

# 인공지능 활용 능력 개발 중급

## [3차시]

# 목차

## I. 머신러닝이란?

1. 머신러닝 정의
2. 머신러닝 특징
3. 머신러닝 사례
4. 머신러닝 알고리즘 종류
5. 머신러닝 프로세스
6. 머신러닝 알고리즘
7. 머신러닝 기초수학
8. 데이터 전처리
9. 편향과 분산

## II. 지도 학습 – 회귀(예측)(선형회귀분석)

1. 회귀(예측) – 선형회귀분석

## III. 지도 학습 – 분류( $\kappa$ -NN, Decision Tree)

1. 분류 –  $\kappa$ -NN
2. 분류 – Decision Tree(의사결정나무)

## IV. 비지도 학습( $\kappa$ -means Clustering)

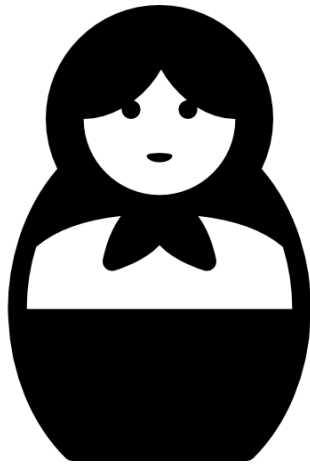
1. 비지도 학습 –  $\kappa$ -means Clustering

# 머신러닝 정의

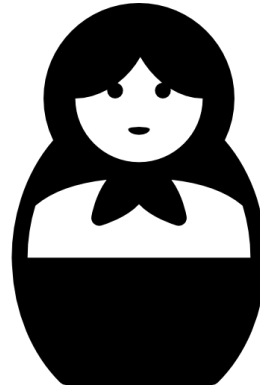
- 머신러닝은 데이터로부터 학습하도록 컴퓨터를 프로그래밍 하는 과학
  - "머신러닝은 명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야"
    - 아서 새뮤얼(Arthur Samuel), 1959
  - "어떤 작업  $T$ 에 대한 컴퓨터 프로그램의 성능을  $P$ 로 측정했을 때 경험  $E$ 로 인해 성능이 향상됐다. 이 컴퓨터 프로그램은 작업  $T$ 와 성능 측정  $P$ 에 대해 경험  $E$ 로 학습한 것 "
    - 톰 미첼(Tom Mitchell), 1997
- 스팸 필터는(사용자가 스팸이라고 지정한) 스팸 메일과 일반 메일의 샘플을 이용해 스팸 메일 구분법을 배울 수 있는 머신러닝 프로그램
  - 작업  $T$ 는 새로운 메일이 스팸인지 구분하는 task
  - 경험  $E$ 는 훈련 데이터(training data)
  - 성능측정  $P$ 는 직접 정의해야 하며, 이 성능 측정을 정확도 (accuracy)라고 부르며 분류 작업에 자주 사용됨
- 데이터마이닝 : 머신러닝 기법을 적용하여 대용량의 데이터로부터 숨겨진 패턴을 발견하는 학문

# 머신러닝 정의

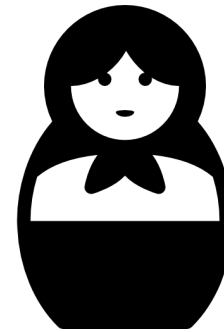
머신 러닝(Machine Learning), 인공 지능(Artificial Intelligence), 딥러닝(Deep Learning)



인공지능  
(AI)



머신러닝  
(ML)



딥러닝  
(DL)

# 머신러닝 정의

▪ Machine Learning : 단어 그대로 **기계를 학습**한다

① 머신이란?

인간이 제공한 데이터 간의 관계를 표현할 수 있는 모델 (=함수)

② 학습이란?

데이터를 가장 잘 표현할 수 있는 모델을 찾는 것 (=모델의 파라미터 최적화)

③ HOW?

통계적인 방법 혹은 경사하강법을 이용해 최적의 파라미터를 찾음

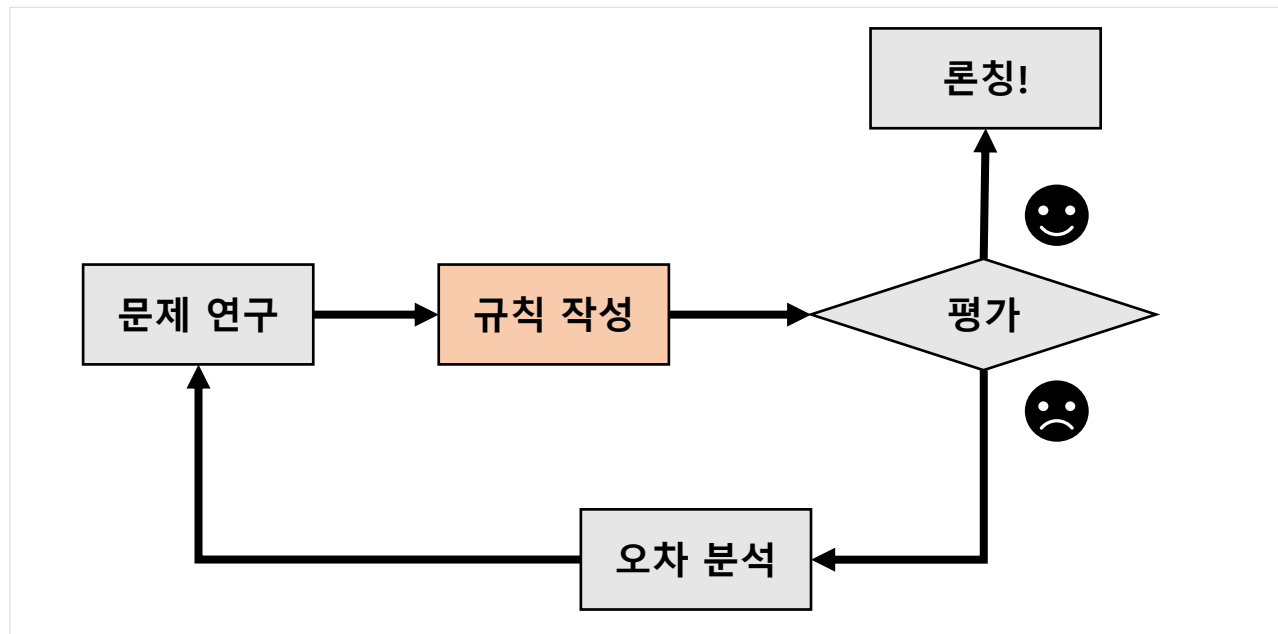
# 머신러닝 정의

- 어떤 형태의 데이터가 머신에게 주어지는 지에 따라 다음의 세부 분야들로 분류됨



# 머신러닝 특징

- 전통적 프로그래밍 기법으로는 규칙이 점점 길고 복잡해지므로 유지 보수하기 매우 어려움
  - 머신러닝 기법에 기반을 둔 스팸 필터는 일반 메일에 비해 스팸에 자주 나타나는 **패턴을 감지**하여 어떤 단어와 구절이 스팸 메일을 판단하는데 좋은 기준인지 **자동으로 학습**
- 전통적인 방식으로는 너무 복잡하거나 알려진 알고리즘이 없는 분야(예:음성인식)

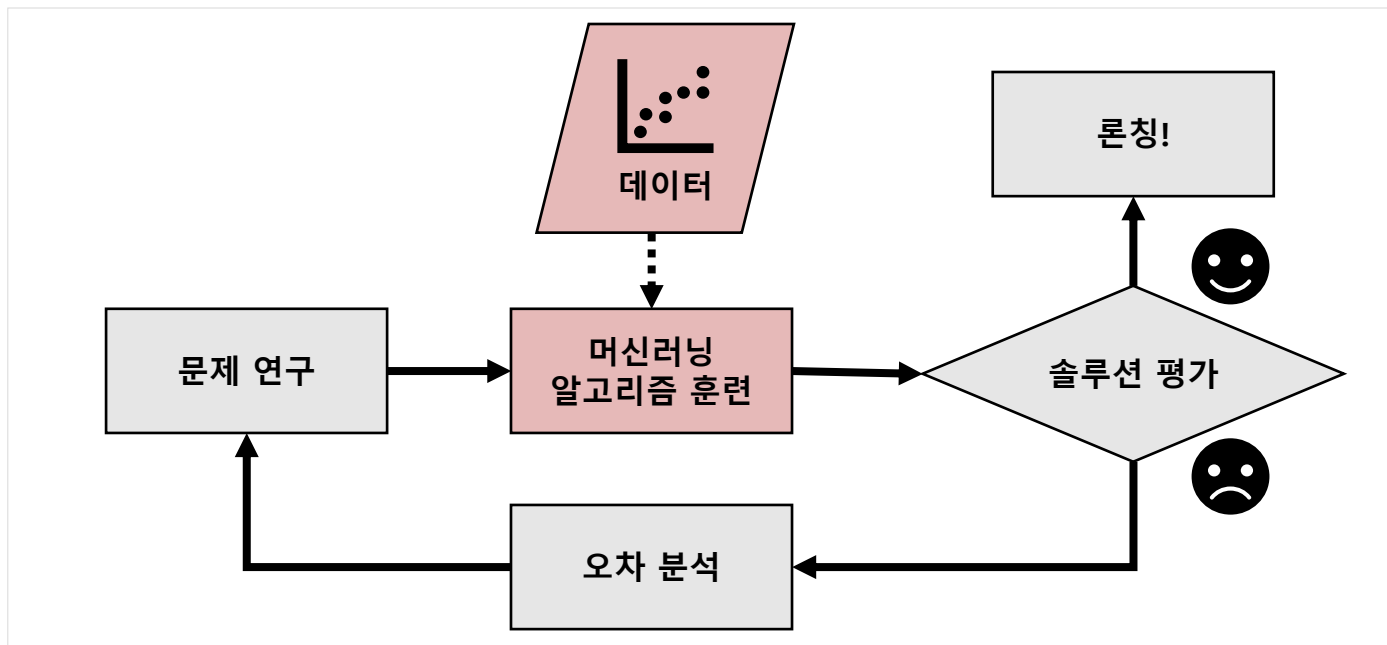


전통적인 접근방법

# 머신러닝 특징

## ■ 머신러닝의 장점

- 기존 솔루션으로는 많은 수동 조정과 규칙이 필요한 문제: 하나의 머신러닝 모델이 코드를 간단하게 만들고 전통적인 방법보다 더 잘 수행되도록 할 수 있음
- 전통적인 방식으로는 해결 방법이 없는 복잡한 문제: 가장 뛰어난 머신러닝 기법으로 해결 방법을 찾을 수 있음
- 유동적인 환경: 머신러닝 시스템은 새로운 데이터에 적응 가능
- 복잡한 문제와 대량의 데이터로부터 자동적으로 통찰 얻을 수 있음



머신러닝 접근방법

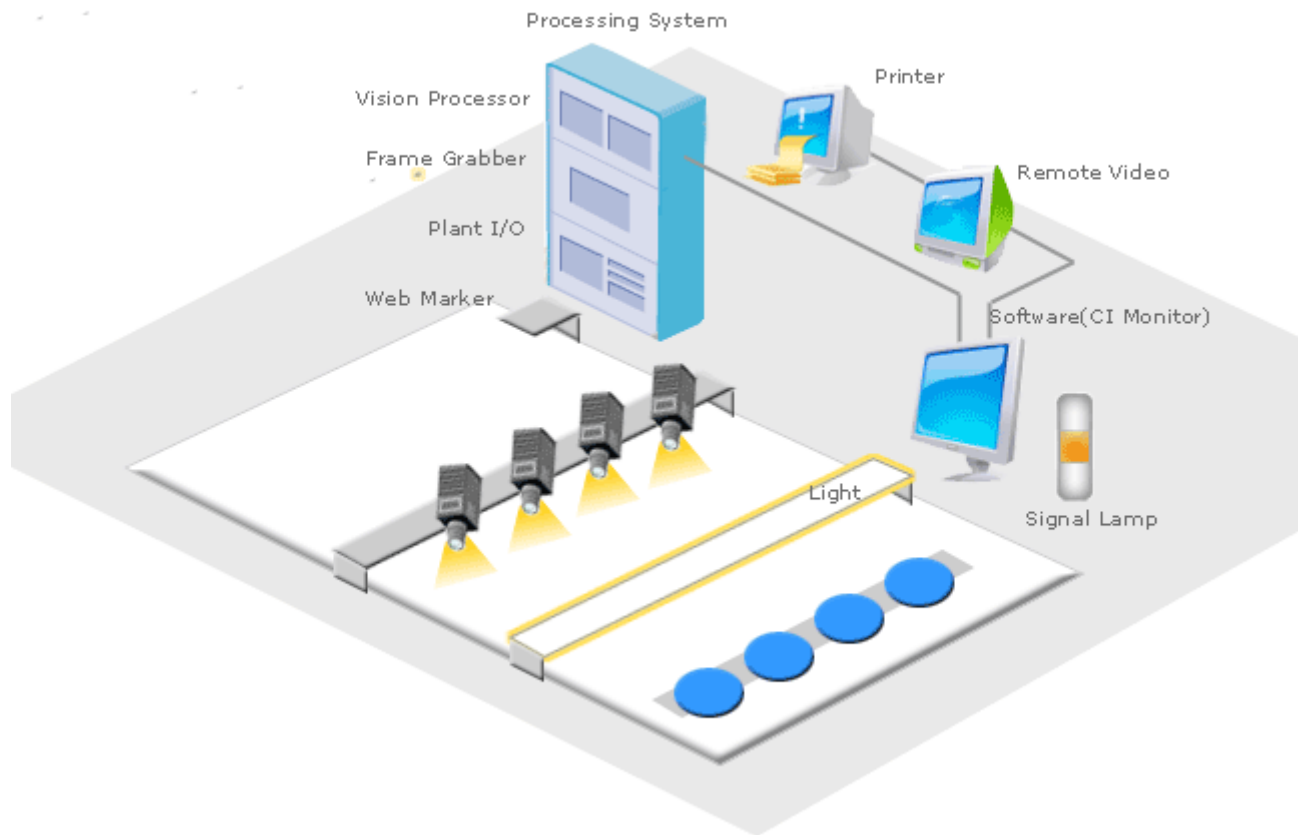


# 머신러닝 사례

- 이미지 분류 작업: 생산 라인에서 제품 이미지를 분석해 자동으로 분류
- 시맨틱 분할 작업: 뇌를 스캔하여 종양 진단
- 텍스트 분류(자연어 처리): 자동으로 뉴스 기사 분류
- 텍스트 분류: 토론 포럼에서 부정적인 코멘트를 자동으로 구분
- 텍스트 요약: 긴 문서를 자동으로 요약
- 자연어 이해: 챗봇(chatbot) 또는 개인 비서 만들기
- 회귀 분석: 회사의 내년도 수익을 예측하기
- 음성 인식: 음성 명령에 반응하는 앱
- 이상치 탐지: 신용 카드 부정 거래 감지
- 군집 작업: 구매 이력을 기반으로 고객을 나누고 각 집합마다 다른 마케팅 전략을 계획
- 데이터 시각화: 고차원의 복잡한 데이터셋을 명확하고 의미 있는 그래프로 표현하기
- 추천 시스템: 과거 구매 이력을 기반으로 고객이 관심을 가질 수 있는 상품 추천하기
- 강화 학습: 지능형 게임 봇(bot) 만들기

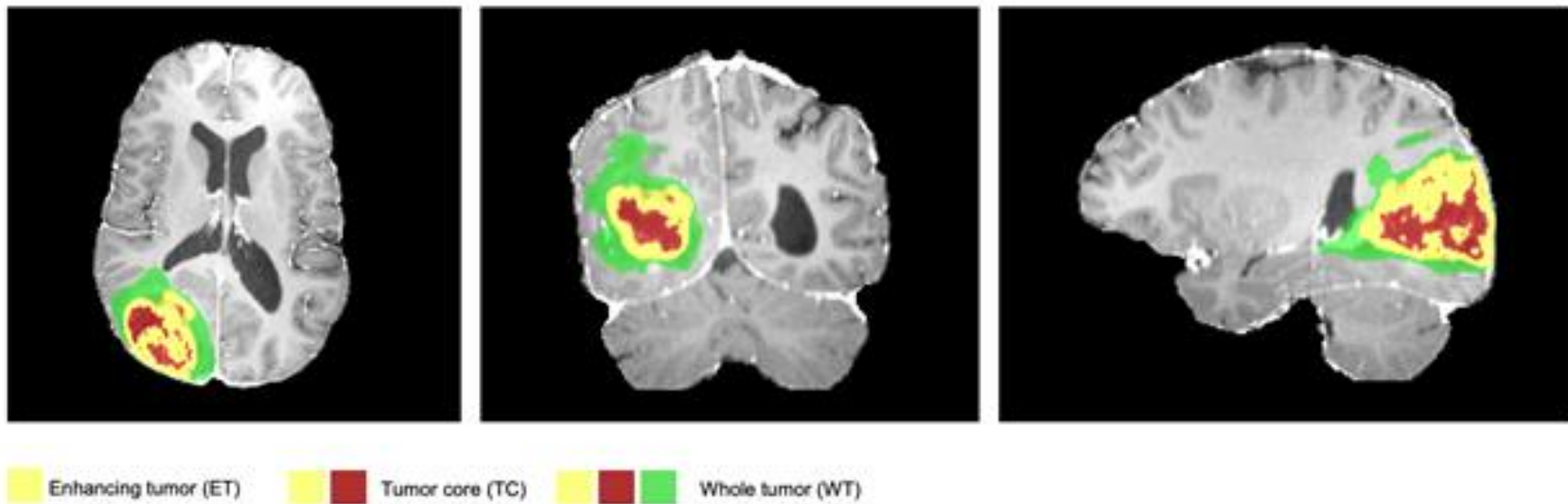
# 머신러닝 사례

- 이미지 분류 작업: 생산 라인에서 제품 이미지를 분석해 자동으로 분류



# 머신러닝 사례

- 시맨틱 분할 작업: 뇌를 스캔하여 종양 진단



# 머신러닝 사례

## ■ 텍스트 분류(자연어 처리): 자동으로 뉴스 기사 분류

뉴스홈 | 최신기사

## 연합뉴스, 인공지능 기사요약 서비스 첫선

송고시간 | 2021-01-11 14:52      日本語



서명덕 기자  
기자페이지

**학습·훈련 거친 AI가 '세 줄 요약문' 하루 900여 건 생성  
중인터넷과 공동 개발...구글 딥러닝 기술 'BERT' 응용**

(서울=연합뉴스) 서명덕 기자 = 연합뉴스가 국내 언론 최초로 인공지능으로 요약하는 '세 줄 요약' 서비스를 11일 시작했다.

연합뉴스는 검색포털 '줌'을 운영하는 중인터넷과 공동으로 기사 요약 AI 기술을 구현해 홈페이지 주요 기사를 대상으로 요약문 서비스를 선보였다.

기사 요약문 서비스는 연합뉴스 주요 기사를 대상으로 하루 평균 900여 건 실시간 제공되며 각 기사 본문 상단에서 아이콘을 눌러 내용을 확인할 수 있다.

요약

✕

연합뉴스가 국내 언론 최초로 인공지능(AI)을 통해 기사 본문을 자동으로 요약하는 '세 줄 요약' 서비스를 11일 시작했다.

연합뉴스는 검색포털 '줌'을 운영하는 중인터넷과 공동으로 기사 요약 AI 기술을 구현해 홈페이지 주요 기사를 대상으로 요약문 서비스를 선보였다.

