

머신러닝(Machine Learning)

목 차

1-1 machine learning(기계학습)이란?

1-2 머신러닝으로 무엇이 가능한가?

1-3 용어 이해하기 – 모델, 모델링, 샘플, 데이터 포인트

1-4 머신러닝과 데이터 분석의 비교

1-5 머신러닝의 구분

1-6 지도학습, 비지도학습, 비정형분석

1-7 데이터 마이닝(머신러닝) 수행 단계

01 machine learning(기계학습)이란?

- ▶ 인공지능의 한 분야로, 컴퓨터가 학습 할 수 있는 알고리즘과 기술을 개발하는 분야를 말한다
(위키 백과 참조)



02 머신러닝으로 무엇이 가능한가?

(가) 편지 봉투에 손으로 쓴 우편번호 숫자 판별

(나) 의료 영상 이미지에 기반한 종양 판단

(다) 의심되는 신용카드 거래 감지

=> 신용카드 거래 내역이 입력이 되고 부정 거래인지가 **출력**이 된다.



(라) 블로그 글의 주제 구분

=> 많은 양의 텍스트 데이터를 요약하고 그 안에 담긴 **핵심 주제**를 찾기.

(마) 고객들을 취향이 비슷한 그룹으로 묶기

=> 어떤 고객들의 취향이 비슷한지 비슷한 취향의 고객을 그룹으로 묶고 싶을 때,

02 머신러닝으로 무엇이 가능한가?

(바) 비정상적인 웹 사이트 접근 탐지

=> 정상 패턴과 비정상 패턴을 찾아본다.

03 용어 이해하기

▶ 모델, 모형, 모델링(Modeling, Modeling)

(가) 정보 시스템 모델링,

가. 데이터 모델링 : 현실세계의 복잡한 데이터들을 컴퓨터 정보 구조로 변환시키는 과정

(나) 수학적 모델링,

가. 시스템의 변화를 나타내는(예측하는) 수학적 모델이 방정식으로 표현된다.

(정보통신기술용어해설 참조)

03 용어 이해하기

▶ 데이터 마이닝

(가) 대용량의 데이터로부터 데이터 내에 존재하는 **관계, 패턴, 규칙 등을 탐색하고 모형화**하여 유용한 지식을 추출하는 일련의 과정.

(나) 데이터 마이닝이 소개되기 전의 데이터 분석과 구분 짓는다면, 복잡성 높은 데이터 분석에 **기계학습(machine learning)** 이론이 적용되기 시작함.

(정보통신기술용어해설 참조)



03 용어 이해하기

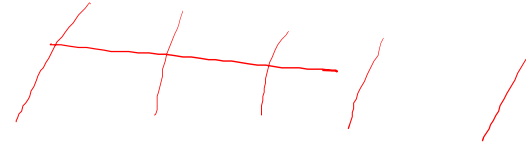
▶ 샘플(sample), 데이터 포인트(data point), 특성(feature)

(가) 샘플(sample) 또는 데이터 포인트

하나의 개체 또는 행을 샘플이라고 말한다.

(나) 특성(feature)

샘플의 속성, 즉 열을 말한다.



(다) 특성 추출(feature extraction) or feature engineering

좋은 입력 데이터를 만들어 내는 것.

04 machine learning(기계학습)이란?

- ▶ 머신러닝과 데이터 마이닝은 종종 같은 방법을 사용.
- ▶ 머신러닝은 훈련 데이터를 통해 학습된 알려진 속성을 기반으로 예측에 초점을 두고 있다.
- ▶ 데이터 마이닝은 데이터의 미처 몰랐던 속성을 발견하는 것에 집중한다.

