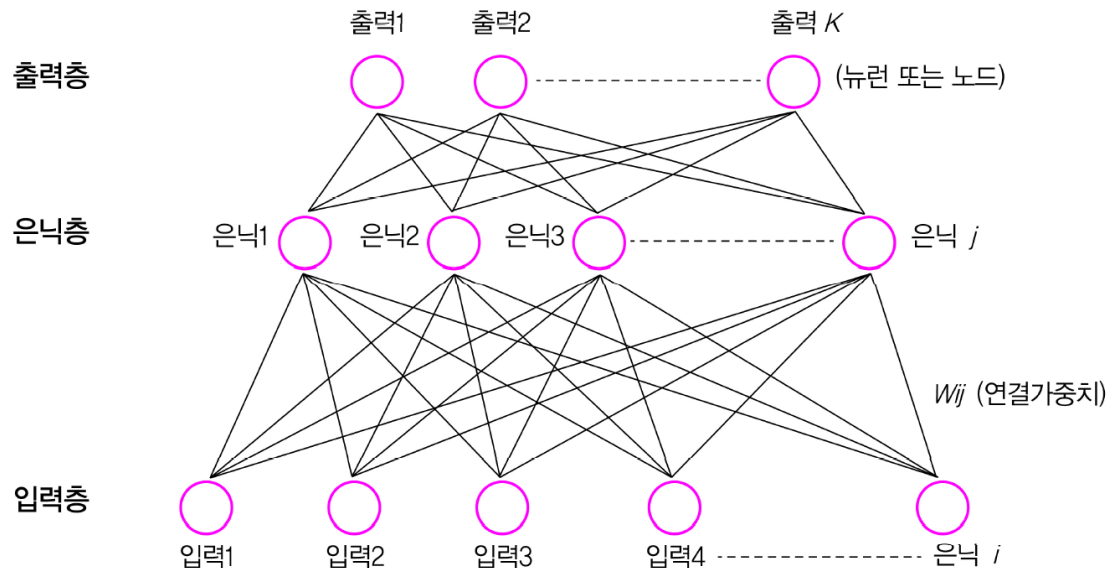
# 4. 인공지능망



## 가. 인공지능망의 개념

### ✓ 인공지능망 모형

- 가장 일반적인 인공지능망 모형은 다계층 퍼셉트론 모형



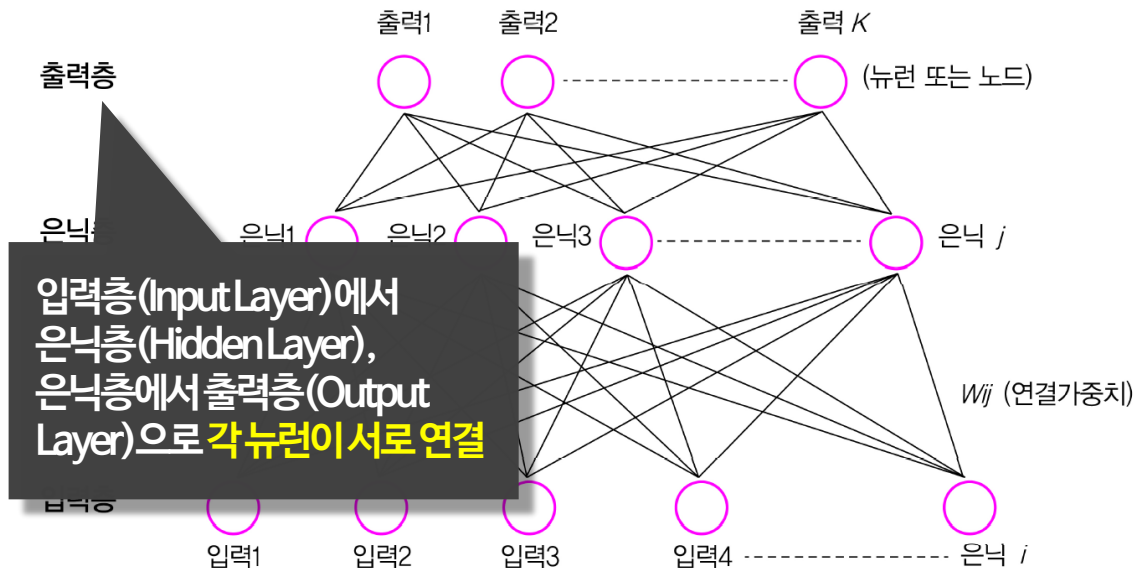
# 4. 인공지능망



## 가. 인공지능망의 개념

### ✓ 인공지능망 모형

- 가장 일반적인 인공지능망 모형은 다계층 퍼셉트론 모형



## 4. 인공지능경망



### 나. 인공지능경망의 특징

- 인간의 뇌처럼 다양한 뉴런이 서로 연결된 구조를 이용하여 의사결정이 이루어지고 있는 구조를 이용
- 인공지능경망은 자료의 관련성을 나타내 줄 수 있는 기법으로 뇌의 신경시스템을 응용하여 예측을 최대화하기 위한 조직화를 찾기 위해 반복적으로 학습하는 원리 이용
- 인공지능경망은 복잡하고 비선형적이며 관계성을 갖는 다변량을 분석 가능

# 4. 인공지능경망



## 나. 인공지능경망의 특징

### 장점

- » 회귀분석과 같은 선형기법과 비교하여 비선형기법으로서의 예측력이 뛰어남
- » 자료에 대한 통계적 분석 없이 결정을 수행 가능
- » 통계적 기본가정이 적고 유연하여 다양하게 활용
- » 데이터 사이즈가 작은 경우, 불완전데이터, 노이즈 데이터가 많은 경우 인공지능경망 모델의 성능이 일반적으로 다른 기법과 비교해서 우수하다고 평가됨

### 단점

- » 모델이 제시하는 결과에 대해서 왜 그런 결과가 나오는지에 대한 원인을 명쾌하게 설명할 수 없음
- » 모델의 학습에 시간이 과도하게 많이 소요됨
- » 전체적인 관점에서의 최적해가 아닌 지역 내 최적해가 선택 가능
- » 과적합화(Overfitting)가 될 수 있음