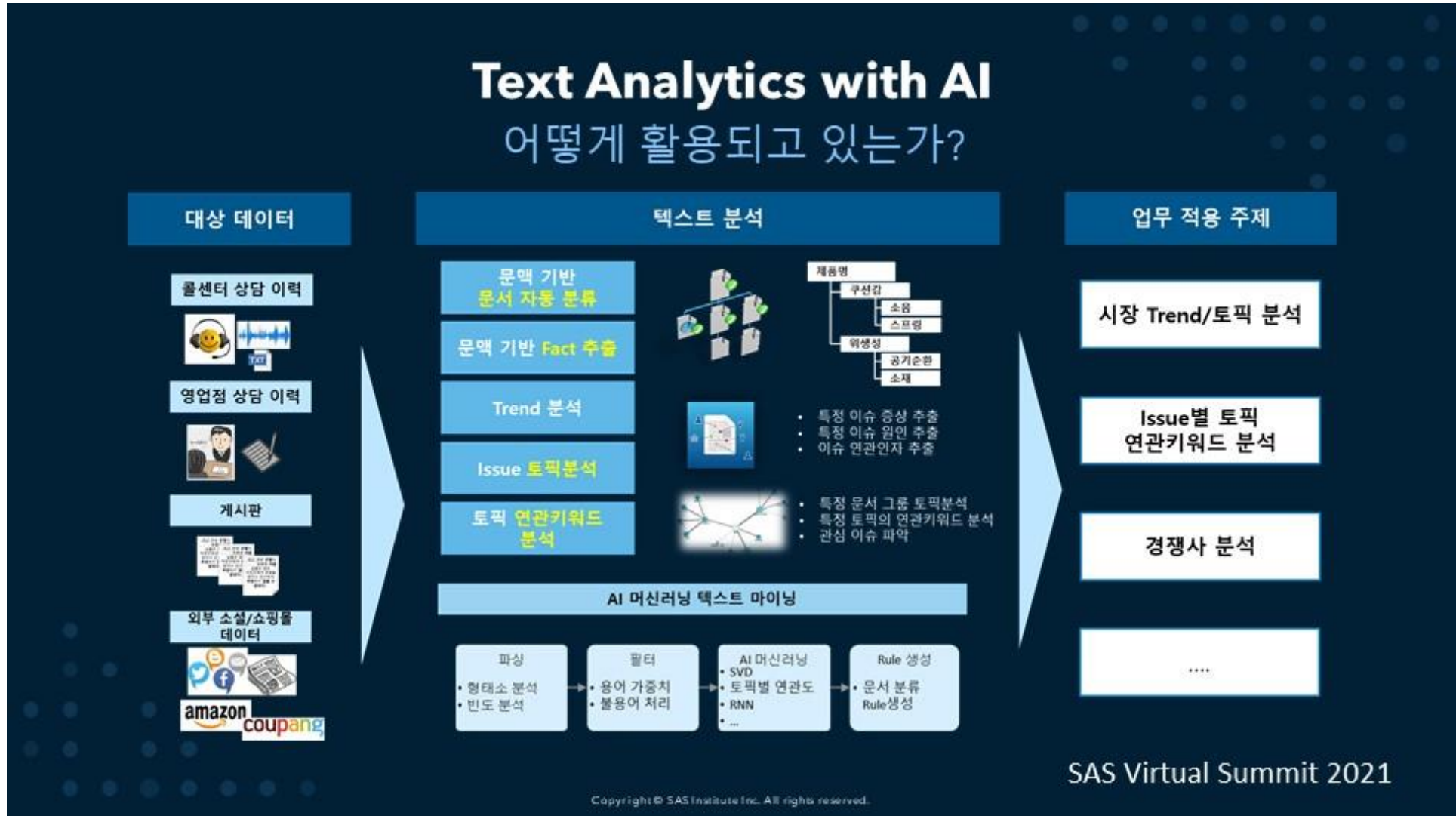
검색 포털이 국내 언론사와 공동 협력하여 딥러닝 연구 결과물을 상용화한 것은 이번이 처음이다.

① 인공지능이 자동으로 줄인 '세 줄 요약' 기술을 사용합니다. 전체 내용을 이해하기 위해서는 기사 본문과 함께 읽어야 합니다. 제공 - 연합뉴스&중인터넷

합뉴스 →

# 머신러닝 사례

- 텍스트 분류: 토론 포럼에서 부정적인 코멘트를 자동으로 구분



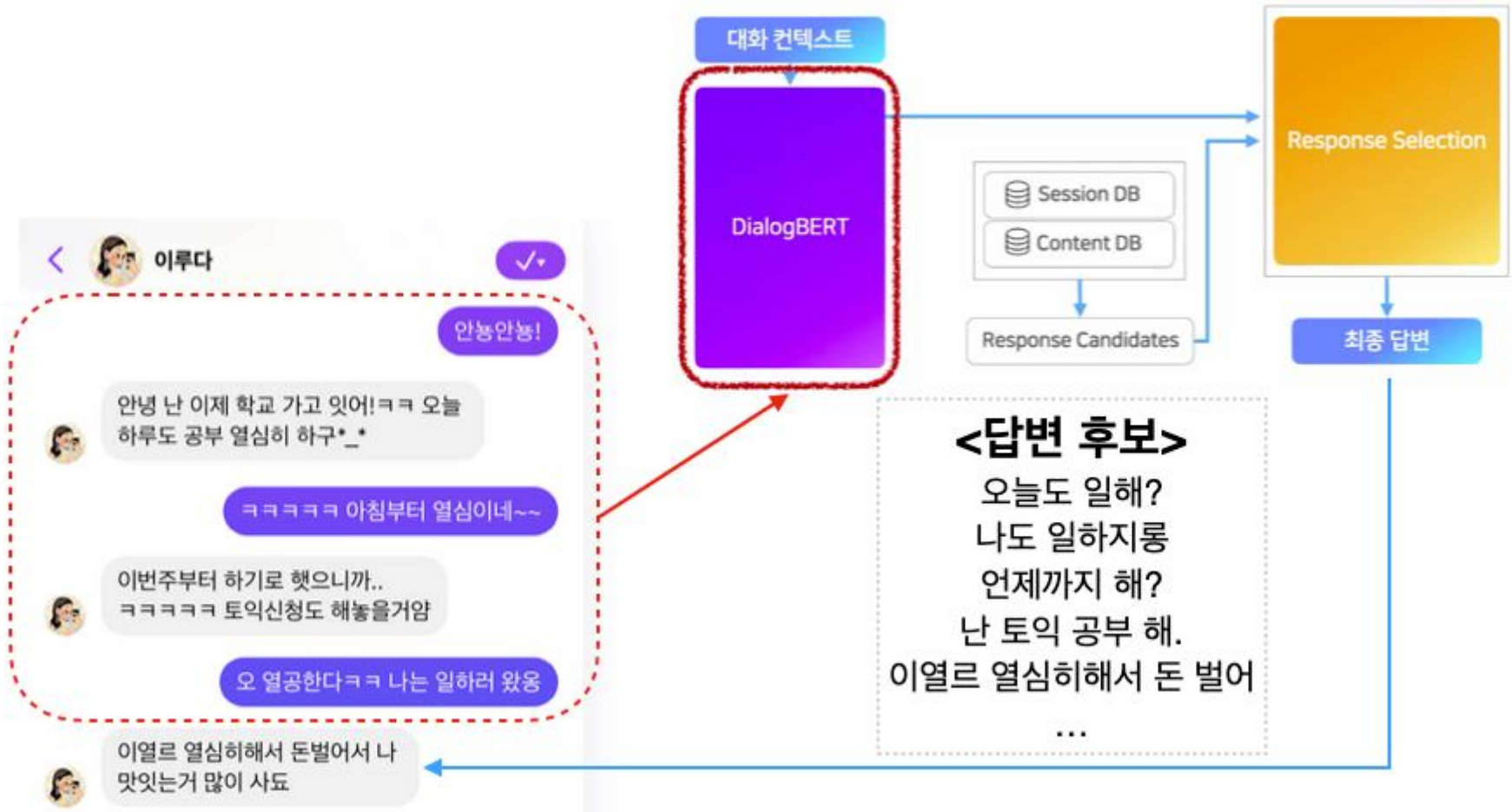
# 머신러닝 사례

- 텍스트 요약: 긴 문서를 자동으로 요약



# 머신러닝 사례

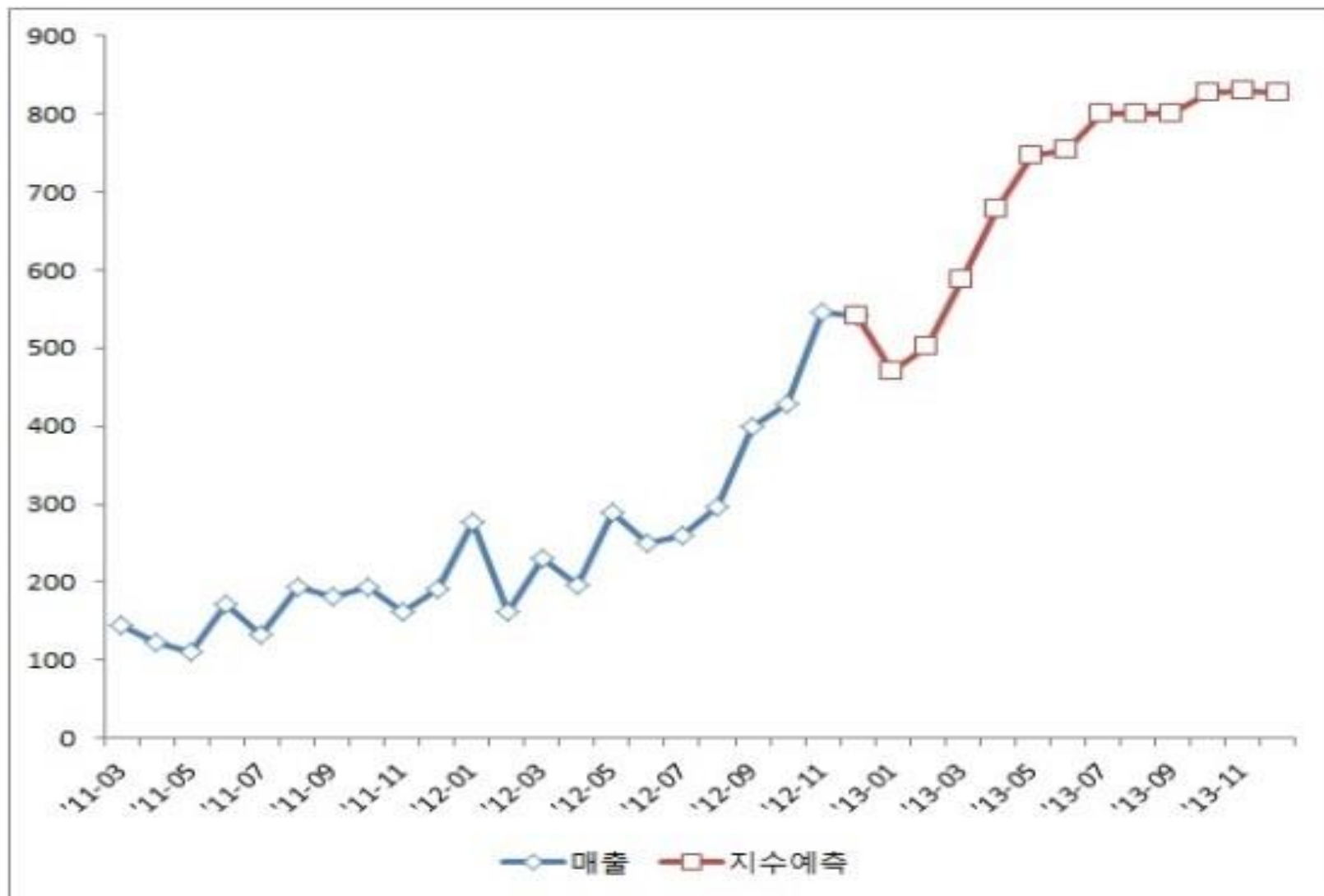
## ■ 자연어 이해: 챗봇(chatbot) 또는 개인 비서 만들기





# 머신러닝 사례

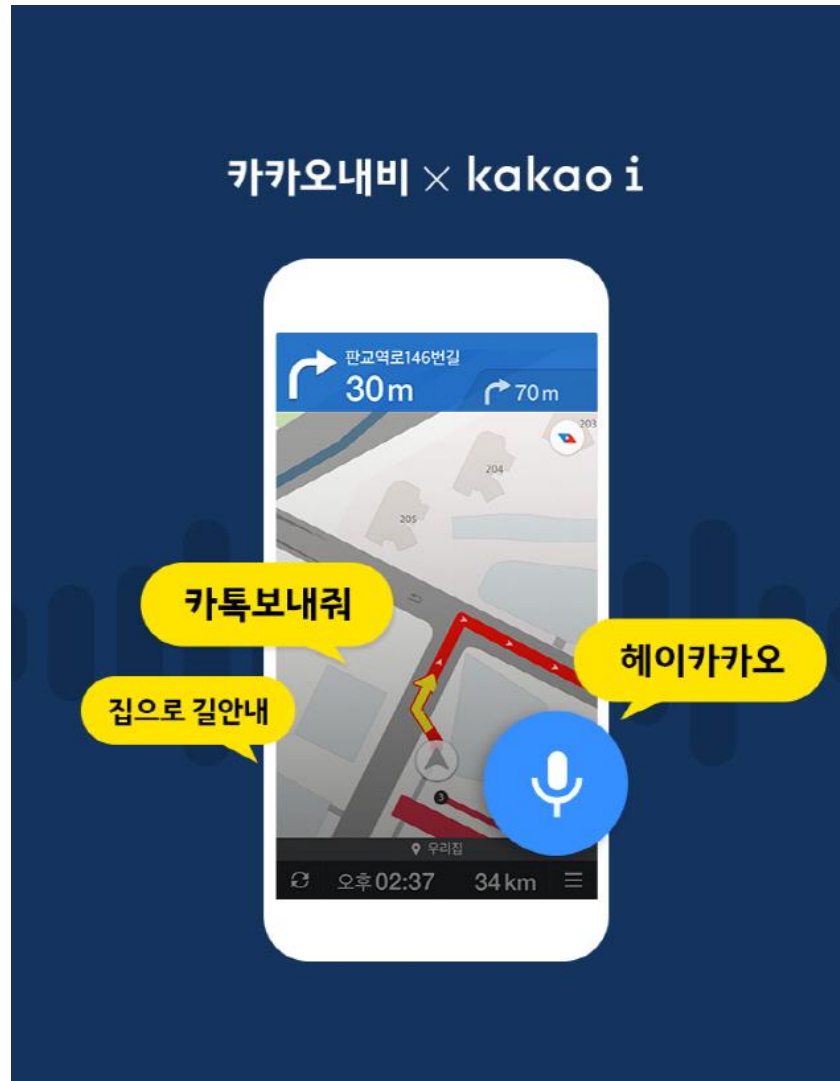
## ■ 회귀 분석: 회사의 내년도 수익을 예측하기





# 머신러닝 사례

- 음성 인식: 음성 명령에 반응하는 앱




# 머신러닝 사례

- 이상치 탐지: 신용 카드 부정 거래 감지



# 머신러닝 사례

- 군집 작업: 구매 이력을 기반으로 고객을 나누고 각 집합마다 다른 마케팅 전략을 계획

<b>문제제기</b>	유사한 특성을 가지고 있는 고객들끼리 어떻게 같은 군집으로 분류할 것인가?
	

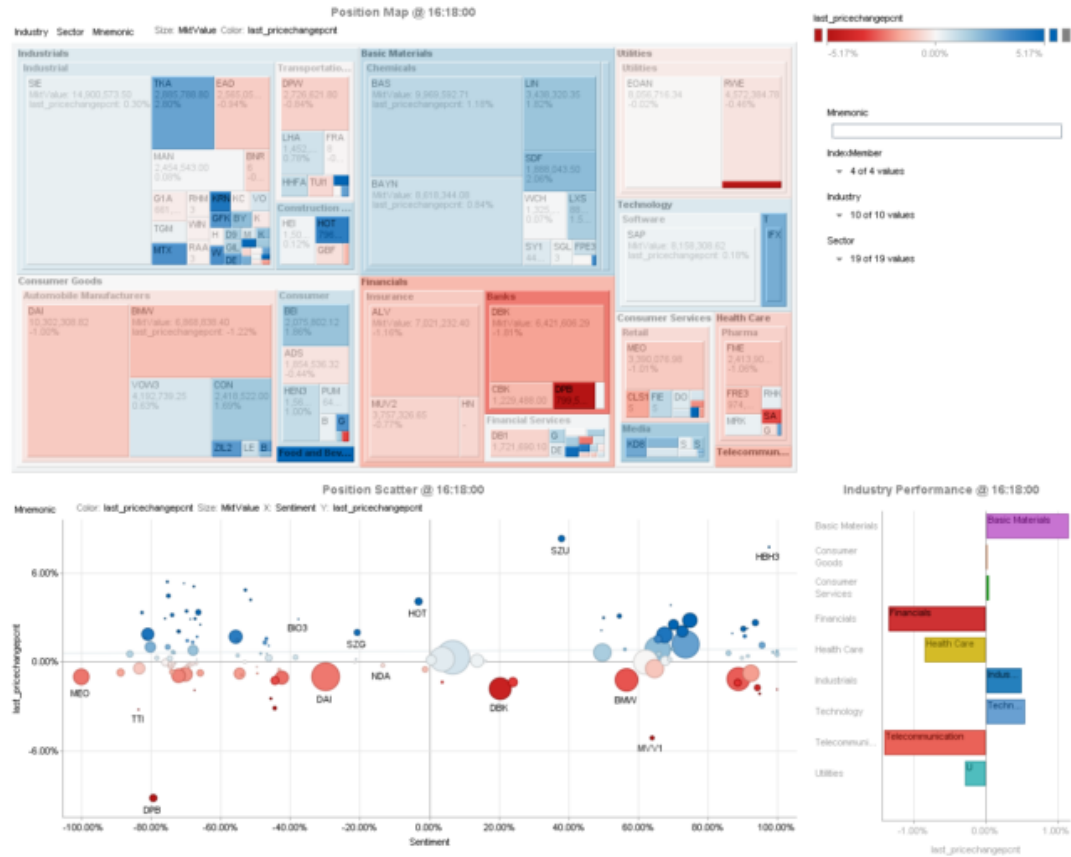
<b>분석결과</b>	서로 유사한 특성을 가지는 고객군 군집도출	
		
집단1	집단2	집단3

<b>활용1</b>	고객의 세분화
고객집단별 인구통계학적·사회적·행태적 특성 파악 집단의 프로파일 작성 및 특성파악	

<b>활용2</b>	군집별로 추가분석을 함
자료의 전반적인 분포형태를 파악함 군집별로 추가적인 분석을 함	

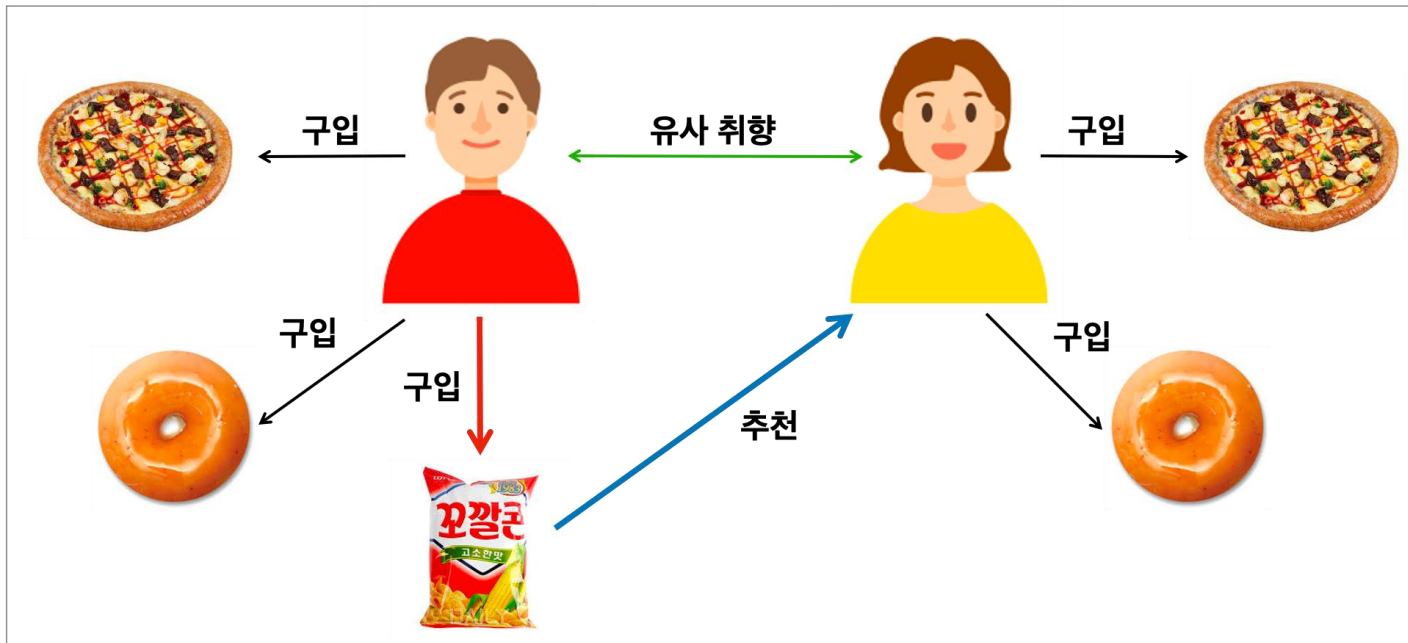
# 머신러닝 사례

- 데이터 시각화: 고차원의 복잡한 데이터셋을 명확하고 의미 있는 그래프로 표현하기



# 머신러닝 사례

- 추천 시스템: 과거 구매 이력을 기반으로 고객이 관심을 가질 수 있는 상품 추천하기



# 머신러닝 사례

- 강화 학습: 지능형 게임 봇(bot) 만들기



# 머신러닝 알고리즘 종류

## ■ 넓은 범주의 분류

- 감독하에 훈련하는 것인지 그렇지 않은 것인지: 지도, 비지도, 강화 학습
- 실시간으로 점진적인 학습을 하는지 아닌지: 온라인 학습과 배치 학습
- 단순히 알고 있는 데이터 포인트와 새 데이터 포인트를 비교하는 것인지 아니면 훈련 데이터셋에서 패턴을 발견하여 예측 모델을 만드는지: 사례 기반 학습과 모델 기반 학습