python

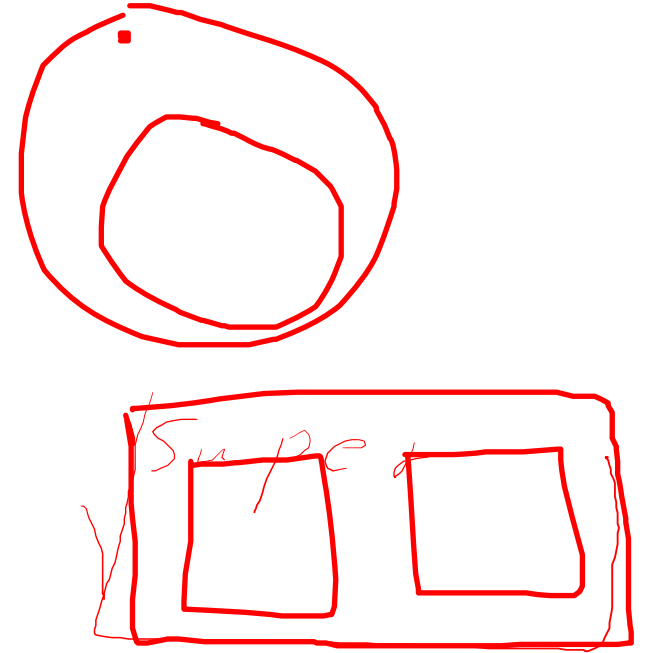
R

(위키 백과 참조)

05 머신러닝의 구분

▶ 지도학습(supervised learning)

- (가) 예측하고자 하는 **목표(Target)가 존재**한다.
학습하고자 하는 데이터의 정답이 있다.
교사의 역할이 존재.



▶ 비지도학습(unsupervised learning) or 자율학습

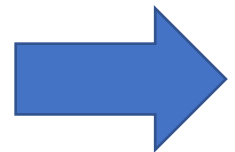
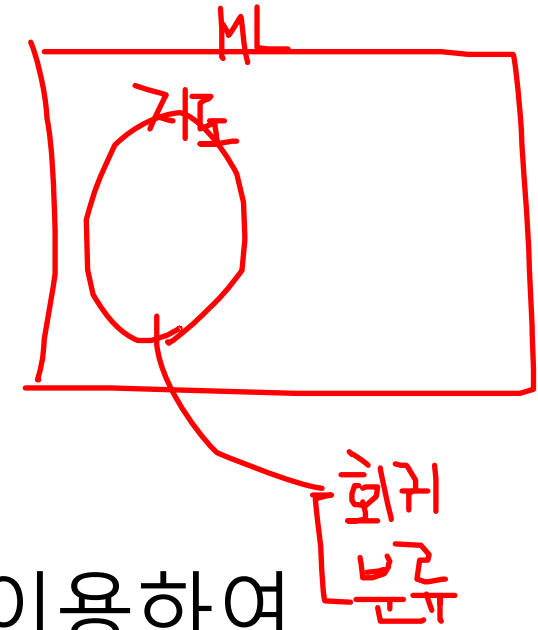
- (나) **목표(Target)가 존재**하지 않는다. 교사 역할이 없음.

05 머신러닝의 구분

▶ 대표적인 지도학습(supervised learning)

(가) Regression – (회귀) – 수치형 변수

(나) Classification – (분류) – 범주형 변수



공통점 : 입력 및 목표 변수의 값을 이용하여
주어진 **입력변수에 대한 목표변수의 값을 예측**하는
모형을 개발한다.

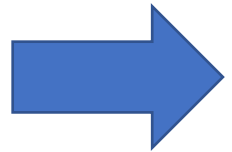


05 머신러닝의 구분

▶ 대표적인 지도학습(supervised learning)