## 4. 인공지능경망



### 다. 인공지능경망 활용 분야

- 인공지능경망은 타마호크라는 첨단 미사일의 눈에 장착이 되어 미사일의 눈의 역할을 한 것으로 언론에 처음으로 알려짐
- 이후 경영학에서는 부도 예측을 통해서 널리 알려지는 계기가 됨



# 4. 인공지능경망



## 다. 인공지능경망 활용 분야

- » 마케팅(marketing)  
고객분류, 소비패턴분석, 이탈고객방지, 타겟마케팅, 광고전략, 판매예측 등에 이용
- » 회계(accounting)  
탈세자 파악에 사용되며 요즘 회계에서 중요한 분야인 회계 감사에 활용
- » 재무(finance)
  - 부도예측, 사기적발, 신용분석, 대출결정, 경제예측에 사용
  - 로보어드바이저(robo-advisor)는 재무분석가(financial analyst)의 기능을 수행하여 주식의 추천을 도와주고, 주가의 변동성 예측, 자산의 운영 조언
- » 인사관리(human resource)  
직원의 실적 예측, 직원의 적정 업무에의 배치, 직원 교육 등에 활용
- » 생산관리(manufacturing)  
작업관리, 고장예방, 수요예측, 최적 길 찾기, 최적 순차 방문 스케줄링, 작업 스케줄링 등에 활용
- » 음성인식, 지문인식, 글자인식 등을 통해서 생활 곳곳에서 활용

## 4. 인공지능경망



### 라. 인공지능경망 사례

- 사례는 앞서 의사결정나무에서와 같은  
기업생존/부도 예측 문제를 인공지능경망 모델로 구현함

#### ✓ 기업생존/부도 예측변수

- |          |            |
|----------|------------|
| » 차입금의존도 | » 총자본경상이익율 |
| » 자기자본비율 | » 유동비율     |
| » 이자보상비율 | » 총부채회전율   |
| » 부채비율   | » 순운전자본비율  |