## ■ 지도 학습과 비지도 학습

- 지도 학습
- 비지도 학습
- 강화학습

## ■ 배치 학습과 온라인 학습

- 배치 학습
- 온라인 학습

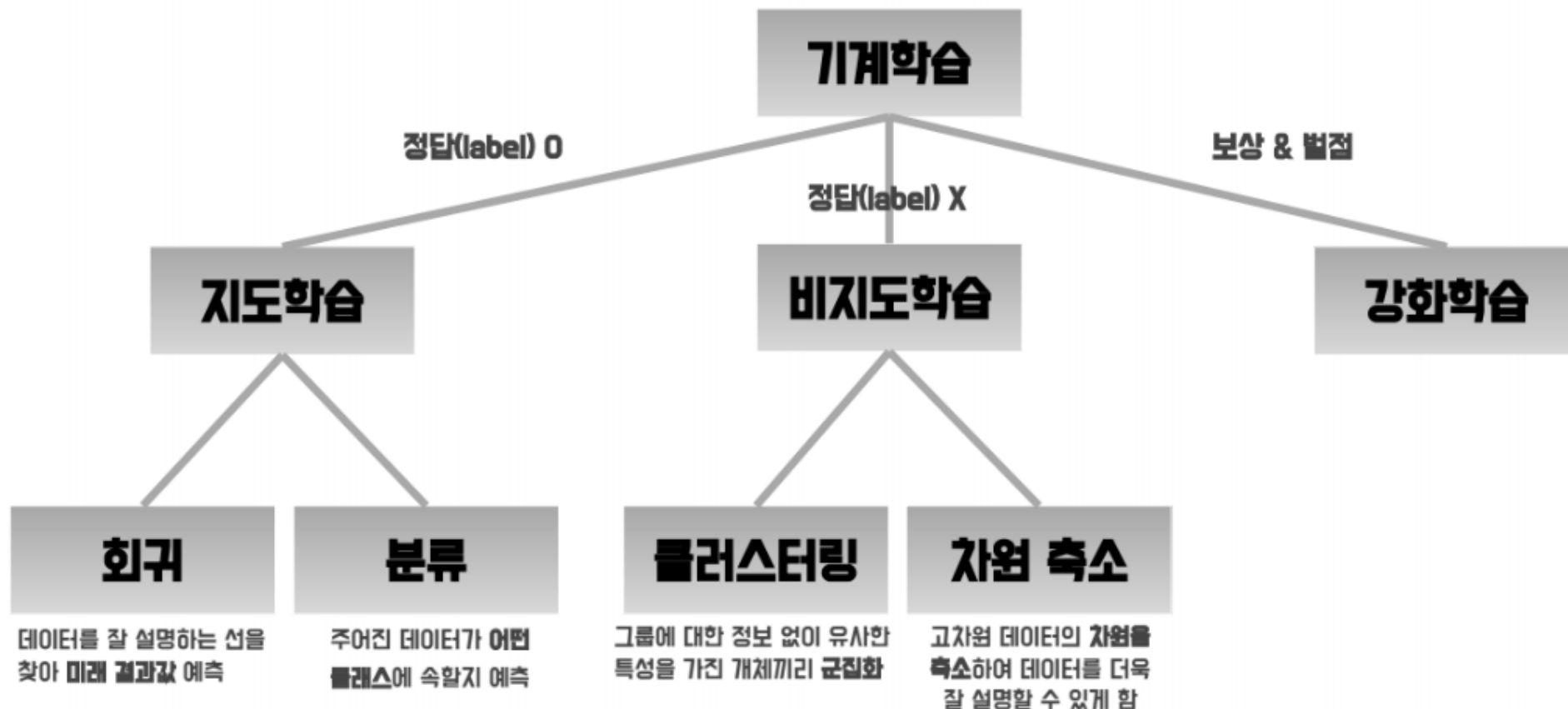
## ■ 사례 기반 학습과 모델 기반 학습

- 사례 기반 학습
- 모델 기반 학습

# 머신러닝 알고리즘 종류

## 머신러닝 분류

⇒ 훨씬 더 많은 종류가 있지만, 대표적인 기법을 바탕으로 큰 틀을 이해하자!





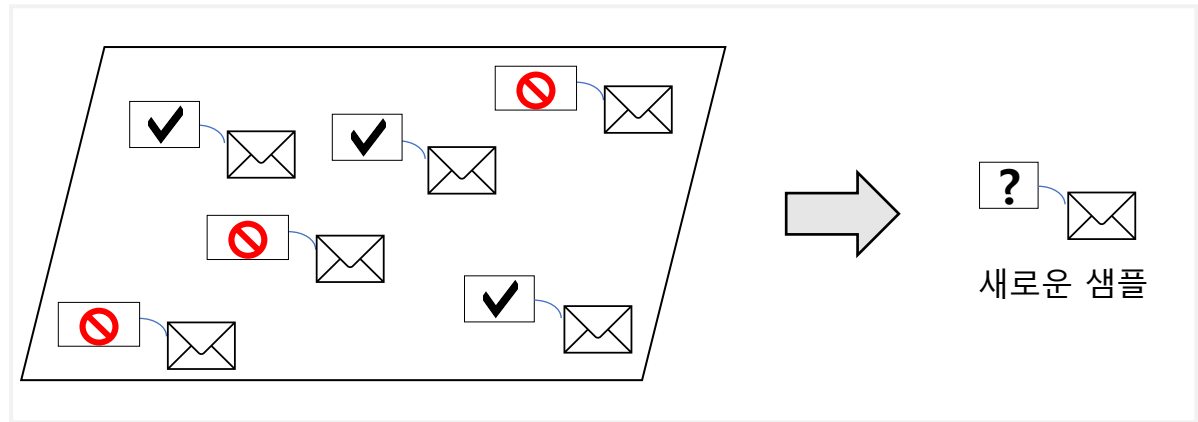
# 머신러닝 알고리즘 종류

## ■ 지도학습

- 알고리즘에 주입하는 훈련 데이터에 레이블(label)이라는 원하는 답이 포함되어 있음
  - 회귀(예측)
  - 분류

## ■ 대표적인 지도 학습 알고리즘

- K-최근접 이웃 (K-NN)
- 선형회귀
- 로지스틱 회귀
- 서포트 벡터 머신
- 결정 트리와 랜덤 포레스트
- 신경망



지도학습 사례

# 머신러닝 알고리즘 종류

## ■ 회귀(Regression)

1. 입력값 : 연속값(실수형), 이산값(범주형) 등 모두 가능
2. 출력값 : 연속값(실수형)
3. 모델 형태 : 일반적인 함수 형태 (*eg.*  $y = w_1x + w_0$ )

## ■ 분류(Classification)

1. 입력값 : 연속값(실수형), 이산값(범주형) 등 모두 가능
2. 출력값 : 이산값(범주형)
3. 모델 형태 : 이진 분류라면 Sigmoid 함수,  
다중 분류라면 Softmax 함수 꼭 포함

# 머신러닝 알고리즘 종류

## ■ 지도 학습 - 회귀(예측) 예제

### ■ 와인품질 측정 방식

- Orley Ashenfelter(1999)
- 프랑스 보르도 지방 날씨 데이터 분석하여 와인품질 공식 완성
- $\text{품질} = 12.145 + 0.00117 * \text{겨울 강우량} + 0.06140 * \text{생장기 평균 기온} - 0.00389 * \text{추수기 강우량}$
- 전문가들에 의해 점수화된 품질값(y) 존재
- 기존 전문가들의 평가와 와인 품질 측정 방식의 장단점?

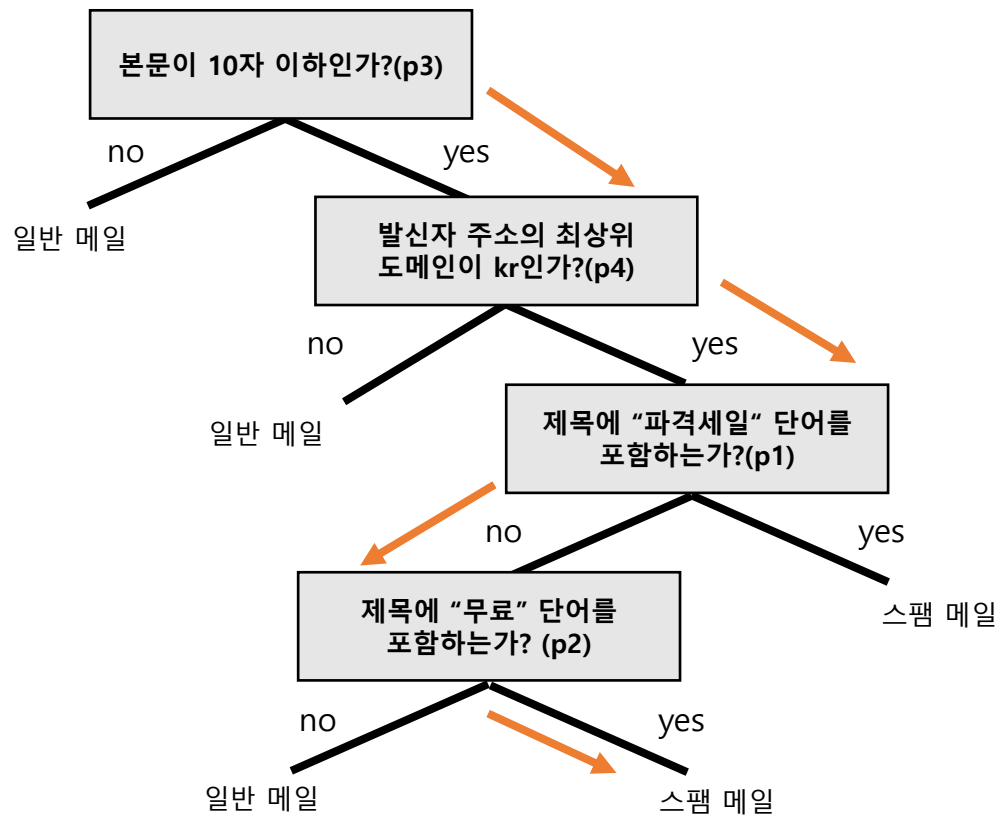


# 머신러닝 알고리즘 종류

## ■ 지도 학습 - 분류 예제

### ■ 스팸메일 분류 문제

- 메일의 내용을 분석하여 일반 메일과 스팸메일 분류



# 머신러닝 알고리즘 종류

## ■ 비지도 학습

- 훈련 데이터에 레이블이 없어서, 시스템이 아무런 도움 없이 학습
- 대표적인 비지도 학습 알고리즘

- 군집

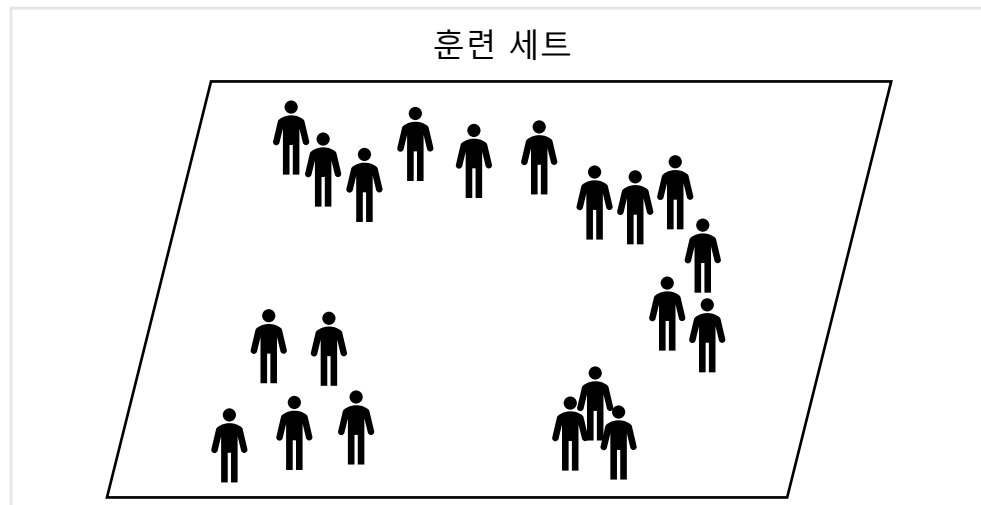
- K-평균
- DBSCAN
- 계층 군집 분석
- 이상치 탐지와 특이치 탐지
- One-class SVM
- Isolation Forest

- 차원 축소

- 주성분 분석(PCA)
- 커널 PCA
- 지역적 선형 임베딩
- t-SNE

- 연관성 규칙 학습

- 어프라이어리(Apriori)
- 이클렛(Eclat)