(가) Regression – (회귀) – 수치형 변수

(나) Classification – (분류) – 범주형 변수



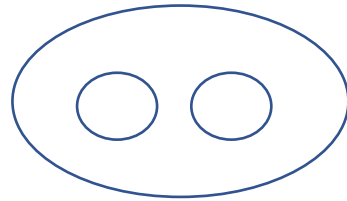
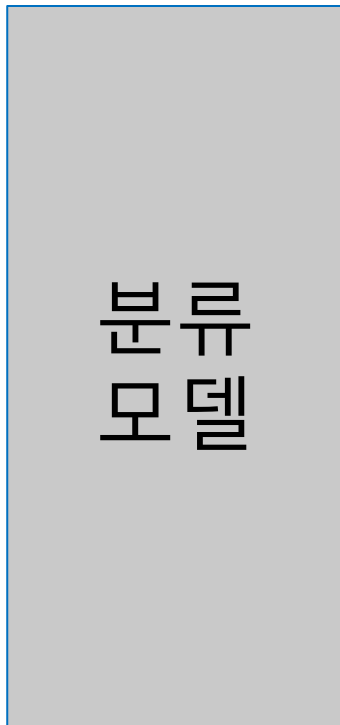
차이점 :

A. 목표 변수의 형태가 회귀의 경우 연속형이다.

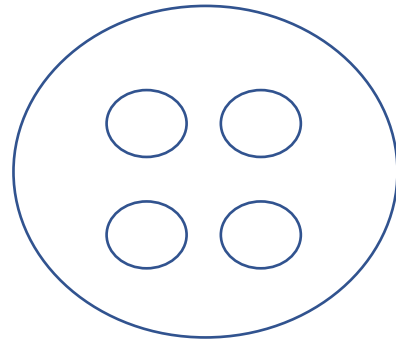
B. 분류의 경우는 범주형이다.(고정되어 있음)