# 4. 인공지능경망



## 라. 인공지능경망 사례

### ④ 신경망모형 분석 결과

- » Estimated Accuracy : 74.8
- » Input Layer : 8 Neurons
- » Hidden Layer 1 : 3 Neurons
- » Output Layer : 2 Neurons
- » Relative Importance of Inputs
  - 차입금의존도 0.532354
  - 자기자본비율 0.4585
  - 이자보상비율 0.386533
  - 부채비율 0.29293
  - 총자본경상이익율 0.216319
  - 유동비율 0.134382
  - 총부채회전율 0.107228
  - 순운전자본비율 0.072239

- 인공지능경망의 예측되는 정확도, 인공지능경망의 구조, 중요한 예측변수들을 제시
- 인공지능경망 모형은 8개의 변수 각각에 한 개씩 Input 노드 부여

# 4. 인공지능경망

## 라. 인공지능경망 사례

④ 신경망 모델

입력층(Input Layer)은 8개의 Input Node로 구성

Estimated Accuracy	0.95
Input Layer	: 8 Neurons
Hidden Layer 1	: 3 Neurons
Output Layer	: 2 Neurons
Relative Importance of Inputs	
- 하위급이익률	0.532354
- 부채비율	0.4585
- 하위급이익률	0.386533
- 부채비율	0.29293
- 총자본경상이익율	0.216319
- 유동비율	0.134382
- 총부채회전율	0.107228
- 순운전자본비율	0.072239

결과층(Output Layer)은 2개의 Node로 구성

- 인공지능경망의 예측되는 정확도, 인공지능경망의 구조, 중요한 예측변수들은
- 은닉층(Hidden Layer)은 3개의 node로 구성
- 인공지능경망 모형은 8개의 변수 각각에 한 개씩 Input 노드 부여

# 4. 인공지능경망



## 라. 인공지능경망 사례

### ④ 신경망모형 분석 결과

- » Estimated Accuracy : 74.8
- » Input Layer : 8 Neurons
- » Hidden Layer 1 : 3 Neurons
- » Output Layer : 2 Neurons
- » Relative Importance of Inputs
  - 차입금의존도 0.532354
  - 자기자본비율 0.4585
  - 기업생존/부도에 가장 큰 영향을 미치는 중요한 변수 0.6533
  - 총자산경상이익율 0.216319
  - 유동비율 0.134382
  - 총부채회전율 0.107228
  - 순운전자본비율 0.072239

- 인공지능경망의 예측되는 정확도, 인공지능경망의 구조, 중요한 예측변수들을 제시
- 인공지능경망 모형은 8개의 변수 각각에 한 개씩 Input 노드 부여

## 4. 인공지능경망



### 라. 인공지능경망 사례

#### ④ 신경망 훈련결과

부도여부		정상	부도	전체
정상	Count	99	26	125
부도	Count	48	121	169
전체	Count	147	147	294

#### ④ 신경망 테스트결과

부도여부		정상	부도	전체
정상	Count	34	15	49
부도	Count	39	58	97
전체	Count	73	73	146

## 4. 인공지능경망



### 라. 인공지능경망 사례

#### ④ 인공지능경망의 판별분석과 예측 정확도 비교 표

모형	훈련자료	테스트자료
판별분석	62.2%	60.3%
인공지능경망모형	74.8%	63.0%

- 인공지능경망 기법의 예측 정확도를 다른 모델의 예측 정확도와 비교하여 표로 제시
- 기업부도예측분석기법으로 널리 사용되어지고 있는 판별분석을 이 사례의 기업부도예측에 적용
- 인공지능경망모형의 예측력이 63%로 판별분석기법의 60.3% 보다 높아 인공지능경망기법이 기업부도예측에 판별분석기법보다 우수한 기법임