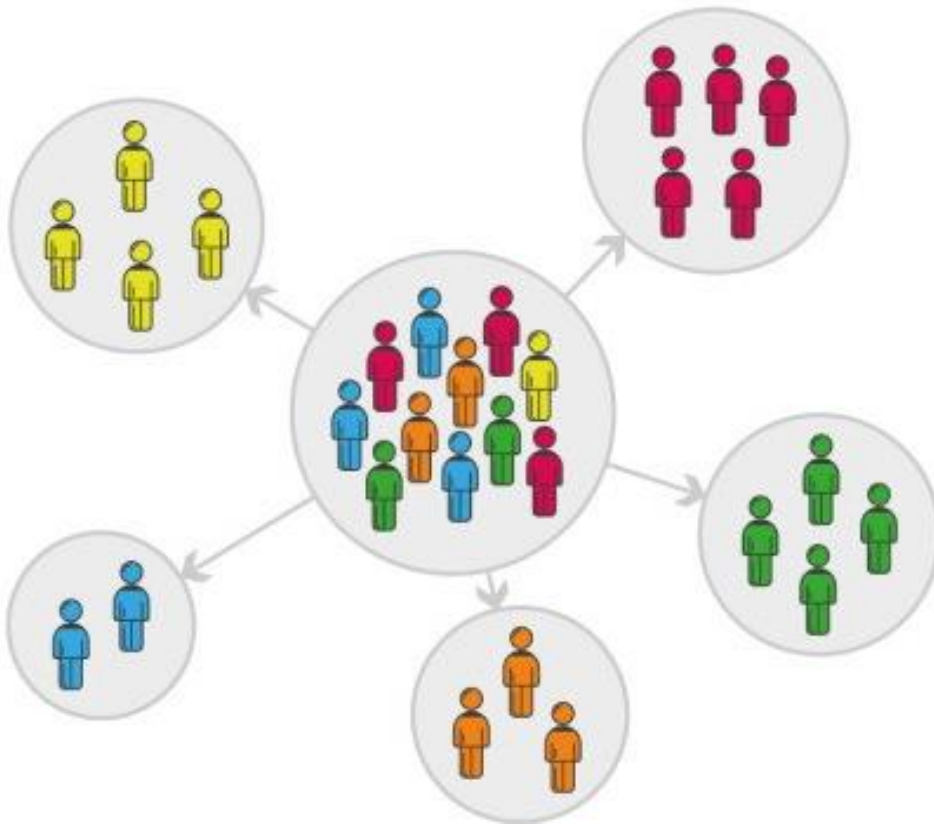
비지도학습 사례

# 머신러닝 알고리즘 종류

## ■ 비지도학습 예제

### ■ 군집분석

- 유사한 고객들로 군집화를 수행하여 고객을 세분화



빠른 배송 및 편리한 주문에 관심 지속적인  
관계유지, 많은 추천

브랜드 추구, 가격에 덜 민감,  
가장 높은 수입, 보통 남성

배송시간, 제품의 손상가능성에  
덜 민감, 낮은 소비 증가폭

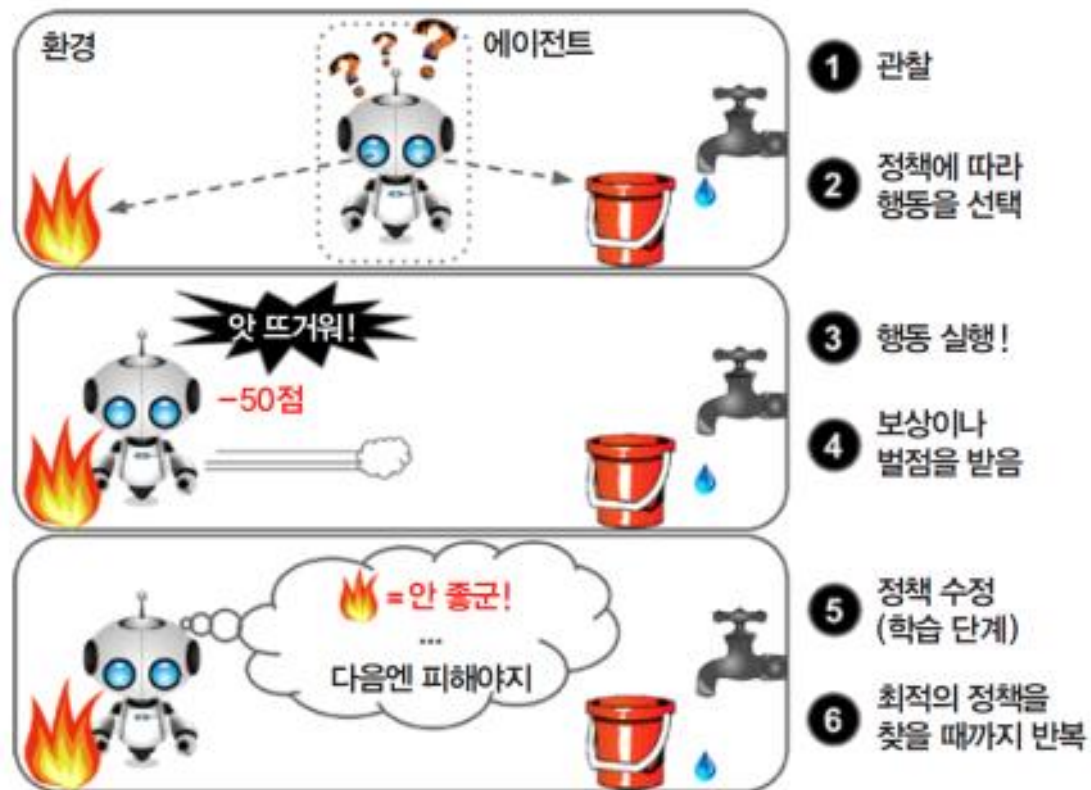
프로모션 등의 영향을 많이 받음  
적은 수입, 추천 거의 없음

가격이 중요, 구매 적음  
제품의 손상가능성은 중요하지 않음

# 머신러닝 알고리즘 종류

## ■ 강화 학습

- 주어진 환경에서 에이전트가 최대의 보상(최소의 벌점)을 얻기 위해 최상의 정책을 학습



# 머신러닝 알고리즘 종류

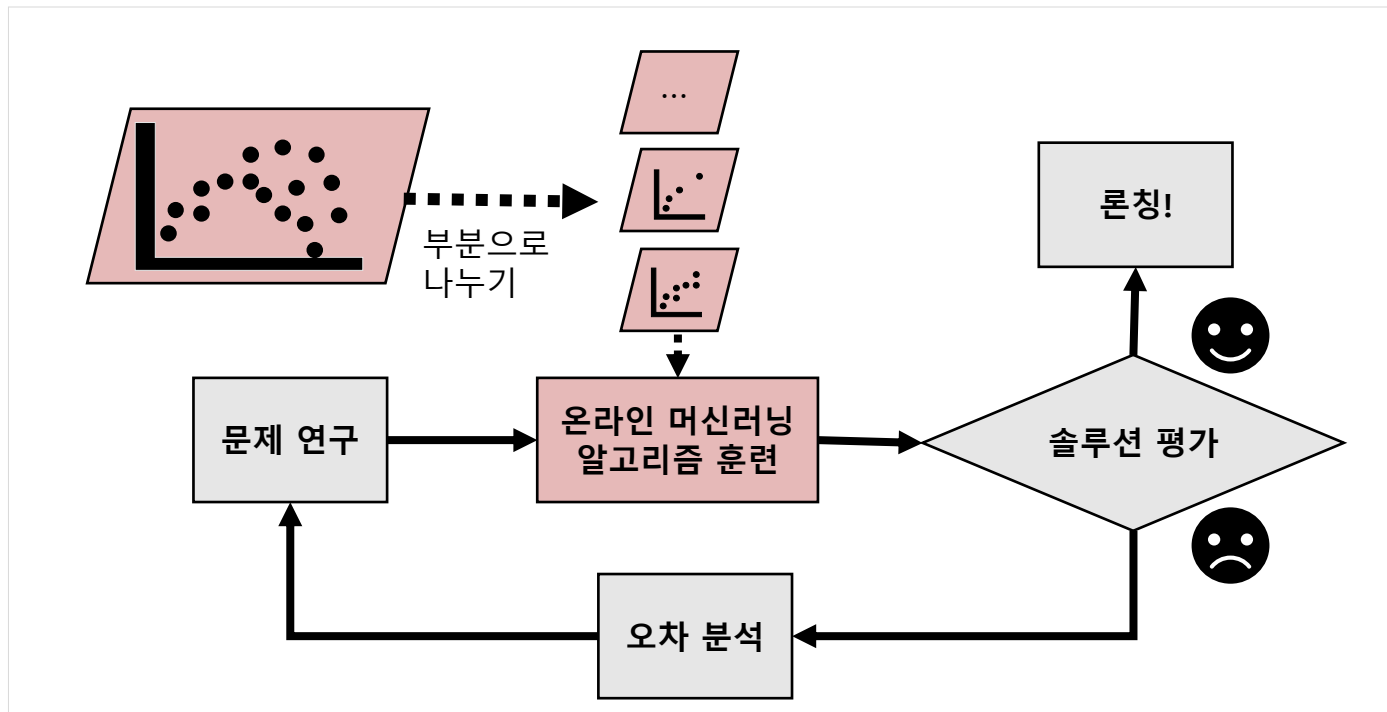
## ■ 배치 학습과 온라인 학습

### ■ 배치 학습

- 시스템이 가용한 모든 데이터를 이용하여 학습하는 오프라인 학습법
- 새로운 데이터가 들어오게 되면 이전데이터와 합해서 모델을 새롭게 학습

### ■ 온라인 학습

- 데이터를 순차적으로 한 개씩 또는 미니배치(mini-batch)라 부르는 작은 묶음 단위로 주입하여 시스템을 훈련



온라인 학습 사례



# 머신러닝 알고리즘 종류

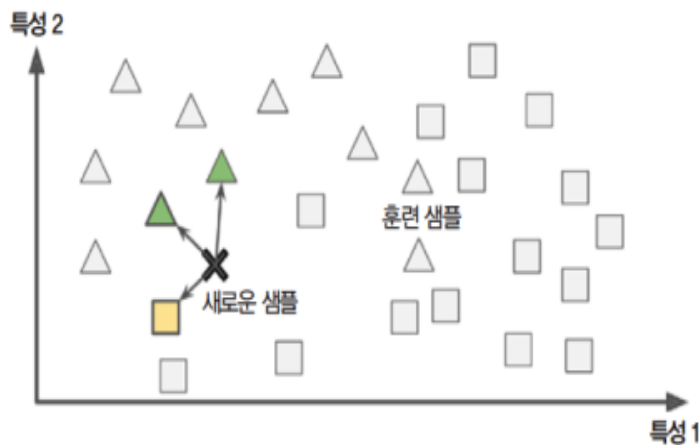
## ■ 사례 기반 학습과 모델 기반 학습

### ■ 사례 기반 학습

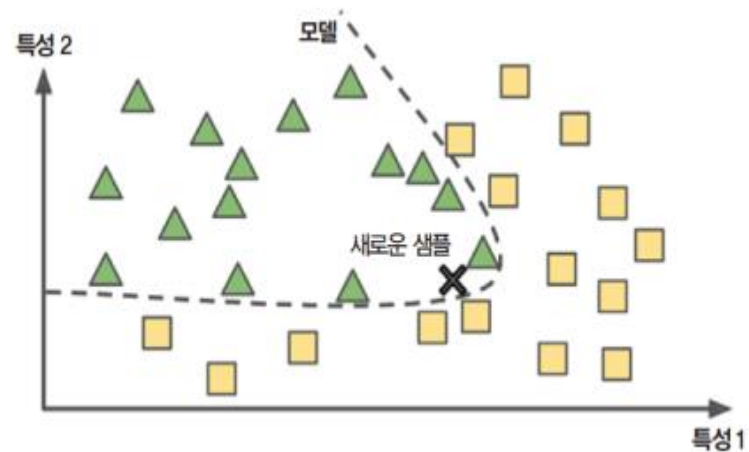
- 훈련데이터에서 가장 가까운(유사한) 샘플을 찾아서 이를 바탕으로 신규 데이터의 예측 수행

### ■ 모델 기반 학습

- 데이터를 이용하여 모델을 만들고 이를 바탕으로 신규 데이터의 예측을 수행(대부분의 알고리즘)



사례 기반 학습



모델 기반 학습

# 머신러닝 프로세스

## ■ 프로세스

단계	설명
비즈니스 이해	<ul style="list-style-type: none"><li>- 비즈니스 목표를 이해하고, 이를 데이터 수집 목표로 정의</li><li>- 비즈니스에 영향을 주는 중요한 항목 도출</li></ul>
데이터 이해	<ul style="list-style-type: none"><li>- 초기 데이터를 수집하고, 데이터의 품질 정의</li><li>- 가설을 위한 데이터 셋 정의</li></ul>
데이터 준비	<ul style="list-style-type: none"><li>- 분석 모델링에 필요한 데이터 추출</li></ul>
모형	<ul style="list-style-type: none"><li>- 분석 기법을 선택하고, 분석에 필요한 최적 변수 설정</li><li>- 분석 모델 구축</li></ul>
평가	<ul style="list-style-type: none"><li>- 분석 모델에 대해 평가하고, 비즈니스 목표를 달성할 분석 모델 선정</li><li>- 전체 프로세스를 재검토하고, 다음 단계를 결정</li></ul>
적용	<ul style="list-style-type: none"><li>- 분석 모델링을 통해 획득한 지식 가공</li><li>- 보고서 작성 및 시각화</li></ul>

# 머신러닝 알고리즘

## ■ 용어정리 (1)

알고리즘 (algorithm)	특정 데이터마이닝 기법, 예를 들어 분류나무, 판별분석 등을 실행하기 위해 사용되는 특정 절차
독립변수 (independent variable)	보통 X로 표기되며, 속성(attribute), 특성(feature), 예측변수(predictor), 입력변수(input variable) 또는 데이터베이스 관점에서 필드(field)라고도 함
종속변수 (dependent variable)	보통 Y로 표기되며, 지도학습으로 예측되는 변수, 반응변수(response variable), 출력변수(output variable), 목표변수 또는 성과변수라고도 함
변수 (variable)	입력변수(X)와 출력변수(Y)를 모두 포함하는 레코드의 측정치를 말함
관측치 (observation)	고객,거래 등의 측정치를 갖는 분석의 단위로서 사례(case), 레코드(record), 패턴, 또는 행(row)이라고도 함(각 행은 레코드를, 각 열은 변수를 의미)
차원 (dimension)	(독립) 변수의 개수

# 머신러닝 알고리즘

## ■ 용어정리 (2)

	성별	학력	키	몸무게	자녀수	연소득
person1	남성	중졸	180	102	0	4000
person2	여성	고졸	150	43	1	5000
person3	남성	대졸	170	52	1	7000
person4	여성	대학원졸	160	80	2	6500

$$\text{연소득} = \alpha X_{\text{성별}} + \beta X_{\text{학력}} + \gamma X_{\text{키}} + \delta X_{\text{몸무게}} + \varepsilon X_{\text{자녀수}}$$

# 머신러닝 알고리즘

## ■ 용어정리 (3)

### ■ 데이터의 구성

- 데이터는 feature과 label로 구성됨
- 이는 독립변수와 종속변수로도 불림
- 라벨은  $y$ 로 표기하며,  
라벨의 유무로 지도/비지도 학습을 구분함

	혈압	몸무게	나이	지병
길동	130	72	17	N
철수	120	82	34	Y
...	...	...	...	...
영희	150	58	31	N

### ■ Feature (=attribute, 항목)

- 데이터  $x$ 의 특징 혹은 항목을 의미
  - $N$ : 데이터의 샘플개수
  - $D$ : 피처의 개수
- 예) 혈압, 몸무게, 나이

$$X = \begin{bmatrix} x_{11} & \cdots & x_{1D} \\ \vdots & \ddots & \vdots \\ x_{N1} & \cdots & x_{ND} \end{bmatrix}$$

$$x_1 = \begin{bmatrix} x_{11} \\ \vdots \\ x_{N1} \end{bmatrix}$$

$$x_p = \begin{bmatrix} x_{1D} \\ \vdots \\ x_{ND} \end{bmatrix}$$

# 머신러닝 알고리즘

## ■ 용어정리 (4)

- Parameter (=weight, 가중치)
  - 주어진 데이터(입력값) 말고, 모델이 가지고 있는 학습 가능한 (learnable) 파라미터  
ex)  $w_0, w_1, \dots, w_D$
- Hyperparameter (하이퍼 파라미터)
  - 모델 학습에 있어, 인간이 정해야하는 변수들
  - 학습률, 배치 크기 등

# 머신러닝 알고리즘

## ■ 용어정리 (5)

### ■ Input(입력값) vs. Output(출력값)

- Input : 모델(함수)에 입력되는 값으로 데이터의 피쳐 부분( $x$ 로 표기)
- Output : 모델로부터 출력되는 예측값 ( $\hat{y}$ 로 표기)

### ■ 선형 모델 vs. 비선형 모델

- Linear regression (선형 회귀) : 파라미터를 선형 결합식으로 표현  
ex)  $y = w_0 + w_1x_1 + w_2x_2 + \dots + w_Dx_D$ ,  $y = w_0 + w_1x + w_2x^2$
- Nonlinear regression (비선형 회귀) : 선형 결합식으로 표현 불가능한 모델  
ex)  $\log(y) = w_0 + w_1 \log(x)$ ,  $y = \max(x, 0)$

# 머신러닝 기초수학

## ■ 기초수학 (1) : 수치 대푯값

### ■ 평균

- 변량의 총합 ÷ 변량의 총개수
- 대푯값으로 가장 많이 쓰이는 값
- 10, 12, 11, 9, 30
- 평균은 극단값이 있을 경우 대푯값으로 적절하지 못함

### ■ 중앙값

- 자료값을 크기순으로 작은값부터 차례로 나열하여 한 가운데 있는 값
- 2, 3, 5, 5, 6 or 2, 3, 4, 5, 6, 6
- 평균의 단점을 극복

### ■ 최빈값

- 자료에 있는 값들 중에서 같은 값으로 제일 많이 나오는 자료값
- 2, 3, 3, 3, 5, 6, 6
- 최빈값이 많이 쓰이는 자료는 수자로 된 자료보다는 선호도와 같은 정성적 자료 등에 적합



# 머신러닝 기초수학

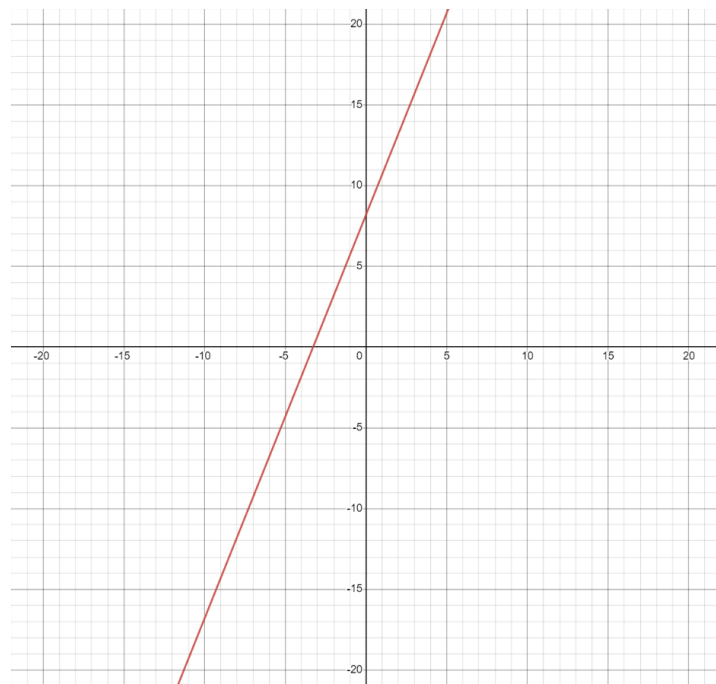
## ■ 기초수학 (2)

### ■ 함수

- 두 집합 사이의 관계 혹은 규칙
- $y = f(x)$ 의 식으로 표현, 이 때  $x$ 는 입력 값,  $y$ 는 출력값

### ■ 일차 함수

- $y$ 가  $x$ 에 대한 일차식으로 표현된 경우
- $y = ax + b$  ( $a \neq 0$ )
- $a$ 를 기울기,  $b$ 를 절편으로 표현

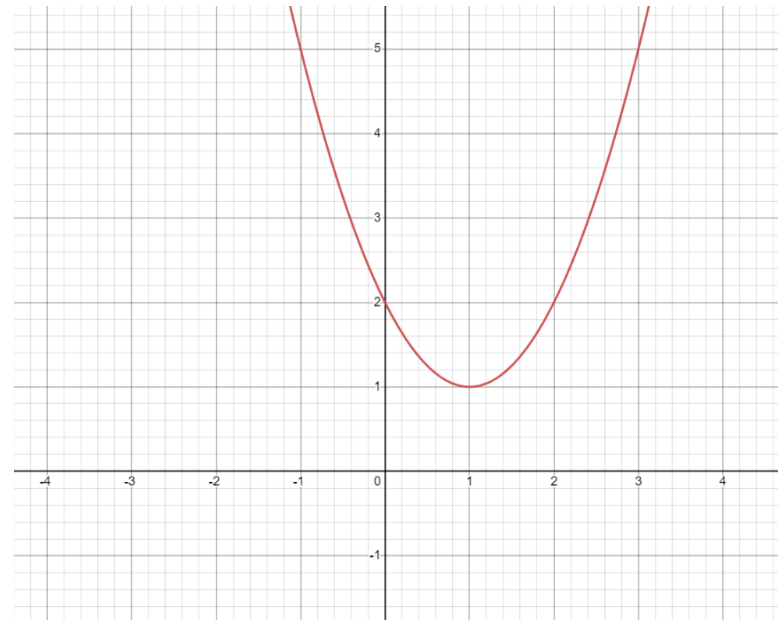


# 머신러닝 기초수학

## ■ 기초수학 (3)

### ■ 이차 함수

- $y$ 가  $x$ 에 대한 **이차식**으로 표현된 경우
- $y = a(x - p)^2 + q$  ( $a \neq 0$ )



# 머신러닝 기초수학

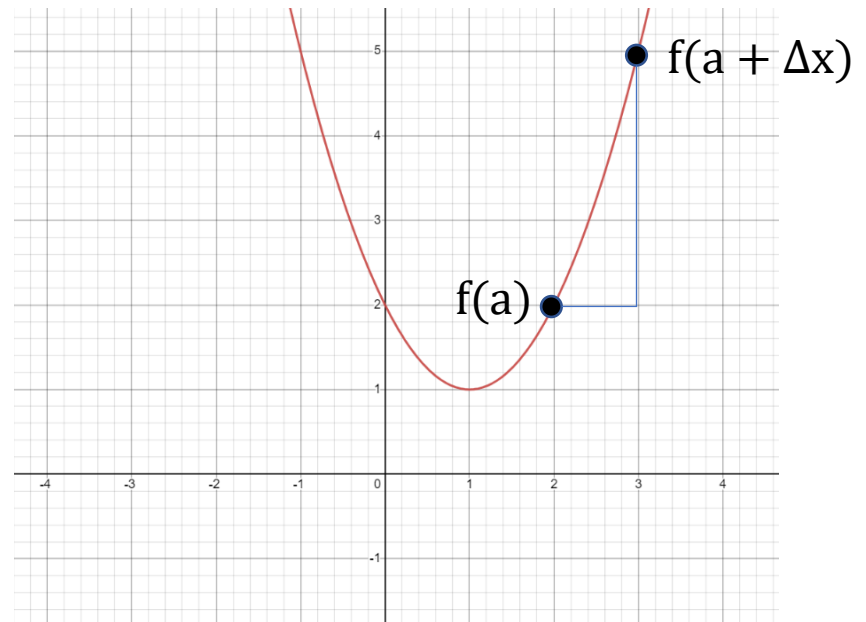
## ■ 기초수학 (4)

### ■ 순간 변화율

- $x$ 의 값이 미세하게 변화했을 때,  $y$ 의 변화율

- $\lim_{\Delta x \rightarrow 0} \frac{f(a+\Delta x) - f(a)}{\Delta x}$

- 어떤  $x$ 값( $=a$ )에서의 그래프와 맞는 접선의 기울기



# 머신러닝 기초수학

## ■ 기초수학 (5)

### ■ 미분

- 함수  $f(x)$ 를 미분한다는 것은 함수의 순간 변화율을 구한다는 뜻
- $f'(x)$  또는  $\frac{d}{dx}f(x)$ 로 표기
- ex)  $f(x) = ax$ ,  $f(x) = x^a$

### ■ 함수의 최솟값

- 함수의 최솟값에서의 미분값(순간 변화율)은 항상 0임
- 이를 바탕으로 파라미터의 최적값을 구할 수 있음

# 데이터 전처리

## ■ 변수 종류

### ■ 주요 분류 : 범주형 vs 수치형

#### • 수치형

- 연속형 : 절대적인 기준값이 존재하고, 사칙연산(+, -,  $\times$ ,  $\div$ )이 가능 (산의 높이, 몸무게, 키 등)
- 정수형 : 관찰대상의 상대적인 차이만 나타내며, 가감(+, -) 연산만 가능 (리커트 5점, 7점 척도)

#### • 범주형

- 명목변수(nominal variable) : 순위가 없는 경우 (남성, 여성)
- 서수변수(ordinal variable) : 순위가 있는 경우 (낮음, 중간, 높음)

# 데이터 전처리

## ■ 수치형

- 사례 기반 학습

## ■ 범주형

- 대부분의 다른 알고리즘에서는 이진더미를 만들어야 함
  - (더미의 수 = 범주의 수 -1)
- '학생', '무직', '직장인', '은퇴' 값을 갖는 변수의 경우,  
3개의 더미변수 사용 (학생, 직장인, 은퇴)

관측치	변수 (직업)	서수변수	더미변수 (학생여부)	더미변수 (직장인여부)	더미변수 (은퇴여부)
A	학생	1	1	0	0
B	무직	2	0	0	0
C	직장인	3	0	1	0
D	은퇴	4	0	0	1

# 데이터 전처리

## ■ 결측 데이터 처리

### ■ 해법 1 : 삭제

- 만약 적은 수의 레코드가 결측치를 갖는다면, 삭제
- 만약 많은 레코드가 작은 변수 집합에서 결측치라면, 해당 변수들을 제거 (또는 대체값 사용)
- 만약 많은 레코드가 결측치를 갖는다면, 삭제는 부적절

### ■ 해법 2 : 대체

- 결측치를 타당한 대체값으로 대체
- 레코드를 유지하고 (결측이 아닌) 정보의 나머지를 사용

# 데이터 전처리

## ■ 데이터 표준화

- 유클리드 거리를 이용하여 B와 가장 유사한 고객을 결정하시오.