05 머신러닝의 구분

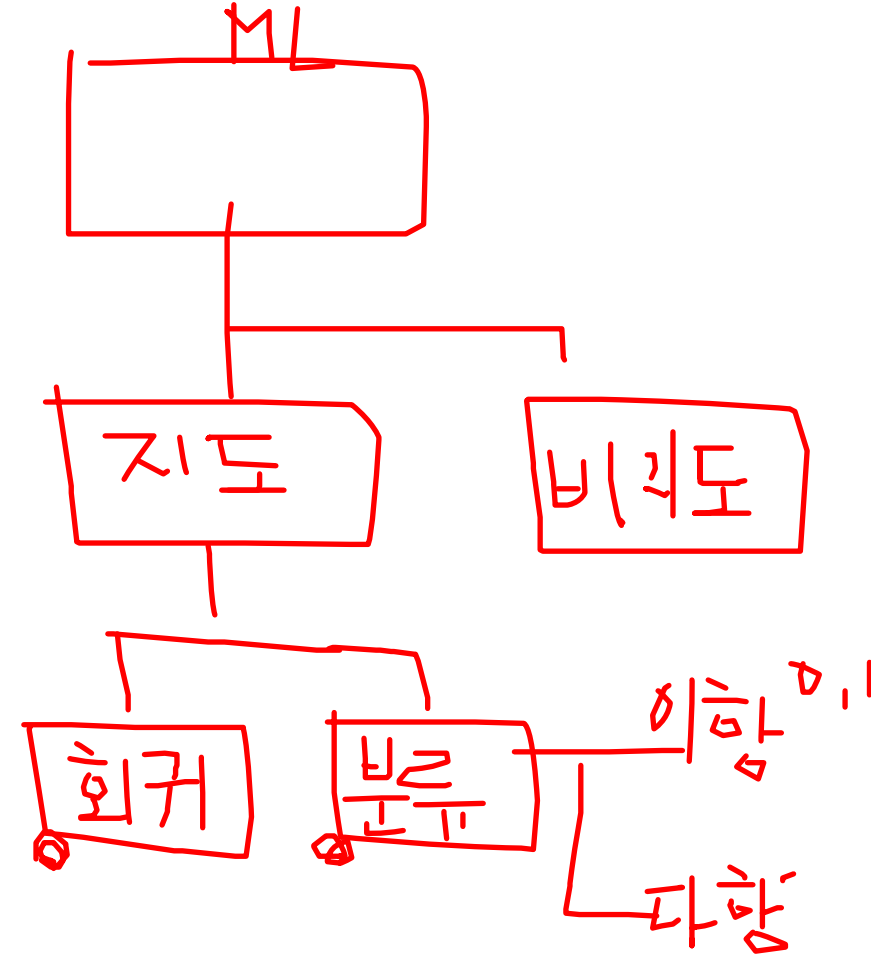
▶ Classification(분류)의 구분 - 이항분류와 다항분류



이항분류
그룹이 2개



다항분류
그룹이 3개 이상

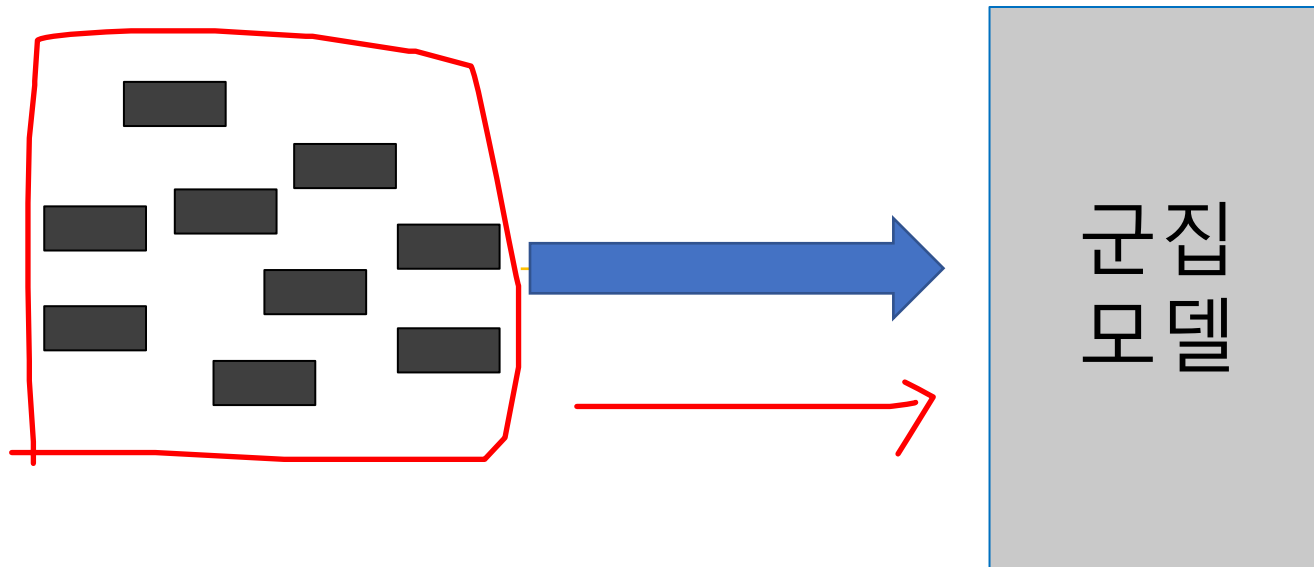


05 머신러닝의 구분 – 비지도학습(unsupervised learning)

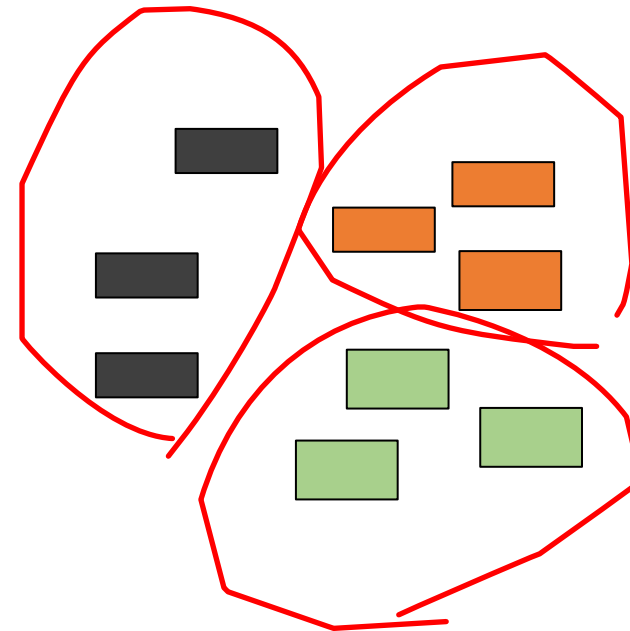
군집은 레이블이 없다.