05

# 사례기반추론

BIG

DATA



## 5. 사례기반추론



### 가. 사례기반추론의 개념

과거에 있었던 사례들의 결과를 바탕으로  
새로운 사례의 결과를 예측하는 기법

## 5. 사례기반추론



### 가. 사례기반추론의 개념

- 과거에 발생한 문제는 미래에 다시 비슷한 형태의 문제로 발생할 가능성이 높고 새로운 문제를 해결할 수 있는 정답이 과거의 문제를 해결했던 정답과 유사할 것이라는 가정을 이용
- 과거 사례들을 저장해 둔 사례기반으로부터 해결하고자 하는 새로운 사례와 가장 유사한 사례를 검색한 후, 유사 사례의 해결책을 바탕으로 당면한 문제의 해결책을 제안하는 과정으로 진행
- 사례기반추론을 이용하기 위해서는 일반적으로 과거의 사례와 사례들 사이의 유사 정도를 측정하기 위한 유사도 척도 준비
  - » 유사도 측정도구는 여러 가지 방법이 제안되고 있지만 일반적으로 근접이웃방법론이 가장 많이 이용

## 5. 사례기반추론



### 가. 사례기반추론의 개념

- 근접이웃방법론으로 입력문제 T와 사례 S에 대한 유사도를 계산

#### ④ 근접이웃방법론을 이용한 유사도 계산방법

T는 입력문제, S는 학습된사례,  
W<sub>i</sub>는 T와 S의 각 속성에 대한 가중치

공식

$$\text{Similarity}(T, S) = \sum_{i=1}^n f(T_i, S_i) \times W_i$$

# 5. 사례기반추론



## 가. 사례기반추론의 개념

- 근접이웃방법론으로 입력문제 T와 사례 S에 대한 유사도를 계산

### ④ 근접이웃방법론을 이용한 유사도 계산방법

» 각 속성에 대한 유사도 함수 f는 자료가 수치형 자료인지 또는 범주형 자료인가에 따라서 계산

수치형 속성인  
경우 함수 f 계산

$$f = 1 - | (\text{비교사례의 속성값} - \text{과거사례의 속성값}) / \text{해당속성의 최대값} |$$

범주형 속성인  
경우 함수 f 계산

$$f = 1 - | \text{비교사례의 속성값} - \text{과거사례의 속성값} |$$

# 5. 사례기반추론



## 나. 사례기반추론 과정

### 01 검색(Retrieve)

- » 대상 문제가 주어지면, 사례 데이터베이스에서 그것을 풀기에 적절한 사례들을 검색

### 02 재사용(Reuse)

- » 이전의 사례로부터 대상문제의 해결 방법을 연결

### 03 수정(Revise)

- » 이전의 해결 방법을 대상의 상황에 연결시킨 후, 그 새로운 해결 방법을 실 세계에서 테스트하고, 필요하다면 수정

### 04 검색(Retrieve)

- » 해법이 성공적으로 대상문제에 적용된 후에, 그러한 새로운 경험이 사례 데이터베이스에 새로운 사례로서 저장

# 5. 사례기반추론



## 다. 사례기반추론 특징

### 장점

- » 인간의 문제 해결 방식과 유사하기 때문에 그 결과를 이해하기 쉬움
- » 새로운 사례를 단순히 저장하는 것만으로도 추가적인 작업 없이 학습 진행
- » 사례기반추론 모델은 구조가 간단하고 이해가 용이
- » 수치형 변수와 범주형 변수 모두가 사용 가능
- » 복잡한 문제를 비교적 적은 정보로 의사결정, 문제해결 가능

### 단점

- » 전통적인 사례기반추론의 경우, 타 인공지능기법 또는 데이터마이닝 기법에 비해 정확도가 상대적으로 크게 떨어짐
- » 사례를 저장하기 위한 공간이 많이 필요
- » 일반화를 위한 학습과정과 해결이 동시에 일어나기 때문에 많은 시간이 소요
- » 사례를 설명하고 있는 속성이 적절하지 못한 경우 성능이 크게 저하