고객	나이	수입(\$)
A	25	49,000
B	56	156,000
C	65	99,000
D	32	192,000
E	41	39,000
F	49	57,000
평균	44.67	98,667
표준 편차	14.98	62,867



# 데이터 전처리

## ■ 데이터 표준화

- 스케일이 큰 변수가 결과에 주된 영향을 미치고 결과가 왜곡될 경우에 사용
- 모든 변수를 같은 척도 안에 넣음
- 표준화 방법

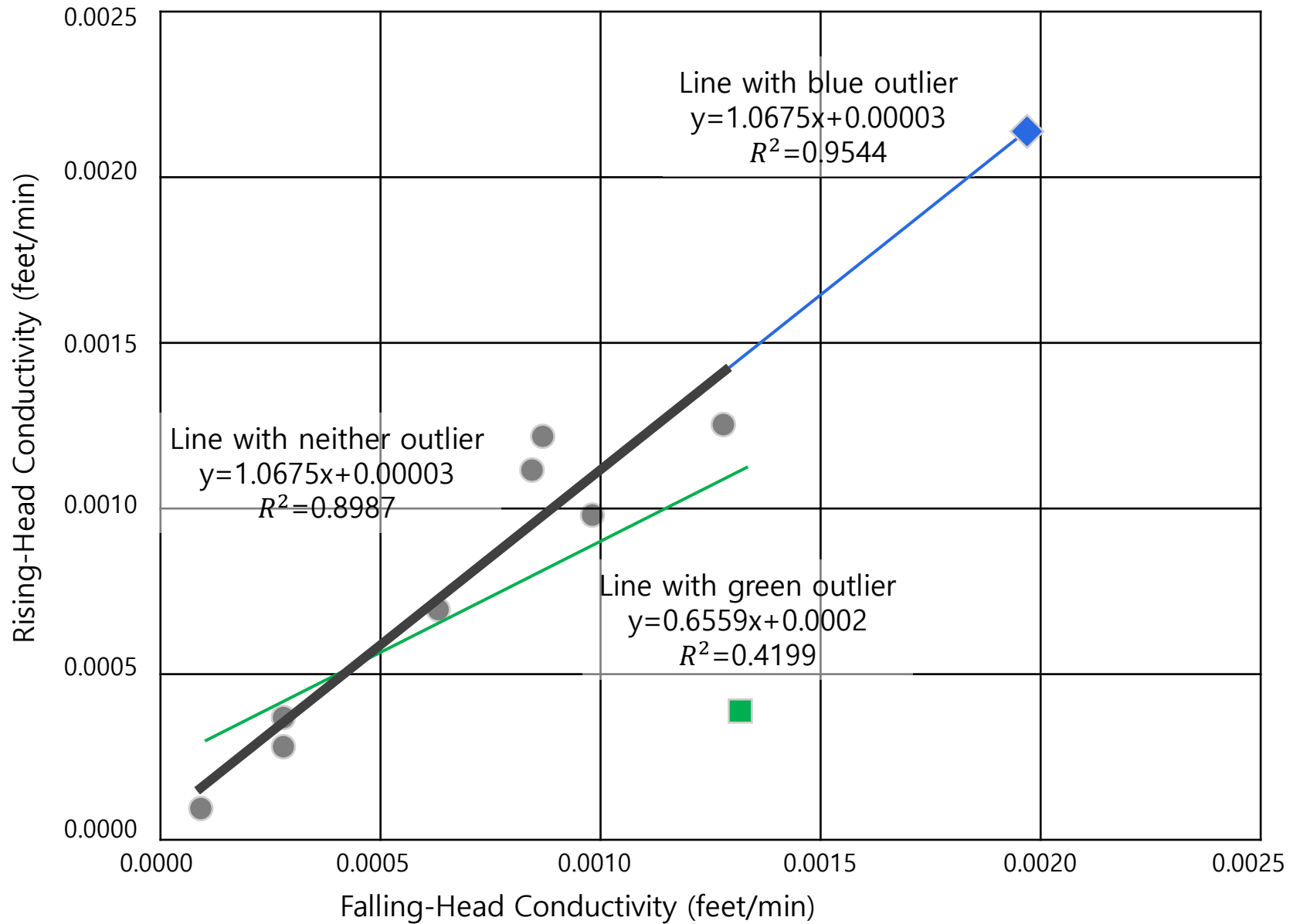
- 표준화 절차 1(Standardization) : 평균을 빼고 표준 편차로 나눔

$$x' = \frac{x - \bar{x}}{\sigma}$$

- 표준화 절차 2(Min-Max Normalization) : 최소값을 빼고 범위로 나눔으로써 0-1 척도로 변환

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

# 데이터 전처리



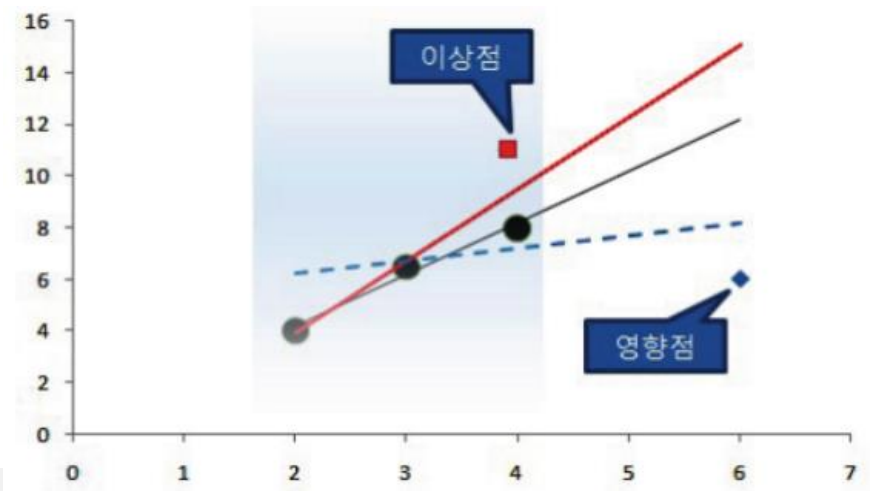
# 데이터 전처리

## ■ 이상치(Outlier) 처리

- 이상치는 “극단적인”, 나머지 데이터에서 멀리 떨어진 관측치 (“멀리 떨어진”이라는 용어는 분명히 모호한 표현)
- 이상치는 모델에 불균형한 영향을 줄 수 있음
- 데이터 전처리에서 중요한 단계가 이상치를 발견하는 것
- 어떤 경우에 이상치를 찾는 것이 데이터마이닝 실행의 목적 (공항 검색절차) 이를 “이상 탐지”(outlier detection)

## ■ 영향점(influential point)과의 차이점은?

- 회귀직선의 기울기 변화에 큰 효과를 미치는 점



# 데이터 전처리

## ■ 과적합(Overfitting)

- 개발된 통계적 모델이 기존 데이터 변수들 사이의 복잡한 관계에 대해서는 매우 잘 설명하지만, 새로운 데이터에는 잘 맞지 않는 현상

## ■ 과적합의 원인

- 너무 많은 예측변수들
- 너무 많은 파라미터들을 가진 (복잡한) 모델

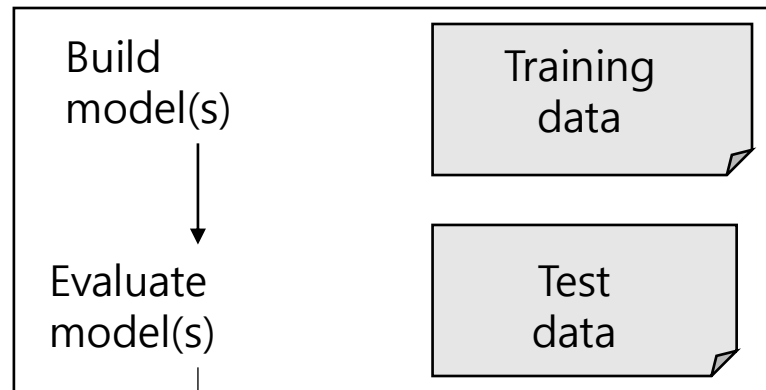
## ■ 과적합의 결과

- 생성된 모델이 신규 데이터에 대해서 제대로 적용되지 않음

# 데이터 전처리

## ■ 과적합 문제 해결방법

- 문제 : 모델이 새로운 데이터에 얼마나 잘 돌아갈 것인가?
- 해법 : 데이터를 두 부분으로 분할
  - 학습 데이터(train data)라는 모델을 개발
  - 평가 데이터(test data)에는 모델을 적용하고, “새로운” 데이터에 대해 일반화 성능 평가

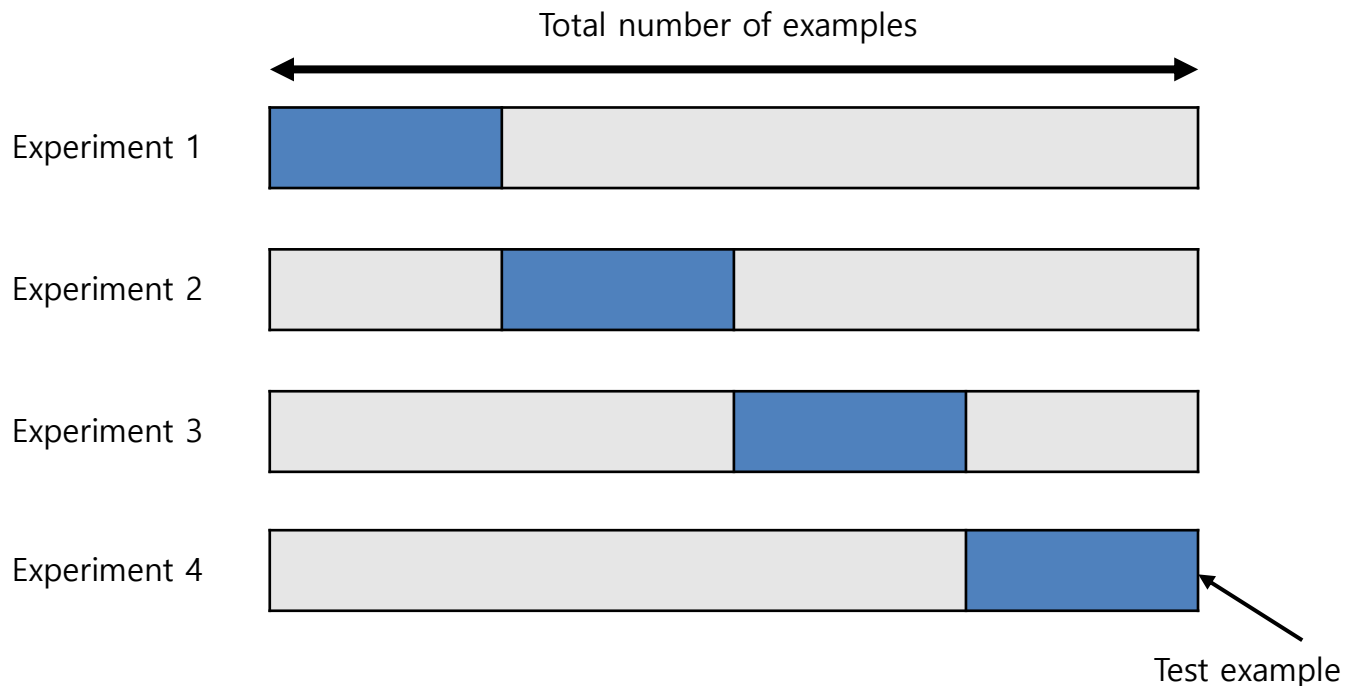


# 데이터 전처리

## ■ 과적합 문제 해결방법

### ■ K-fold cross validation

- 전체 데이터를 k개로 나눈 후, 이중 k-1개를 train data로, 나머지 1개를 test data로 활용하여 총 k번의 실험을 수행한 후, 이들 결과의 평균을 성능평가에 활용
- 모든 데이터가 train과 test에 활용



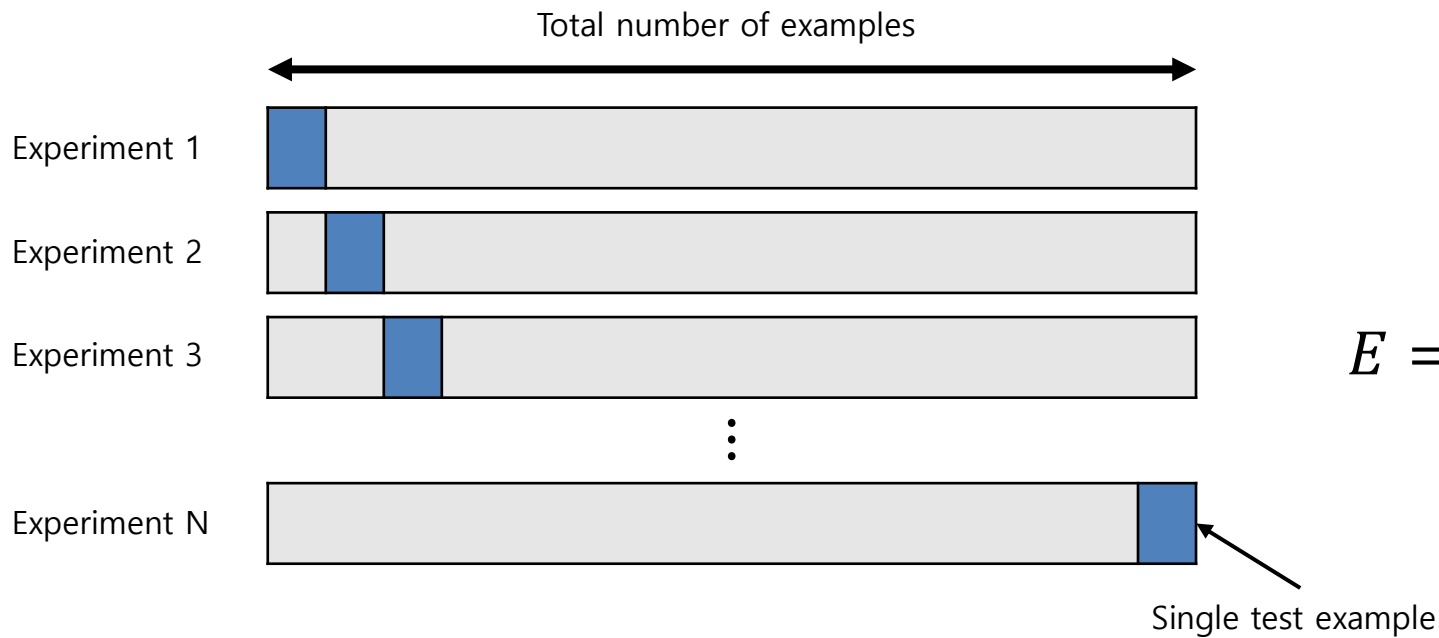
$$E = \frac{1}{K} \sum_{i=1}^K E_i$$

# 데이터 전처리

## ■ 과적합 문제 해결방법

### ▪ Leave-one-out CV

- 전체 데이터를 N개에 대해서, N-1개를 train data로, 나머지 1개를 test data로 활용하여 총 N번의 실험을 수행한 후, 이들 결과의 평균을 성능평가에 활용 (i.e., N-fold CV)



$$E = \frac{1}{N} \sum_{i=1}^N E_i$$

## ■ Training data vs. Test data

### ■ 데이터의 분할

- 입력된 데이터는 학습 데이터와 평가 데이터로 나눌 수 있음
- 학습 데이터는 모델 학습에 사용되는 모든 데이터셋
- 평가 데이터는 오직 모델의 평가만을 위해 사용되는 데이터셋
- 평가 데이터는 절대로 모델 학습에 사용되면 안됨

### ■ 평가 데이터

- 학습 데이터와 평가 데이터는 같은 분포를 가지는가?
- 평가 데이터는 어느 정도 크기를 가져야 하는가?

Train data	Test data
------------	-----------



## ■ Validation data

### ■ 검증 데이터셋

- 모델 학습의 정도를 검증하기 위한 데이터셋
- 모델 학습에 직접적으로 참여하지 못함
- 학습 중간에 계속해서 평가를 하고,  
가장 좋은 성능의 파라미터를 저장해 둬



## ■ Bias and Variance Trade-off (1)

### ■ 모델의 복잡도

- 선형에서 비선형 모델로 갈수록 복잡도가 증가함
- 모델이 복잡할수록 학습 데이터를 다 완벽하게 학습함
- 그러면 좋은가?
  1. 데이터가 너무 적은 상황 (Under-fitting, 과소적합)  
: 데이터의 특성에 비해 모델이 너무 간단함
  2. 학습 데이터에 대해 과하게 학습된 상황 (Over-fitting, 과적합)  
: 학습 데이터 이외의 데이터에 대해선 모델이 잘 작동하지 못함

## ■ Bias and Variance Trade-off (2)

### ■ 편향(bias)과 분산(variance)

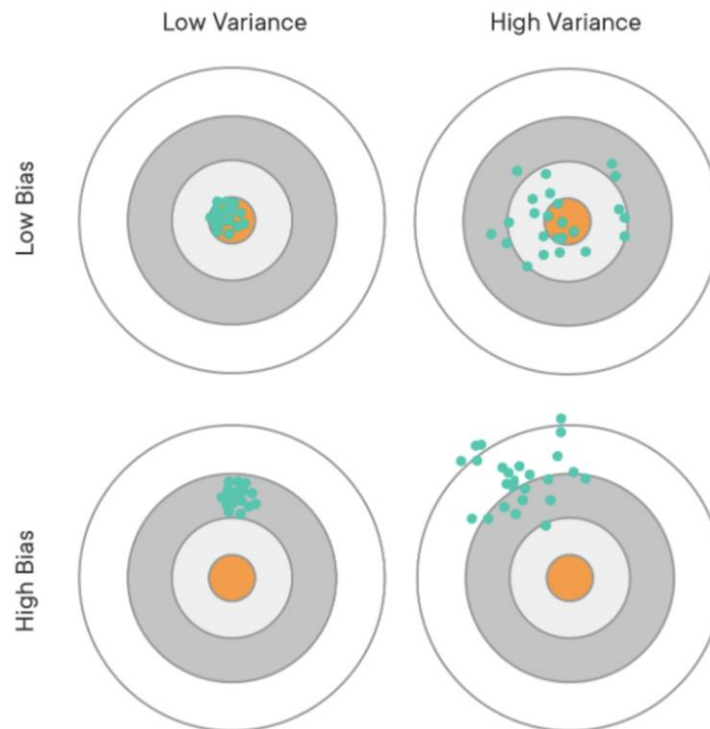
- 편향과 분산은 모두 알고리즘이 가지고 있는 에러의 종류

- $$\begin{aligned} MSE(\hat{\theta}) &= E_{\theta} \left( (\hat{\theta} - \theta)^2 \right) = E \left( (\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta)^2 \right) \\ &= E \left( (\hat{\theta} - E(\hat{\theta}))^2 + 2(\hat{\theta} - E(\hat{\theta}))(E(\hat{\theta}) - \theta) + (E(\hat{\theta}) - \theta)^2 \right) \\ &= E \left( (\hat{\theta} - E(\hat{\theta}))^2 \right) + 2 \left( (E(\hat{\theta}) - \theta)(\hat{\theta} - E(\hat{\theta})) \right) + (E(\hat{\theta}) - \theta)^2 \\ &= E \left( (\hat{\theta} - E(\hat{\theta}))^2 \right) + (E(\hat{\theta}) - \theta)^2 \\ &= Var_{\theta}(\hat{\theta}) + Bias_{\theta}(\hat{\theta}, \theta)^2 \end{aligned}$$

## ■ Bias and Variance Trade-off (3)

### ■ 편향(bias)과 분산(variance)

- 편향과 분산은 모두 알고리즘이 가지고 있는 에러의 종류
- $MSE(\hat{\theta}) = E \left( (\hat{\theta} - E(\hat{\theta}))^2 \right) + (E(\hat{\theta}) - \theta)^2 = Var_{\theta}(\hat{\theta}) + Bias_{\theta}(\hat{\theta}, \theta)^2$

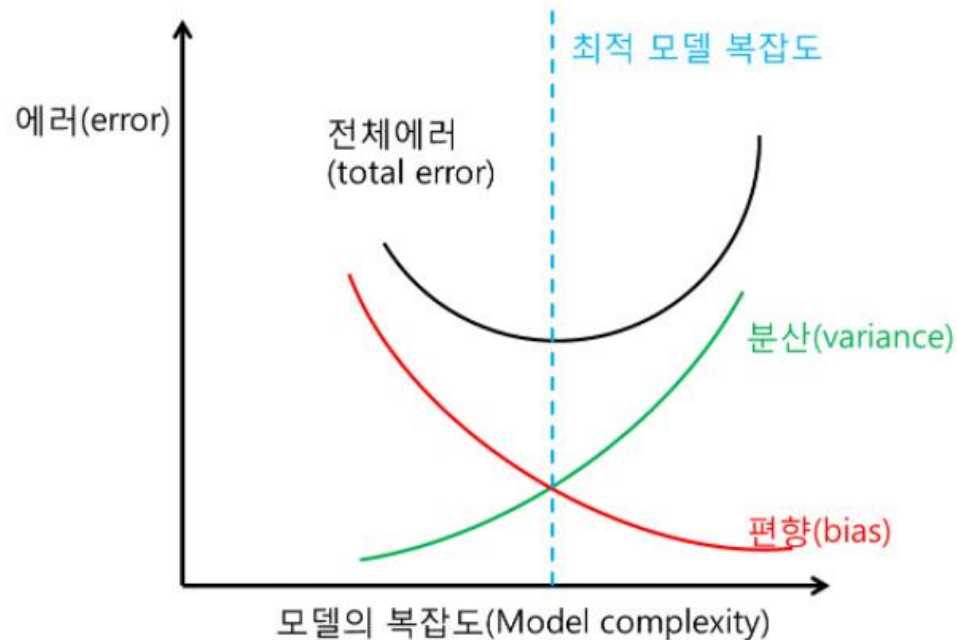


# 편향과 분산

## ■ Bias and Variance Trade-off (4)

### ■ 편향(bias)과 분산(variance)

- 편향은 under-fitting 과 관련 있는 개념
- 분산은 over-fitting과 관련 있는 개념



# 지도 학습 - 회귀(예측) (선형회귀분석)

# 회귀(예측) - 선형회귀분석

- 단순 선형 회귀 (simple linear regression)
  - 피처의 종류가 한 개인 데이터에 대한 회귀 모델
  - $y = w_0 + w_1x$

	공부시간	성적
길동	2시간	40
철수	4시간	58
상민	6시간	64
영희	8시간	75

- 다중 선형 회귀 (multiple linear regression)
  - 피처의 종류가 여러 개인 데이터에 대한 회귀 모델
  - $y = w_0 + w_1x_1 + w_2x_2 + \dots + w_Dx_D$
- 다항 회귀 (polynomial regression)
  - 독립 변수(피처)의 차수를 높인 회귀 모델
  - $y = w_0 + w_1x + w_2x^2 + w_mx^m$