목표 변수가 없다.

군집은 레이블 없이 확보된
데이터의 특성을 분석



서로 유사한 특성을
가진 데이터끼리 그룹화



06 지도학습, 비지도학습, 비정형분석

지도학습(supervised learning)

Classification

✓ Logistic Regression

✓ SVM
(Support Vector machine)

Naïve Bayes
Classification

✓ Neural Net(신경망)

✓ knn - 최근접이웃기법

✓ Decision Tree

✓ 앙상블(RandomForest)

✓ 앙상블(GradientBoosting)

Regression

Regression

knn - k-최근접이웃기법

Neural Net(신경망)

Decision Tree

앙상블(RandomForest)

앙상블(GradientBoosting)

비지도학습 (unsupervised learning)

Clustering

k-mean(K 평균)

계층적 군집분석

DBSCAN

연관성 분석

장바구니 분석

서열 분석

트랜잭션 데이터분석

06 지도학습, 비지도학습, 비정형분석

지도학습

분류분석

회귀분석

비지도학습

군집 분석

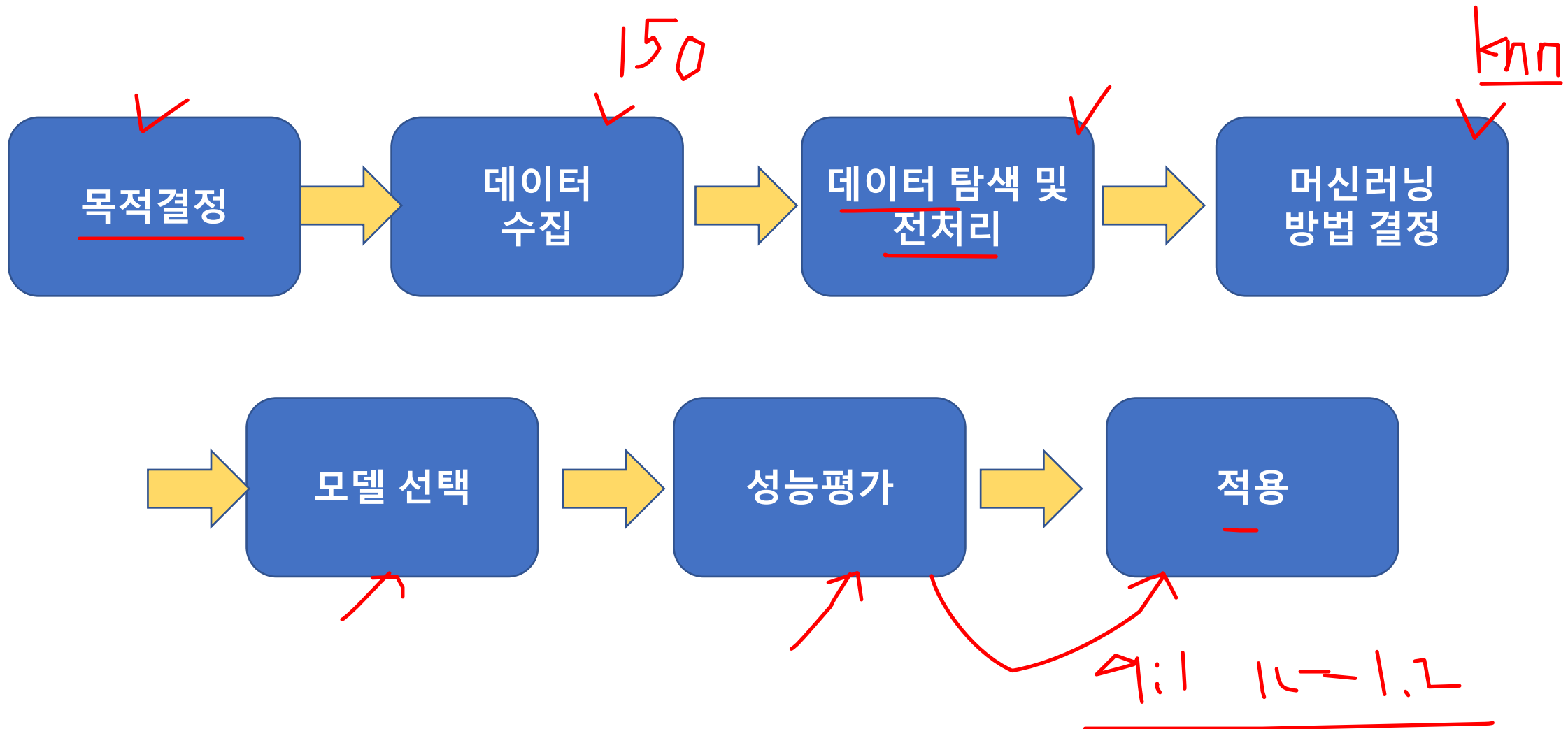
연관성 분석

비정형
분석

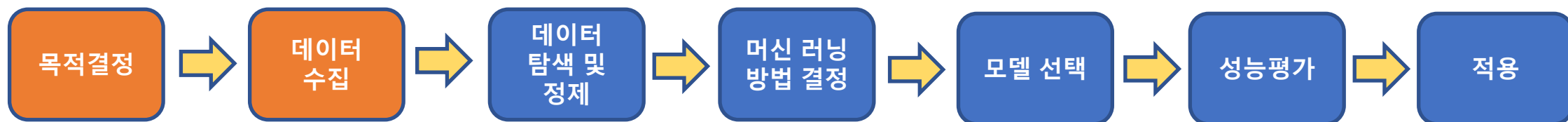
텍스트마이닝

사회연결망 분석

07 데이터 마이닝의 수행단계



07 데이터 마이닝의 수행단계



목적결정

▶ 프로젝트 목적을 계획하고 설정하는 단계

목적을 정하고 관련 데이터를 수집하기도 하지만,
때로는 데이터 수집 후 탐색 과정을 거쳐 문제가 설정되기도 한다.

데이터 수집

▶ 데이터 베이스 또는 분산된 데이터 베이스 이용

▶ 외부 데이터, 내부 데이터

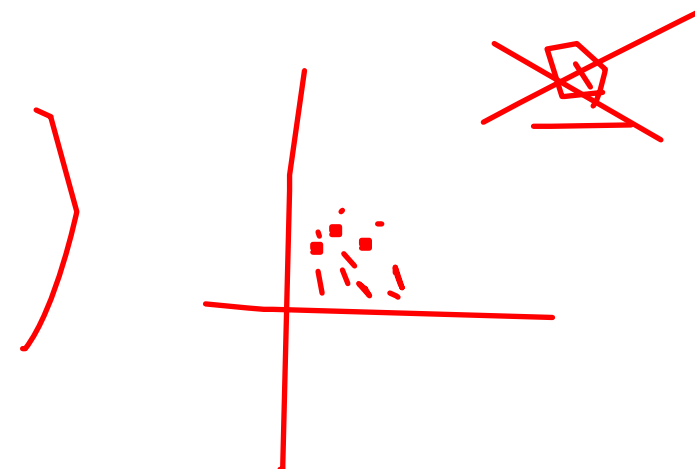
07 데이터 마이닝의 수행단계



데이터탐색 및 정제

▶ 데이터 표준화 및 점검 (시각화)