# 5. 사례기반추론



## 라. 사례기반추론 활용 분야

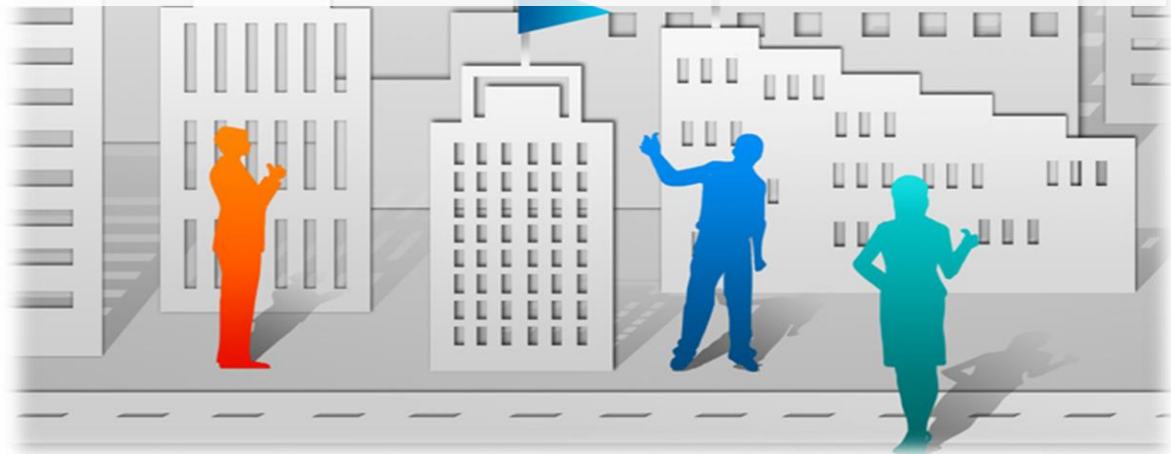
- » 고객응대(call center)
  - 고객의 문의 전화 시 어떻게 답을 주어야 하는지 상담자들을 도와주는 업무
  - 프린터를 구매한 고객이 어떠한 프린터 고장이 있는지를 파악하는데 도움을 줌
- » 진단시스템(diagnosis system)  
공장에서 문제가 발생 시, 자동차 또는 비행기의 문제점이 무엇인지를 진단하는데 사용
- » 전문가 시스템(expert system)  
병원에서 의사가 질병을 파악하는데, 법률가가 과거 판례를 분석 할 때 사례기반 시스템이 널리 쓰임. 많은 병원에서 병을 진단하고 처방을 내리는데 사용
- » 프로젝트관리(project management)  
회사에서 프로젝트 제안서 작성 시 반복되는 유사한 프로젝트들을 검색하여 이를 참조하는데 사용하여 시간과 비용을 줄이고 제안서의 품질을 높임
- » 지식관리 시스템(knowledge management system)  
보다 고도화된 지식관리 시스템에 적용이 되어 고객관리, 고객응대, 고객분류 등에 활용

## 5. 사례기반추론



### 마. 사례기반추론 사례

보험회사인 A사는 청약 프로세스에서 고객들이 보험설계를 마친 후에 프로세스를 종료하지 않고 실제 구매 단계인 청약신청 단계로 유인하기 위해서 개인화된 보험금 지급사례와 보험통계 정보를 제공하는 온라인보험 판매지원시스템을 설계





## 5. 사례기반추론



### 마. 사례기반추론 사례

④ 사례기반추론을 이용한 온라인보험 판매지원시스템을 구성하기 위해서는?

온라인보험 설계를 진행하는 **고객의 프로파일에 기초**해서  
질의를 통해 **관련 사례**들을 추출

추출된 사례들과 고객과의 유사도를 계산하여  
**유사도가 가장 높은 사례**들을 최종 사례로 선정

선정된 사례의 속성 값을 기반으로  
사례의 내용과 일치하는 **보험통계 정보**를 추출

추출된 보험금 지급사례와 보험통계는 고객이 온라인 보험의  
설계를 종료하는 시점에서 **청약신청 단계**로 유인하기 위한  
**정보**로 활용

## 5. 사례기반추론



### 마. 사례기반추론 사례

✓ 사례기반추론을 이용한 온라인보험 판매지원시스템을 구성하기 위해서는?