# 회귀(예측) - 선형회귀분석

## ■ 회귀분석 - 예제

$y$	$x$
5	1
20	4
105	21
22	4.4
10	2

$$y = 5 \times x$$

$y$	$x_1$	$x_2$
9	1	2
24	4	2
111	21	3
30	4.4	4
20	2	5

$$y = 5 \times x_1 + 5 \times x_2$$

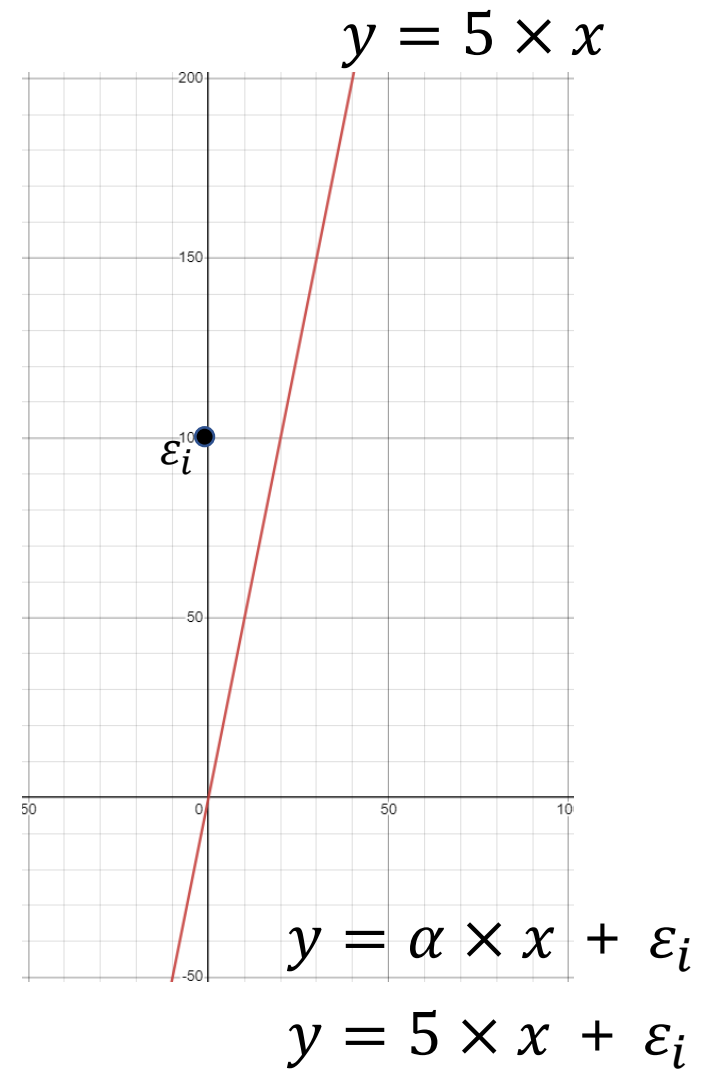
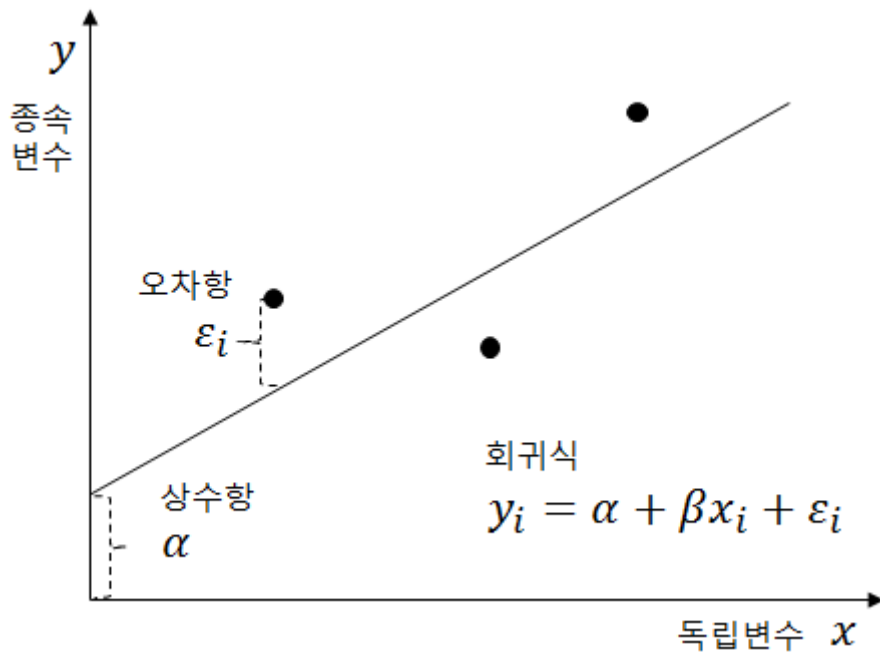
$y$	$x$	$\varepsilon$
6	1	-1
22	4	+2
104	21	-1
20	4.4	-2
13	2	+3

$$y = 5 \times x + \varepsilon_i$$



# 회귀(예측) - 선형회귀분석

## ■ 회귀분석 - 예제



# 회귀(예측) - 선형회귀분석

## ■ 회귀분석이란?

- 연구대상이 되는 시스템에 존재하는 변수들 사이의 함수적인 관계를 규명하기 위해 수학적인 모형을 상정하고, 이 모형을 수집된 자료로부터 추정하는 통계적인 기법

## ■ ‘회귀’ 라는 말의 유래

- 영국의 유전학자 갈튼 (Galton)
- 아버지와 아들의 키의 관계에 대한 연구에서 유래됨  
(아들의 키는 또래의 평균 키로 회귀하려는 경향이 있음)

## ■ 회귀분석의 목적

- 설명모델 : 변수간의 관계를 기술하고 설명
  - 예) 아파트 평수와 전기소모량의 관계
- 예측모델 : 목표변수의 값을 예측
  - 예) 아파트 평수에 따른 전기소모량 예측

# 회귀(예측) - 선형회귀분석

## ■ 설명모델

### ■ 모델 목표

- 설명변수(독립변수, 회귀변수)와 목표변수(종속변수, 반응변수) 사이의 관계 설명
- 데이터를 잘 적합하고 모델에 대한 설명 변수들의 기여 정도를 이해하는 것이 모델의 목표

### ■ 데이터 분석에서 회귀분석을 사용하는데 많이 쓰임

### ■ “적합도 검증” : $R^2$ , 잔차 분석

## ■ 예측모델

### ■ 모델 목표

- 예측변수 값은 있지만, 목표변수의 값이 없는 경우, 다른 데이터로부터 목표변수의 값을 예측
- 예측 정확성을 최적화하는 것이 모델의 목표

### ■ 학습 데이터에서 학습 모델 생성후, 테스트 데이터에서 성능평가

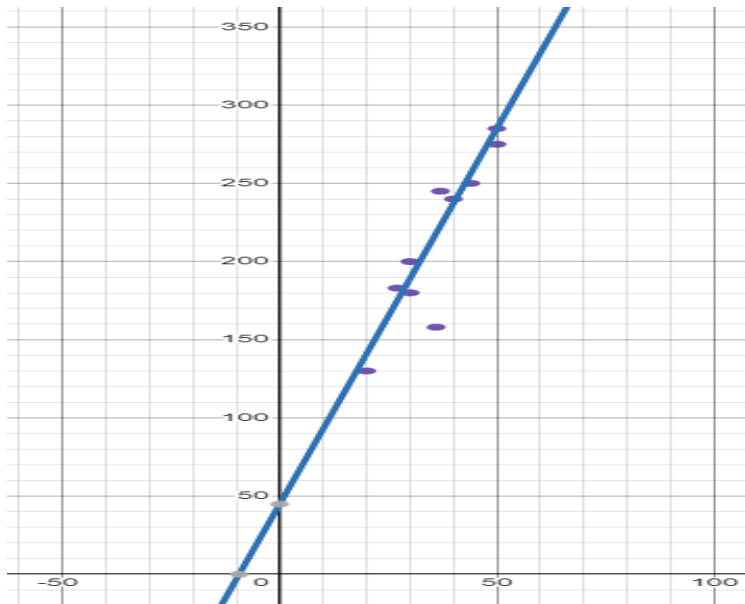
### ■ 예측변수는 설명변수로서의 역할이 주요한 목적이 아님

# 회귀(예측) - 선형회귀분석

## ■ 예제 - 단순회귀분석

- 아파트의 평수와 전기소모량의 관계를 알아보기 위해서 아파트 단지내 여러 가구 중에서 10가구를 임의로 선택하여 다음 자료를 수집하였다.

X(평수)	20	25	27	30	30	37	40	44	50	50
Y(전기소모량:kw)	130	158	183	180	200	245	240	250	285	275



모형가정

$$y = \beta_0 + \beta_1 x + \varepsilon$$

모형추정

$$\hat{y} = 44.8 + 4.81x$$

모형검토

모형의 유의성, 적합성  
검토

모형이용

$$\begin{aligned}\hat{y}_{x=33} &= 44.8 + 4.81 \times 33 \\ &= 203.53(\text{kw})\end{aligned}$$

# 회귀(예측) - 선형회귀분석

## ■ 모형 추정

- 최소제곱법이라는 추정방법을 이용하여  $\beta$ 를 추정한다.
- 최소제곱법(Least Squares Method)
  - 미지의 모수  $\beta_0$ 와  $\beta_1$ 은 오차의 제곱합이 최소가 되도록 하는 값으로 추정하는데, 이러한 방법을 최소제곱법이라고 함

$$L = \sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2$$

- 최소제곱법에 의한  $\beta_0$ 와  $\beta_1$ 의 추정량

- $\hat{\beta}_1 = b_1 = \frac{\sum_{j=1}^n (x_j - \bar{x})(y_j - \bar{y})}{\sum_{j=1}^n (x_j - \bar{x})^2}$

- $\hat{\beta}_0 = b_0 = \bar{y} - b_1 \bar{x}$

- 단,  $\bar{x}$ ,  $\bar{y}$ 는  $x$ 와  $y$ 의 평균치

$$\begin{aligned} L &= \sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j)^2 \\ \frac{\partial L}{\partial \beta_1} &= -2 \sum_{j=1}^n x_j (y_j - \beta_0 - \beta_1 x_j) = 0 \\ \frac{\partial L}{\partial \beta_0} &= -2 \sum_{j=1}^n (y_j - \beta_0 - \beta_1 x_j) = 0 \end{aligned}$$

# 회귀(예측) - 선형회귀분석

## ■ 모형 추정

### ■ 최소제곱 회귀식

- 추정된  $b_0$ 와  $b_1$ 을 이용하여 반응변수  $y$ 에 대한 추정식(회귀식)을 다음과 같이 나타낼 수 있음

$$\hat{y} = b_0 + b_1x$$

- $x = x_j$ 일 때의 반응변수  $y$ 의 추정값은 다음과 같음

$$\hat{y}_j = b_0 + b_1x_j$$

# 회귀(예측) - 선형회귀분석

## ■ 모형 검토

### ■ 분산분석에 의한 모형의 유의성 검토

- 총제곱합(Total Sum of Squares, SSTO)의 분해
- 총제곱합은 회귀식에 의해 설명되는 변동(Regression Sum of Squares, SSR)과 회귀식에 의해 설명되지 않는 잔차변동(Residual Sum of Squares, SSE)으로 분해

SSTO		SSR		SSE
$\sum_{j=1}^n (y_j - \bar{y})^2$	=	$\sum_{j=1}^n (\hat{y}_j - \bar{y})^2$	+	$\sum_{j=1}^n (y_j - \hat{y}_j)^2$

### ■ 결정계수(coefficient of determination)

- 전체 변동 중 회귀식에 의해 설명되는 변동의 비율

$$R^2 = \frac{SSR}{SSTO} = 1 - \frac{SSE}{SSTO}$$

# 회귀(예측) - 선형회귀분석

$y$	$x$
7	1
21	4
105	21
21	4.4
8	2

$$y = 5 \times x + \varepsilon_i$$

$y$	$x_1$	$x_2$
11	1	2
25	4	2
111	21	3
29	4.4	4
19	2	5

$$y = 5 \times x_1 + 2 \times x_2 + \varepsilon_i$$



# 회귀(예측) - 선형회귀분석

## ■ 다중회귀모형

$$\begin{array}{c} \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \\ n \times 1 \end{array} = \begin{array}{c} \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & & & & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \\ n \times (p+1) \end{array} \begin{array}{c} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} \\ (p+1) \times 1 \end{array} + \begin{array}{c} \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix} \\ n \times 1 \end{array}$$

## ■ 예제

$$\begin{pmatrix} Price_1 \\ Price_2 \\ \vdots \\ Price_n \end{pmatrix} = \begin{bmatrix} 1 & \text{평형}_1 & \text{위치}_1 & \cdots & \text{범죄율}_1 \\ 1 & \text{평형}_2 & \text{위치}_2 & \cdots & \text{범죄율}_2 \\ \vdots & & & & \\ 1 & \text{평형}_n & \text{위치}_n & \cdots & \text{범죄율}_n \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$$y_i(Price_i) = \beta_0 + \beta_1 \cdot \text{평형} + \beta_2 \cdot \text{위치} + \cdots + \beta_p \cdot \text{범죄율} + \varepsilon_i = X\beta + \varepsilon$$

# 회귀(예측) - 선형회귀분석

## ■ 다중회귀모형

- 일반적인 회귀모형은 소위 선형 회귀모형(linear regression model)이라고 불리며, 그 일반적인 형태는 다음과 같음

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj} + \varepsilon_j, \quad j = 1, 2, \dots, n \quad (1)$$
$$\varepsilon_j \sim \text{NID}(0, \sigma^2)$$

- 이를 vector/matrix 형태로 나타내면,

$$\begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} = \begin{bmatrix} 1 & x_{11} & x_{21} & \cdots & x_{p1} \\ 1 & x_{12} & x_{22} & \cdots & x_{p2} \\ \vdots & & & & \\ 1 & x_{1n} & x_{2n} & \cdots & x_{pn} \end{bmatrix} \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

$n \times 1 \qquad n \times (p+1) \qquad (p+1) \times 1 \qquad n \times 1$

$$y = X\beta + \varepsilon$$

- 식 (1)에서  $x_{pj}$  는 반드시 원래의 변수만을 의미하는 것은 아님

# 회귀(예측) - 선형회귀분석

## ■ 다중회귀모형

- 예를 들어, 두 개의 변수에 대해 2차 모형을 상정했다면,

$$y_j = \beta_0 + \beta_1 u_{1j} + \beta_2 u_{2j} + \beta_3 u_{1j}^2 + \beta_4 u_{2j}^2 + \beta_5 u_{1j} u_{2j} + \varepsilon_j$$

로 나타낼 수 있는데, 모형(1)로는

$$p = 5$$

$$x_{1j} = u_{1j}, x_{2j} = u_{2j}, x_{3j} = u_{1j}^2, x_{4j} = u_{2j}^2, x_{5j} = u_{1j} u_{2j}$$

# 회귀(예측) - 선형회귀분석

## ■ 최소제곱법에 의한 추정

- 미지의 회귀계수 vector 는  $\beta$  다음 오차제곱합이 최소가 되도록 결정

$$L = \sum_{j=1}^n \varepsilon_j^2 = \sum_{j=1}^n \{y_j - (\beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \cdots + \beta_p x_{pj})\}^2$$

- 최소제곱법에 의한  $\beta$  추정량

$$b = (X'X)^{-1}X'y = \begin{pmatrix} b_0 \\ b_1 \\ \vdots \\ b_p \end{pmatrix}$$

# 회귀(예측) - 선형회귀분석

## ■ 예제 - 다중회귀분석

### ■ 회귀모형

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \varepsilon_j, \quad j = 1, 2, \dots, 25$$

$$y = \begin{pmatrix} 10.98 \\ 11.13 \\ \vdots \\ 11.08 \end{pmatrix} \quad X = \begin{bmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ \vdots & \vdots & \vdots \\ 1 & 28.6 & 22 \end{bmatrix} \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \end{pmatrix} \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_{25} \end{pmatrix}$$

### ■ 최소제곱법에 의한 $\beta$ 추정

$$b = (X'X)^{-1}X'y$$

$$\begin{aligned} &= \left\{ \begin{bmatrix} 1 & 1 & \dots & 1 \\ 35.3 & 29.7 & \dots & 28.6 \\ 20 & 20 & \dots & 22 \end{bmatrix} \begin{bmatrix} 1 & 35.3 & 20 \\ 1 & 29.7 & 20 \\ \vdots & \vdots & \vdots \\ 1 & 28.6 & 22 \end{bmatrix} \right\}^{-1} \times \begin{bmatrix} 1 & 1 & \dots & 1 \\ 35.3 & 29.7 & \dots & 28.6 \\ 20 & 20 & \dots & 22 \end{bmatrix} \begin{pmatrix} 10.98 \\ 11.13 \\ \vdots \\ 11.08 \end{pmatrix} \\ &= \begin{bmatrix} 2.778747 & -0.011242 & -0.106098 \\ & 0.146207 \times 10^{-3} & 0.175467 \times 10^{-3} \\ & & 0.478599 \times 10^{-2} \end{bmatrix} \begin{pmatrix} 235.6 \\ 11821432 \\ 483186 \end{pmatrix} = \begin{pmatrix} 9.1266 \\ -0.0724 \\ 0.2029 \end{pmatrix} = \begin{pmatrix} b_0 \\ b_1 \\ b_2 \end{pmatrix} \end{aligned}$$

(대칭)

# 회귀(예측) - 선형회귀분석

## ■ 예측모형의 성능평가

### ■ 절대 평균오차(MAE : mean absolute error)

- 절대평균오차의 크기를 의미

$$MAE = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| = 1.5$$

관측치	실제값 (y)	예측값 ( $\hat{y}$ )
1	5	4
2	3	6
3	4	5
4	3	2

### ■ 평균제곱오차의 제곱근(RMSE : root-mean-squared error)

- 예측된 변수와 동일한 측정단위 사용

$$MSE = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = 3$$

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} = \sqrt{3}$$

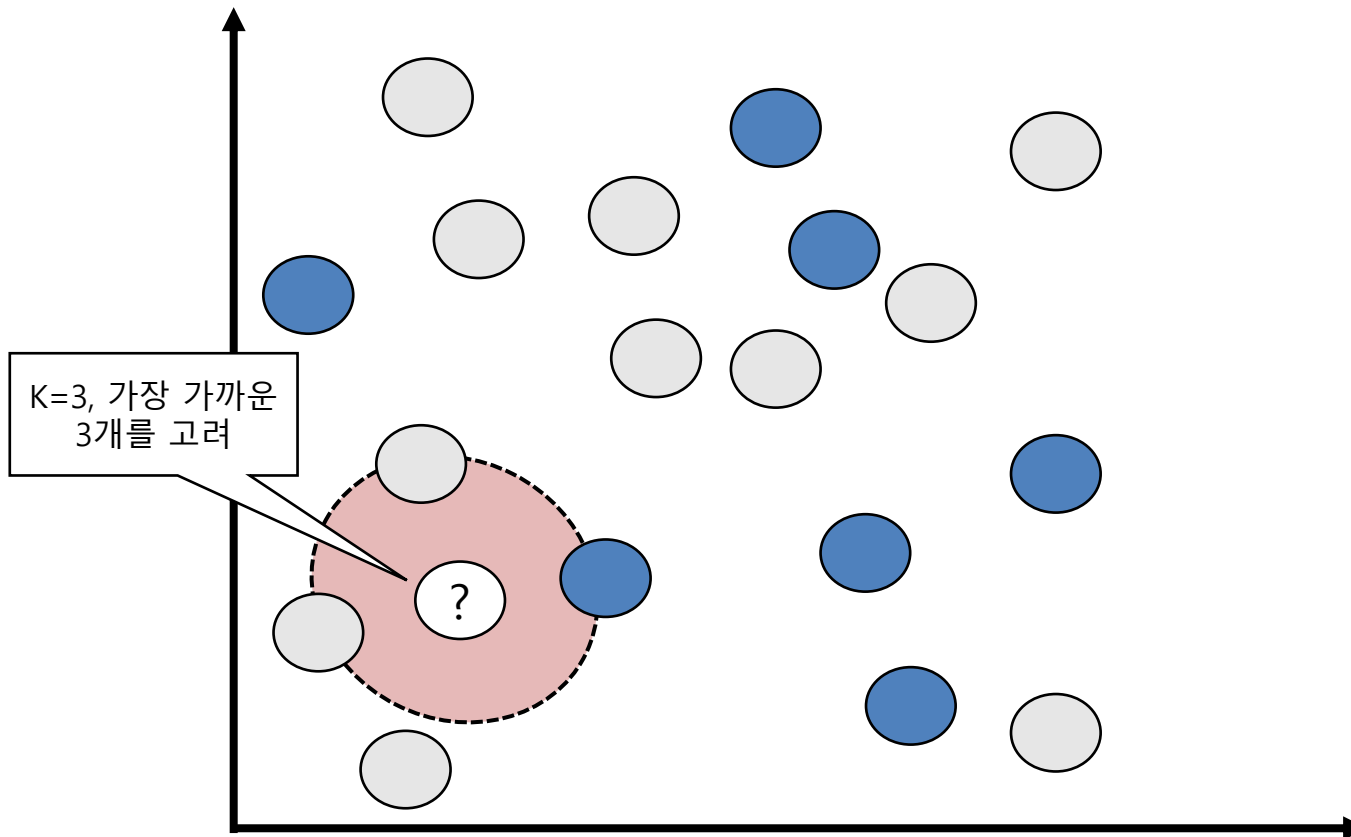
# 지도 학습 - 분류

( $\kappa$  -NN, Decision Tree)

# 분류 - $\kappa$ -NN

## ■ $\kappa$ -NN( $\kappa$ -Nearest Neighbors)

- A man is known by the company he keeps.
- Birds of a feather flock together.
- 모델 추론이 아닌 데이터 추론 데이터에 대한 어떠한 가정도 만들지 않음