- (1) 데이터에 결측치가 존재하는지,
- (2) 모든 값에 상식적인 범위 내에 있는지,
- (3) 이상치가 존재하는지,

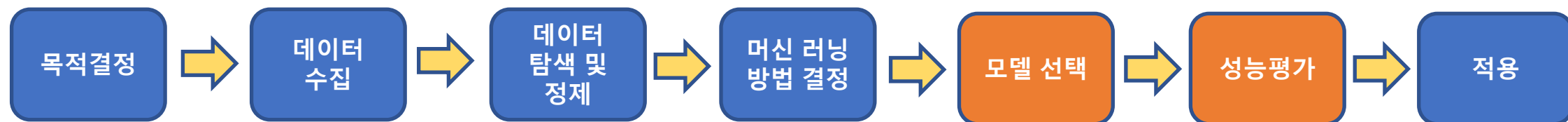


머신 러닝 방법 결정

▶ 머신 러닝 도전 과제(분류, 회귀, 군집화 등)을 결정

▶ 머신 러닝 기법(로지스틱 회귀, 신경망, 계층 군집 등)을 선택하는 단계

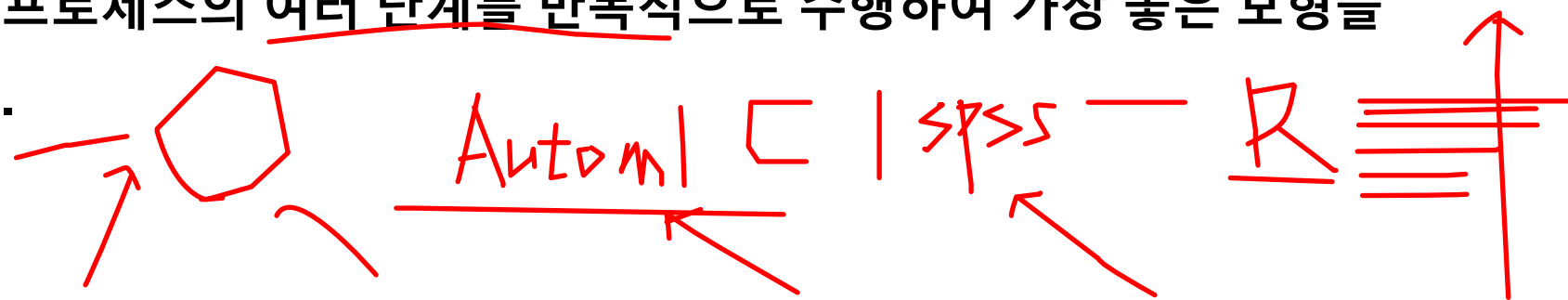
07 데이터 마이닝의 수행단계



모델 선택

▶ 가장 좋은 모형을 찾는 단계

머신러닝 프로세스의 여러 단계를 반복적으로 수행하여 가장 좋은 모형을 찾는 단계.

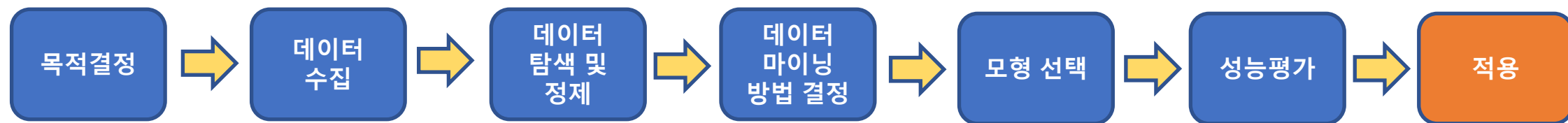


성능평가

▶ 검증 데이터(테스트 데이터)를 이용하여 구축된 모형의 성능을 평가하여 효율적인 모형을 찾는다.

▶ 예측력이 가장 우수한 것을 선택하여 최종 모형 선정

07 데이터 마이닝의 수행단계



적용

▶ 구축된 모형을 운용 시스템에 탑재 후, 실제 의사결정에 적용.

(예제) **구축된 모형을 적용**하여 **구매가능성이 높은 고객을 결정**하고 해당 **고객에게 구매권유 메일**을 보내어 수익창출 가능성을 높인다.

08 History

2019.09.01 06 수행 절차 내용 추가

2020.10.21 전반적인 내용 업데이트, 용어 이해 추가