보험설계 고객 특성과 보험금 지급사례의 고객 특성 간에 유사도를 계산하여 **고객 유사도를 산출**

가장 최근의 사례를 제공하기 위해 사례의 최신도를 곱하여 **최종 유사도를 산출**

사례기반추론을 이용한 A사의 온라인보험 판매지원시스템은 **개인화된 청약유인 정보를 제공하여 고객을 청약신청단계로 유인**

기존의 다른 추천 시스템을 이용한 경우 구매의도가 평균 2.83/4.0로 나타난 반면에 **사례기반추론 시스템을 이용한 경우 구매의도가 3.17/4.0로 더욱 높게 향상**

06

# 군집 분석 (Cluster Analytics)

BIG

DATA

## 6. 군집 분석(Cluster Analytics)

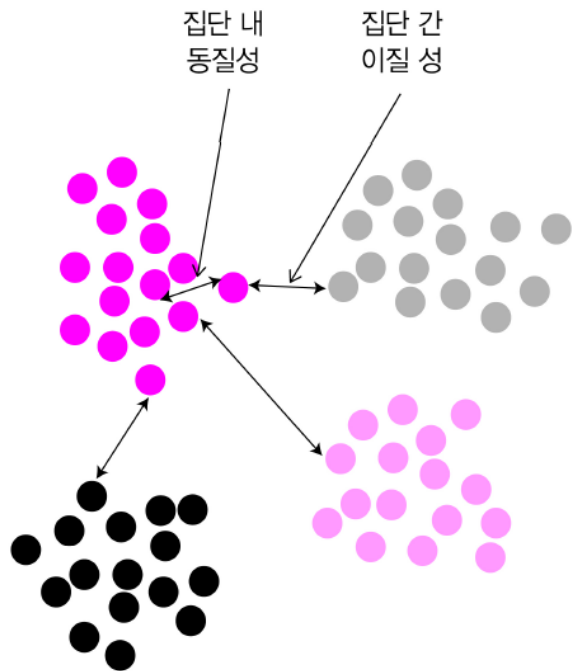


- 전체 데이터를 군집을 통해 잘 구분하는 것으로 다양한 특징을 가진 **관찰 대상으로부터 동일집단으로 분류**하는데 사용
- 유사한 특성을 가진 개체를 합쳐가면서 최종적으로 **유사 특성의 군집을 찾아내는 분류방법**
- 구분하려고 하는 각 군집에 대한 아무런 사전지식이 없는 상태에서 분류하는 것
  - 무감독학습(Unsupervised Learning)에 해당
- ≫ 개체들에 대한 사전지식 없이 유사도에 근거하여 군집들을 구분
- ≫ **군집(cluster)**  
개체 공간에 주어진 유한 개의 개체들이 서로 가깝게 모여서 무리를 이루고 있는
- ≫ **클러스터링(clustering)**  
개체 집합을 군집화 하는 과정