# 분류 - $\kappa$ -NN

## ■ 기본 알고리즘

- 분류해야 할 레코드로부터 근접한  $\kappa$  개의 레코드를 찾는다.
- "근접"은 유사한 예측변수 값 을 가진 레코드를 뜻한다.
- 유사한 레코드들("이웃")이 주로 속한 클래스(우세한 클래스)로 해당 레코드 클래스를 결정 한다. - 다수결의 원칙

## ■ 유사성 측정 방법

- 가장 많이 쓰이는 거리 측정방법은 유클리드 거리

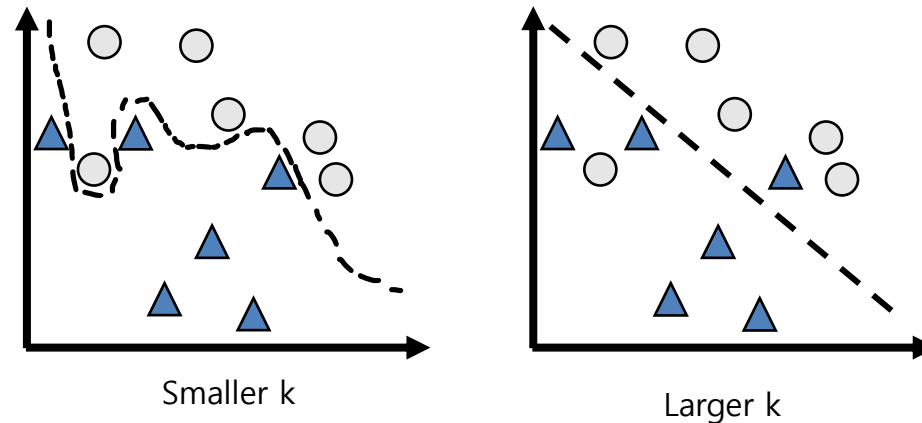
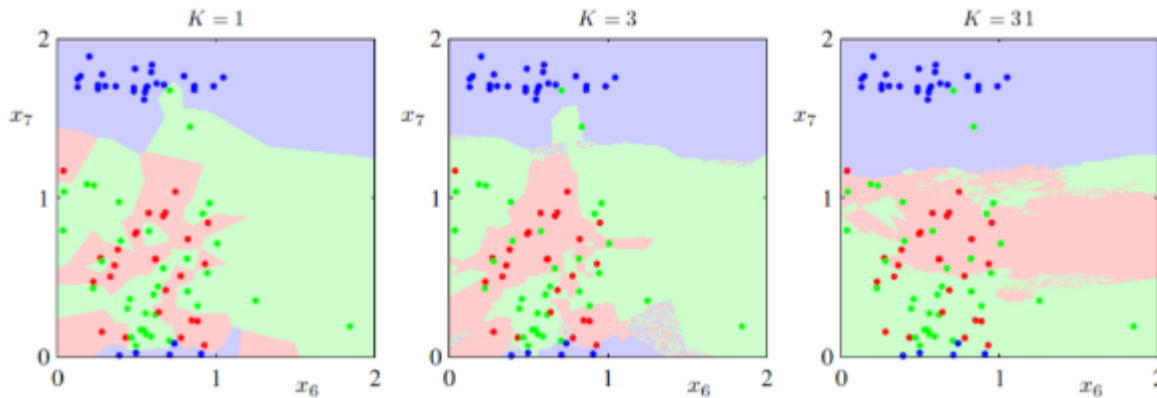
$$\sqrt{(x_1 - u_1)^2 + (x_2 - u_2)^2 + \dots + (x_p - u_p)^2}$$

- 예제 - 토마토(달콤함 = 6, 바삭함 = 4)와의 유사성 측정

음식	달콤함	바삭함	음식종류	유클리드거리(토마토)
포도	8	5	과일	$\sqrt{(6-8)^2 + (4-5)^2} = 2.2$
깍질콩	3	7	야채	$\sqrt{(6-3)^2 + (4-7)^2} = 4.2$
견과	3	6	단백질	$\sqrt{(6-3)^2 + (4-6)^2} = 3.6$
오렌지	7	3	과일	$\sqrt{(6-7)^2 + (4-3)^2} = 1.4$

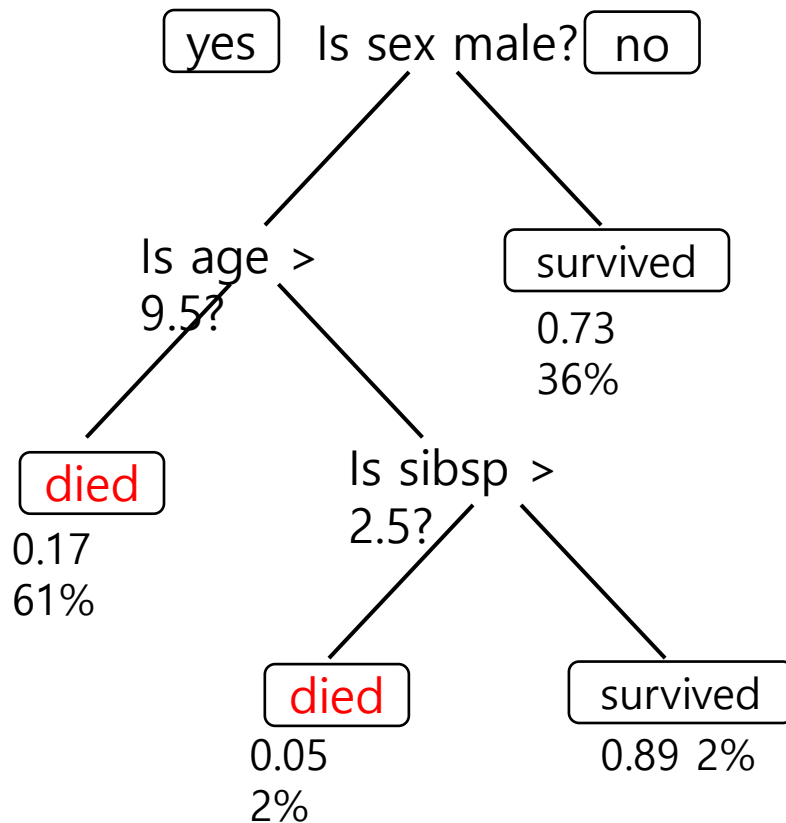
# 분류 - $\kappa$ -NN

- 작은  $\kappa$ 값(1,3,...)은 데이터의 지역적 구조(잡음을 포함하여)를 반영한다.
- 큰  $\kappa$ 값은 지역적 구조에 덜 민감하고 잡음의 영향을 덜 받지만 지역적 구조가 주는 정보를 놓칠 수 있다.



# 분류 - 의사결정나무

## ■ 의사결정나무 예제



변수	기준값
Sex	Male or Female
Age	$\leq 9.5$ or $> 9.5$
Siblings	$\leq 2.5$ or $> 2.5$

<타이타닉호 생존자 분류 의사결정  
나무>

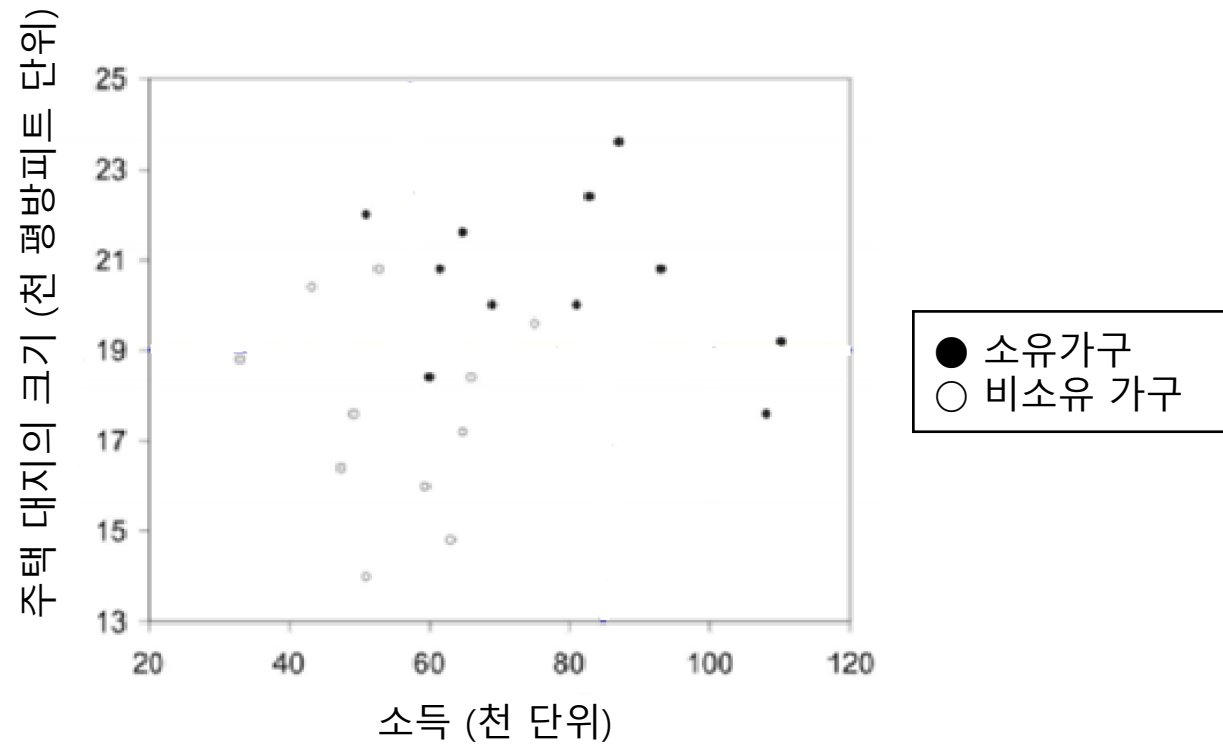
# 분류 - 의사결정나무

## ■ 의사결정나무 (Decision Tree)

- 정의 : 의사결정나무는 의사결정 규칙(Decision rule)을 나무구조로 도표화하여 관심대상이 되는 집단을 몇 개의 소집단으로 분류(Classification)하거나 예측(Prediction)을 수행하는 분석 방법
- 목표 : 일련의 예측변수를 기반으로 데이터를 분류 · 예측
  - 예: "신용카드 발급" 또는 "발급하지 않음" 으로 고객 분류
- 결과물 : 일련의 규칙
  - 예 : IF (Income > 92.5) AND (Education <= 1.5) AND (Family <= 2.5) THEN Class = 0 (non\_acceptor)
- 표현방법 : 규칙을 나무형태의 다이어그램으로 표현
- 대표 알고리즘 : Classification And Regression Tree (CART)

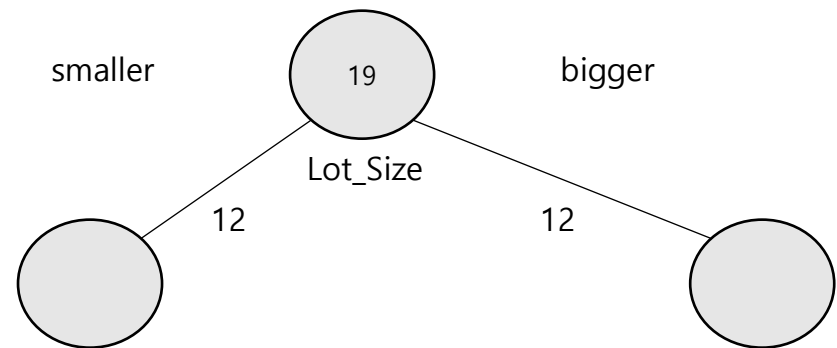
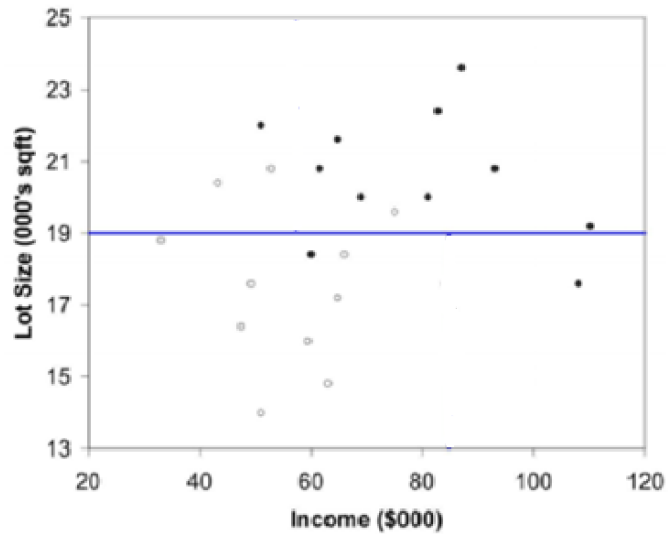
# 분류 - 의사결정나무

## ■ 분할방법



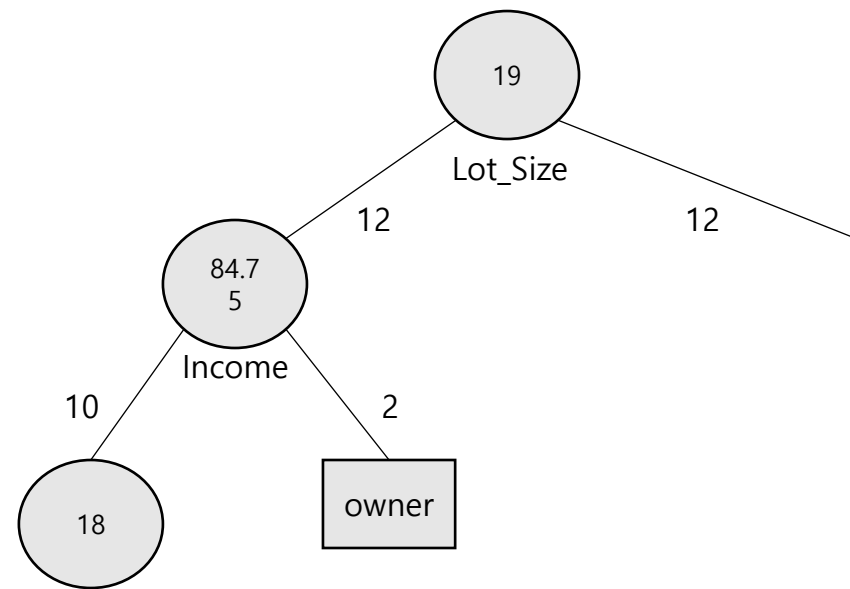
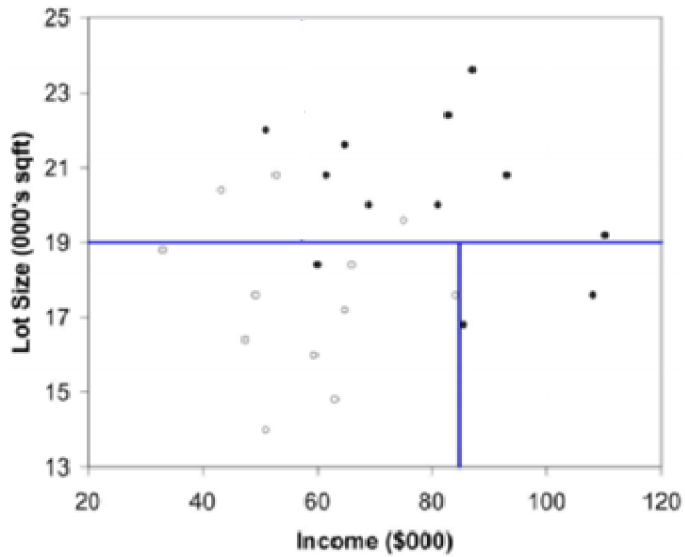
# 분류 - 의사결정나무

## ■ 첫 번째 분할: Lot Size = 19,000



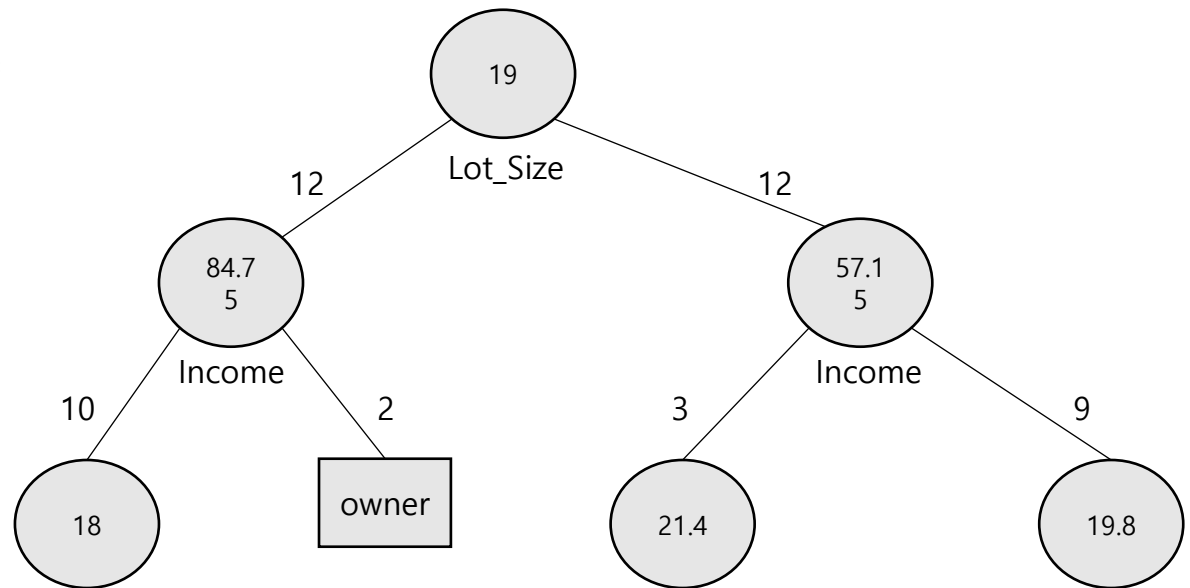
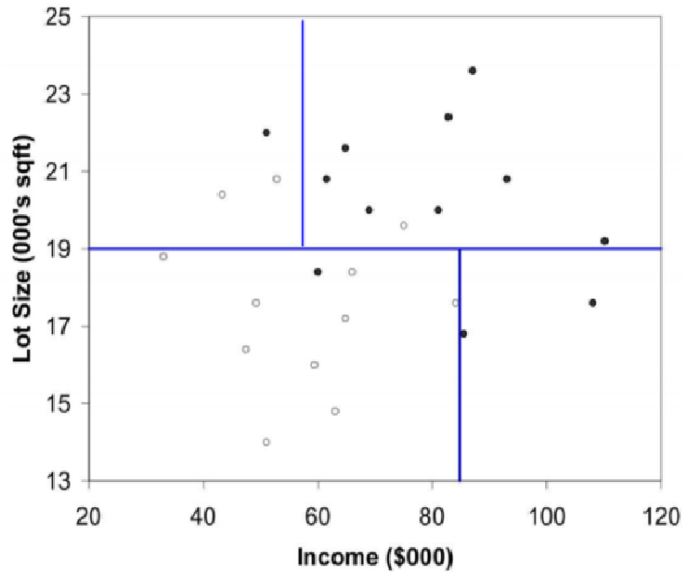
# 분류 - 의사결정나무

- 두 번째 분할: Income = \$84,750



# 분류 - 의사결정나무

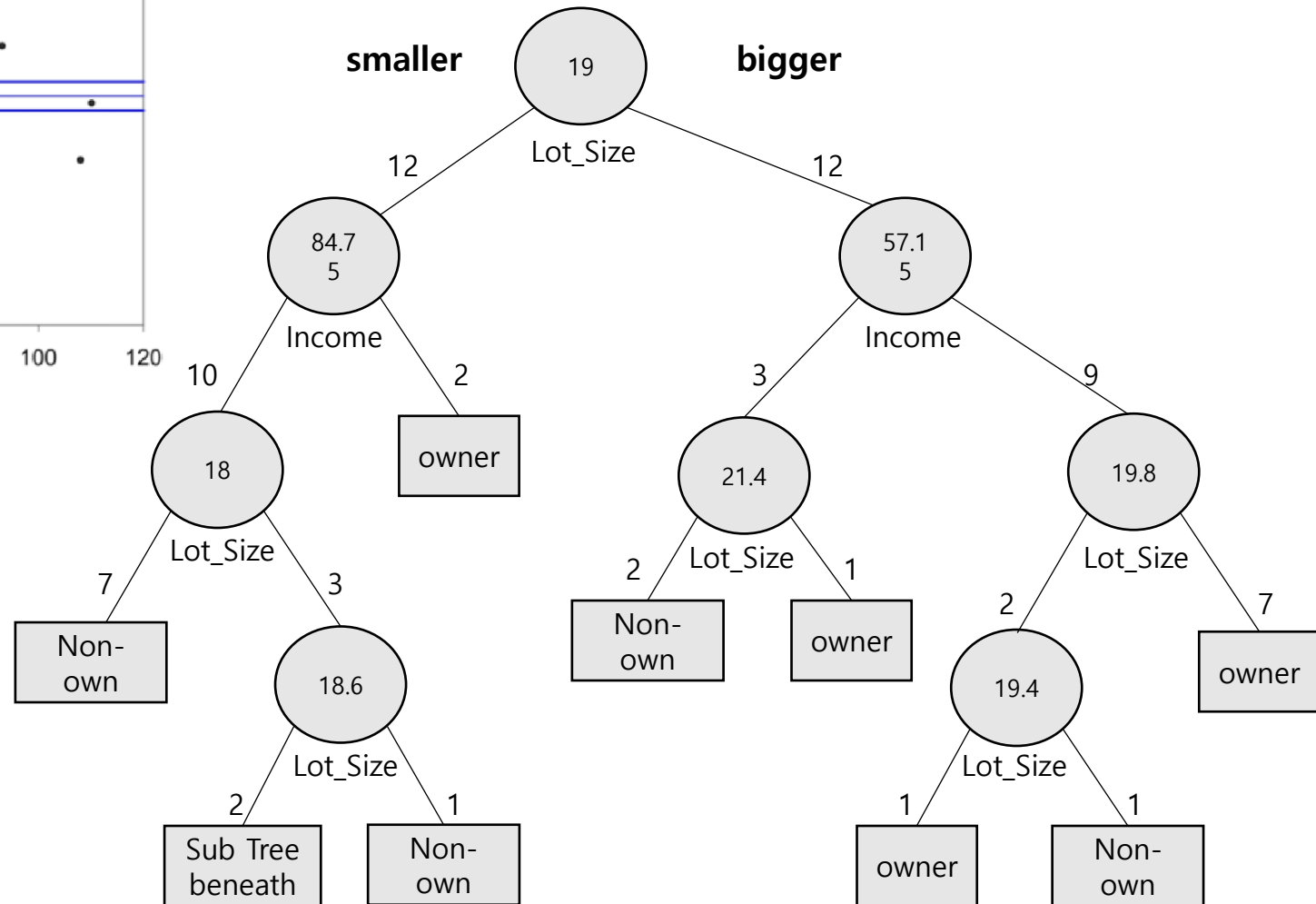
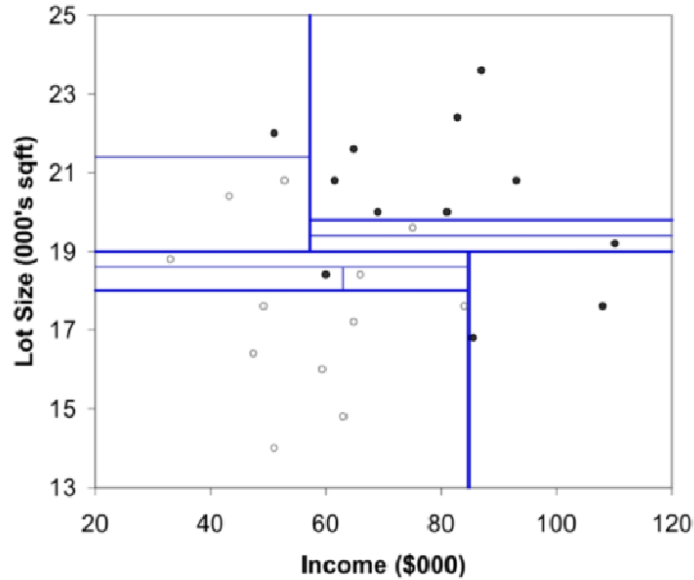
■ 세 번째 분할: Income = \$57,150





# 분류 - 의사결정나무

## ■ 모든 분할후



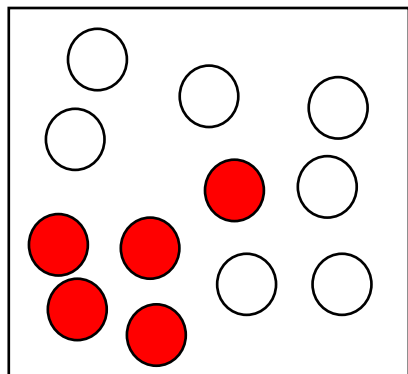
# 분류 - 의사결정나무

## ■ 불순도 측정 - 지니지수

- m개의 관측치를 포함하는 직사각형 A에 대한 지니지수

$$I(A) = 1 - \sum_{k=1}^m p_k^2$$

- $p_k$  : 직사각형 A 내에서 클래스 k에 속하는 관측치의 비율



$$\begin{aligned} I(A) &= 1 - \sum_{k=1}^m p_k^2 \\ &= 1 - \left(\frac{7}{12}\right)^2 - \left(\frac{5}{12}\right)^2 = 0.49 \end{aligned}$$

- 모든 관측치가 같은 클래스에 속할 때  $I(A) = 0$
- 각 클래스에 속한 관측치들의 비율이 같을 때  $I(A)$ 는 최대값 (=0.50) (이진분리 경우)

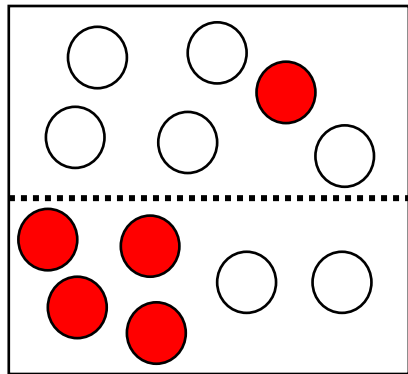
# 분류 - 의사결정나무

## ■ 불순도 측정 - 지니지수

- m개의 관측치를 포함하는 직사각형 d개의 직사각형이 존재할 때의 지니지수

$$I(A) = \sum_1^d \left( r_i \left( 1 - \sum_{k=1}^m p_k^2 \right) \right)$$

- $r_i$  : 전체 관측치중에서  $i$ 번째 직사각형내에 존재하는 관측치의 비율



$$\begin{aligned} I(A) &= \sum_1^d \left( r_i \left( 1 - \sum_{k=1}^m p_k^2 \right) \right) \\ &= \frac{6}{12} \left( 1 - \left( \frac{5}{6} \right)^2 - \left( \frac{1}{6} \right)^2 \right) + \frac{6}{12} \left( 1 - \left( \frac{2}{6} \right)^2 - \left( \frac{4}{6} \right)^2 \right) \\ &= \frac{1}{2} \left( 1 - \left( \frac{5}{6} \right)^2 - \left( \frac{1}{6} \right)^2 \right) + \frac{1}{2} \left( 1 - \left( \frac{2}{6} \right)^2 - \left( \frac{4}{6} \right)^2 \right) \\ &= 0.36 \end{aligned}$$

- Information Gain(IG) : 분할을 통해 얻어지는 지니지수의 감소량  
-  $IG = 0.49 - 0.36 = 0.13$

# 분류 - 의사결정나무

## ■ 예제 - 분류나무

- 목표변수는 '신용상태', 설명변수는 '급여형태', '나이', '직업' 등이라고 할 때 여러 설명 변수 중,  $\Delta G$ 가 최대화 되는 설명변수를 구하고자 합니다.  
먼저, '급여형태'에 따른 Gini index의 감소량을 구하시오.

	신용 상태(나쁨)	신용 상태(좋음)	합계
급여형태(주급)	143	22	165
급여형태(월급)	25	133	158
합계	168	155	323



$$G = 1 - (168/323)^2 - (155/323)^2 = 0.4992$$
$$G_L = 1 - (143/165)^2 - (22/165)^2 = 0.2311$$
$$G_R = 1 - (25/158)^2 - (133/158)^2 = 0.2664$$

Gini index의 감소량은

$$\Delta G = 0.4992 - ((165/323) \times 0.2311 + (158/323) \times 0.2664) = 0.2508$$

- 다른 설명변수에 대해서도  $\Delta G$ 를 구한 후, 가장 큰 값을 갖는 설명변수를 사용하여 분리

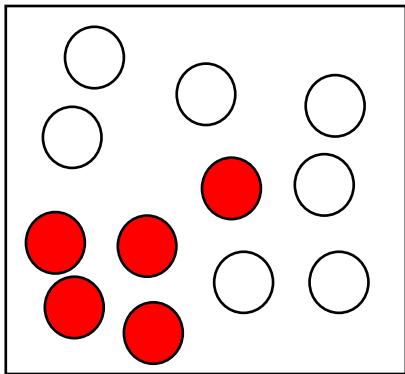
# 분류 - 의사결정나무

## ■ 불순도 측정 - 엔트로피

- m개의 관측치를 포함하는 직사각형 A에 대한 엔트로피

$$\text{Entropy}(A) = - \sum_{k=1}^m p_k \log_2(p_k)$$

- $p_k$  : 직사각형 A 내에서 클래스  $k$ 에 속하는 관측치의 비율



$$\begin{aligned} \text{Entropy}(A) &= - \sum_{k=1}^m p_k \log_2(p_k) \\ &= - \left( \frac{7}{12} \right) \log_2 \left( \frac{7}{12} \right) - \left( \frac{5}{12} \right) \log_2 \left( \frac{5}{12} \right) \\ &= 0.98 \end{aligned}$$

- 엔트로피값은 0(가장 순수)과  $\log_2(m)$  (클래스 비율이 같을 경우) 사이에 분포

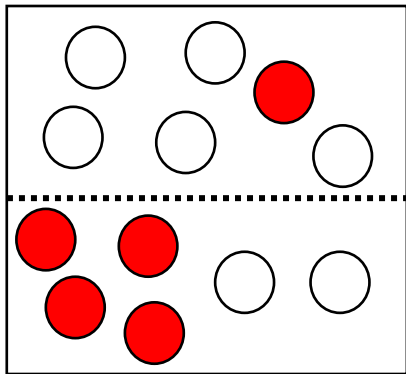
# 분류 - 의사결정나무

## ■ 불순도 측정 - 엔트로피

- m개의 관측치를 포함하는 직사각형 d개의 직사각형이 존재할 때의 엔트로피

$$\text{Entropy}(A) = \sum_1^d \left( r_i \left( - \sum_{k=1}^m p_k \log_2(p_k) \right) \right)$$

- $r_i$  : 전체 관측치중에서  $i$ 번째 직사각형내에 존재하는 관측치의 비율



$$\begin{aligned} \text{Entropy}(A) &= \sum_1^d \left( r_i \left( - \sum_{k=1}^m p_k \log_2(p_k) \right) \right) \\ &= \frac{6}{12} \left( - \left( \frac{5}{6} \right) \log_2 \left( \frac{5}{6} \right) - \left( \frac{1}{6} \right) \log_2 \left( \frac{1}{6} \right) \right) \\ &\quad + \frac{4}{12} \left( - \left( \frac{2}{6} \right) \log_2 \left( \frac{2}{6} \right) - \left( \frac{4}{6} \right) \log_2 \left( \frac{4}{6} \right) \right) \\ &= 0.78 \end{aligned}$$

- 분할을 통해 얻어지는 엔트로피의 감소량  $\text{Information Gain} = 0.98 - 0.78 = 0.20$

# 분류 - 의사결정나무

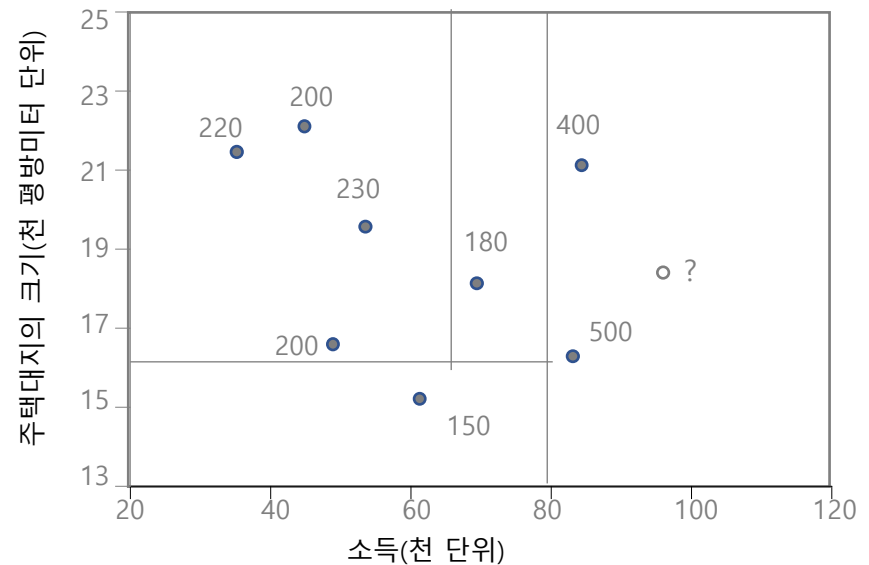
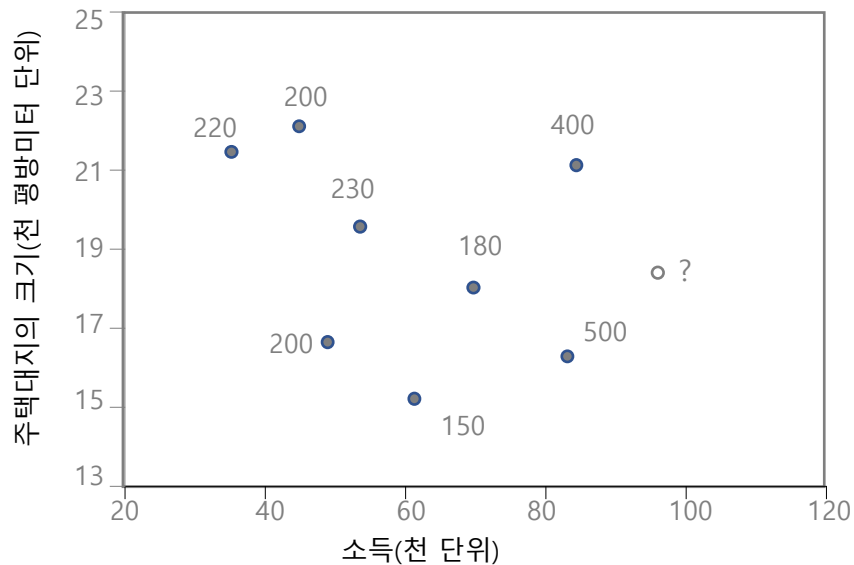
## ■ 분석절차

1. 목표변수를 잘 분리하는 설명변수 선택
2. 적절한 분리기준을 바탕으로 자식마디의 순도가 최대화되도록 분리지점 선택
3. 의사결정 나무 생성
4. 부적절한 나뭇가지는 제거 (가지치기)
5. 최종 의사결정나무 선정
6. 분류규칙(Rule) 도출
7. 분류(Classification) 및 예측(Prediction)

# 분류 - 의사결정나무

## ■ 회귀나무

- 수치형(연속형) 목표변수 사용
- 분리 절차는 분류나무와 유사
- 회귀나무의 불순도는 분산으로 측정
- 불순도를 최소화하도록 분리기준 선택
- 예측값은 직사각형에서 수치형 타깃 변수들의 평균으로 계산

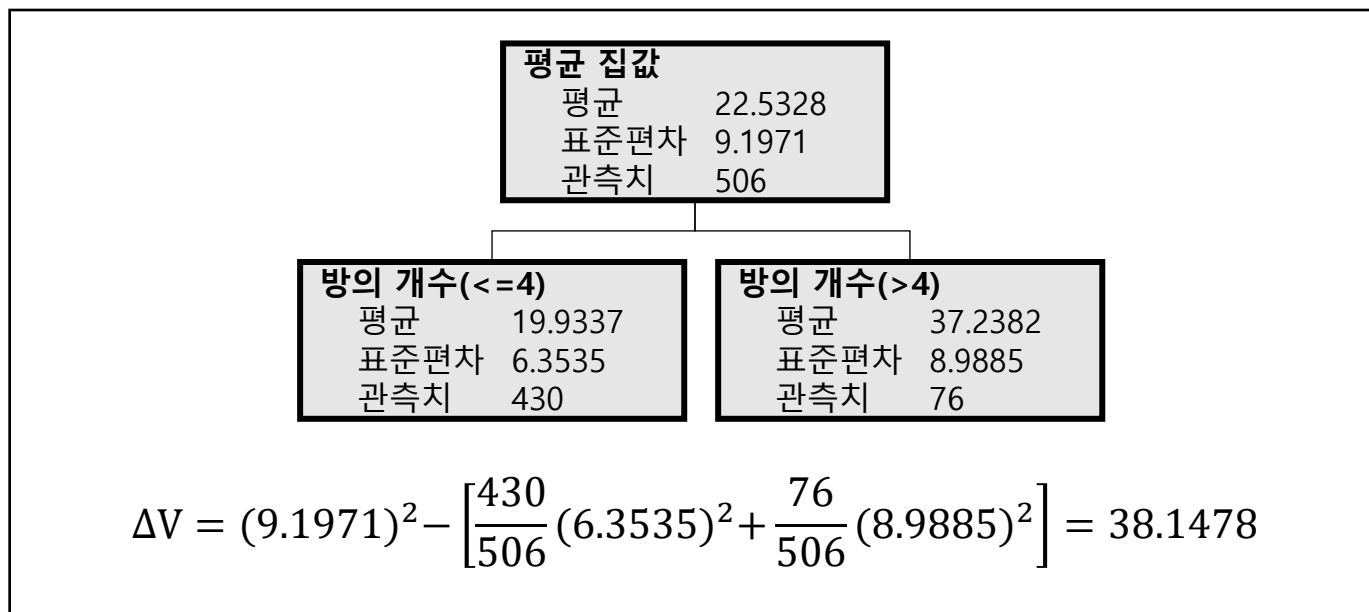




# 분류 - 의사결정나무

## ■ 예제 - 회귀나무

- 목표변수 '평균집값'에 영향을 미치는 설명변수를 선택하기 위한 과정은 아래 그림과 같음



# 비지도 학습 ( $\kappa$ – means Clustering)

# 비지도 학습 - $\kappa$ -means Clustering

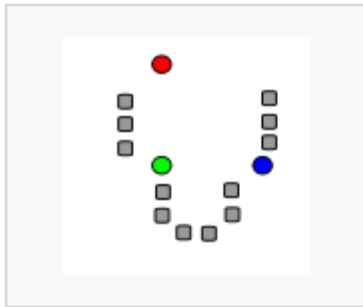
## ■ 군집분석(클러스터링) 정의

- 개인 또는 개체 중에서 유사한 것들을 몇 개의 집단으로 그룹화 하여, 각 집단의 성격을 파악함으로써 데이터 전체의 구조에 대한 이해를 돕고자 하는 탐색적인 분석방법
  - 데이터 자체에만 의존하여 자료를 탐색하고 요약하는 분석기법
  - 사전에 정의된 어떠한 특수한 목적이 없음
    - 예를 들어, '제품 구매/비구매에 가장 많이 영향을 주는 변수는 무엇인가' 와 같은 분석의 목적이 없음
- 전체 데이터를 군집을 통해 잘 구분하는 것이 군집분석의 최대 목적
  - 동일한 군집의 개체들은 유사한 성격을 갖도록, 서로 다른 군집에 속한 개체들은 서로 다른 성격을 갖도록 군집이 형성되어야 함
  - 관찰단위의 성격을 표현하는 알맞은 변수를 선택한 후에, 주어진 변수들을 이용해 각 관찰단위가 서로 얼마나 유사한지 또는 유사하지 않은 지를 측정할 수 있는 측도 필요

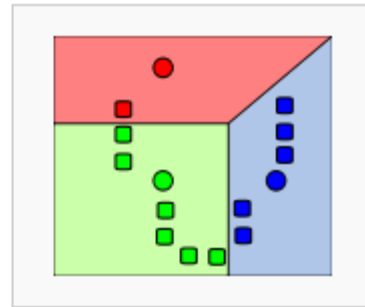
# 비지도 학습 - $\kappa$ -means Clustering

## ■ 절차

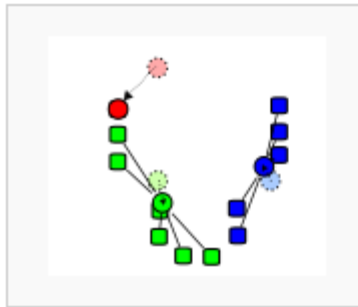
Demonstration of the standard algorithm



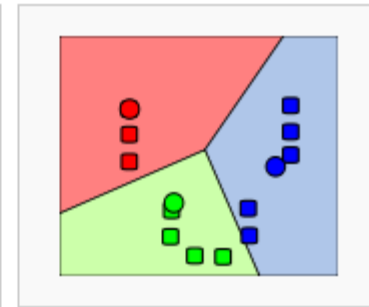
1.  $k$  initial "means" (in this case  $k=3$ ) are randomly generated within the data domain (shown in color).



2.  $k$  clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.



3. The centroid of each of the  $k$  clusters becomes the new mean.



4. Steps 2 and 3 are repeated until convergence has been reached.

군집의 수  $\kappa$  결정 :  $\kappa=3$   
최초 군집 기준값 결정

각 관측 값들을  
가장 가까운 중심의  
군집에 할당

개체를 각 군집에  
할당  
각 군집의 중심  
재계산

개체를 각 군집에  
재할당  
각 군집의 중심 재계산  
경계가 변경되지 않으면  
종료

# Thank you