## 6. 군집 분석(Cluster Analytics)



### ④ 군집분석에서의 군집 특성



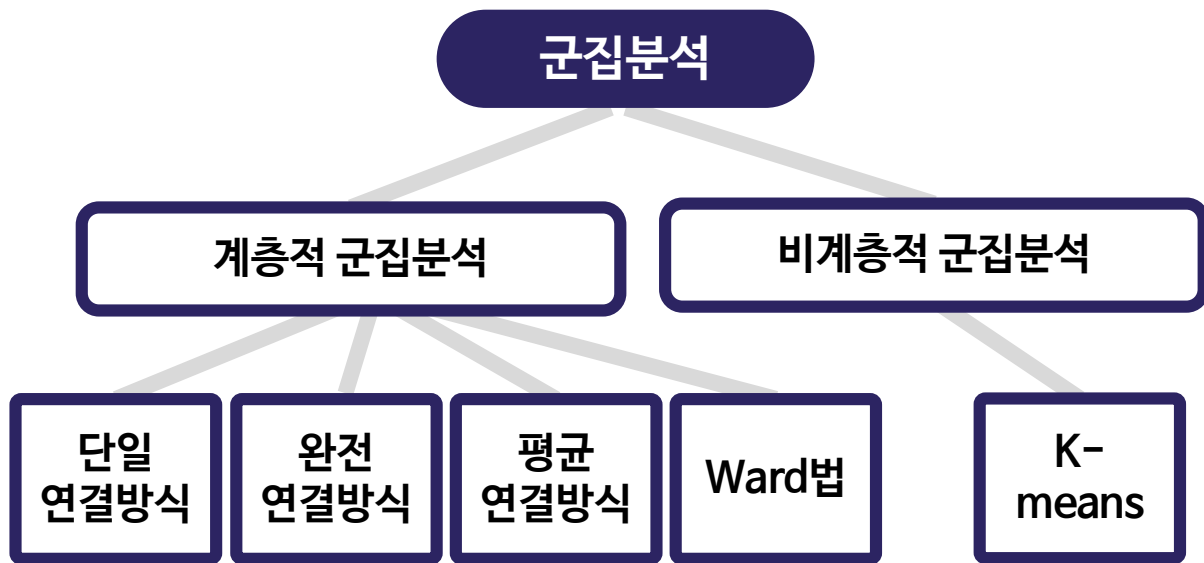
- » 군집간의 유사도를 평가하기 위해서 여러 가지의 거리 측정 함수 사용
- » 유클리언 거리 (Euclidean distance), 마할라노비스 거리 (Mahalanobis distance), 헤밍 거리 (Hamming distance) 등이 사용

## 6. 군집 분석(Cluster Analytics)



### 가. 계층적 군집분석

#### ☑ 군집분석의 분류

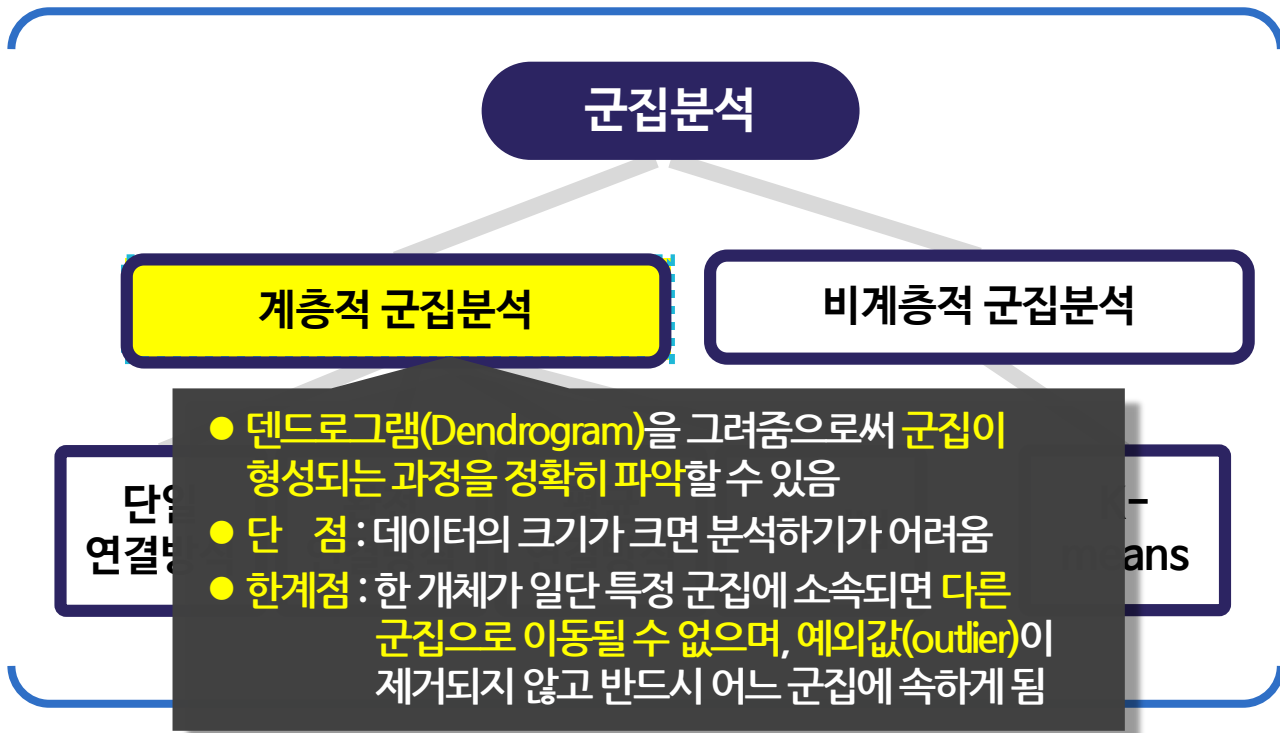


## 6. 군집 분석(Cluster Analytics)



### 가. 계층적 군집분석

#### ④ 군집분석의 분류



## 6. 군집 분석(Cluster Analytics)



### 나. 비계층적 군집분석

군집의 수가 한 개씩 감소하는 것이 아니라  
사전에 정해진 군집의 숫자에 따라 대상들이  
군집들에 할당되는 것

- 단점

많은 데이터를 빠르고 쉽게 분류할 수 있으나 군집의 수를 미리  
정해주어야 하고, 군집을 형성하기 위한 초기값에 따라  
군집결과가 달라짐



## 6. 군집 분석(Cluster Analytics)



### 다. 군집분석의 활용 분야

- divide & conquere 방법

고객을 몇 개의 그룹으로 나누고 각 그룹에 맞추어 전략을 짜는 것

예

광고 전략을 만들 때  
모든 고객에게 똑같은 광고 메시지를 보내는 것보다 고객을  
특성에 따른 군집화를 하여 각 그룹 특성에 맞게 타겟 마케팅을  
하는 것이 광고의 효율성을 높일 수 있다는 것이다.

07

# 머신러닝 (machine learning)

BIG

DATA

# 7. 머신러닝(machine learning)



## 가. 머신러닝의 개념

- 머신(기계), 즉 컴퓨터를 인간처럼 학습시켜 스스로 규칙을 형성할 수 있지 않을까? 하는 시도에서 시작
- 통계적인 접근 방법 사용

예

“독감이 걸린 사람은 대부분 열이 많이 나고 오한이 있고 구토 증상이 있었다”라는 통계에 기반하여 독감을 진단하는 것이다.

예

인공지능의 한 분야인 머신러닝을 통해서 수신한 이메일이 스팸인지 아닌지를 구분할 수 있도록 훈련할 수 있는 것이다.

- 경험에 해당하는 데이터가 가장 중요한데, 좋은 품질의 데이터를 많이 가지고 있다면 보다 높은 성능을 끌어낼 수 있음

# 7. 머신러닝(machine learning)



## 나. 머신러닝의 활용 분야

- 이메일 스팸 필터링 시스템, 쇼핑몰이나 영화 연관 추천 시스템, 문자 인식, 자연어 처리, 연관 검색어 처리 등
- 샘플 데이터를 통해서 지속적으로 알고리즘을 학습해 나감으로써, 최적의 알고리즘을 찾아가고 결국 적절한 답을 찾아내도록 하는 것



쇼핑몰 추천 시스템의 경우 사용자의 구매 패턴을 군집화 하여, 유사한 패턴을 찾아냄으로써 적절한 상품 추천

- ➔ 40대 미혼/남자/연수입 5000만원/차량 보유 사용자  
카메라, 배낭 등을 구매
- ➔ 여행 상품을 구매할 확률이 높음
- ➔ 사용자에게 여행 상품 추천

- 인공지능, 검색엔진, 광고, 마케팅, 로봇, 인사활동 등 거의 모든 시스템에 활용 될 듯