빅 데이터 분석 개요

한눈에 보는 머신러닝

- 머신 러닝이란?
- 왜 머신러닝을 사용하는가?
- 애플리케이션 사례
- 머신 러닝 시스템의 종류
  - 지도 학습과 비지도 학습
  - 배치 학습과 온라인 학습
  - 사례 기반 학습과 모델 기반 학습
- 머신러닝의 주요 도전 과제
  - 충분하지 않은 양의 훈련데이터
  - 대표성 없는 훈련 데이터
  - 낮은 품질의 데이터
  - 관련 없는 특성
  - 훈련 데이터 과대적합
  - 훈련 데이터 과소적합
  - 한걸음 물러서서
- 테스트와 검증
  - 하이퍼파라미터 튜닝과 모델선택

머신러닝이란?

머신러닝이란?

정의자	내용
아서 새뮤얼, 1959	명시적인 프로그래밍 없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구영역
톰 미첼, 1997	어떤 작업 T 에 대한 프로그램의 성능을 P로 측정했을 때 경험 E로 인해 성능이 향상됐다면 이 컴퓨터 프로그램은 작업 T와 성능측정 P에 대해 경험 E로 학습한 것이다.
일반적 정의	데이터에서부터 학습하도록 컴퓨터를 프로그래밍하는 과학
	애플리케이션을 수정하지 않고도 데이터를 기반으로 패턴을 학습하고 결과를 예측하는 알고리즘 기법을 통칭함

머신러닝이란?

스팸 메일 케이스

용어	내용
훈련 정의	스팸 메일과 일반 메일의 샘플을 이용해 스팸 메일 구분법을 배울 수 있는 머신 러닝 프로그램
훈련 세트	시스템이 학습하는 데 사용하는 샘플 세트
훈련 사례(샘플)	훈련 세트에 포함되어 있는 각각의 훈련 데이터
작업(T)	새로운 메일이 스팸 메일인지 구분하는 작업
경험(E)	스팸메일 분류기의 훈련에 사용된 메일의 샘플-훈련데이터
측정 기준(P)	정확도:정확히 분류된 메일의 비율, 분류작업에 주로 사용

머신러닝이란?

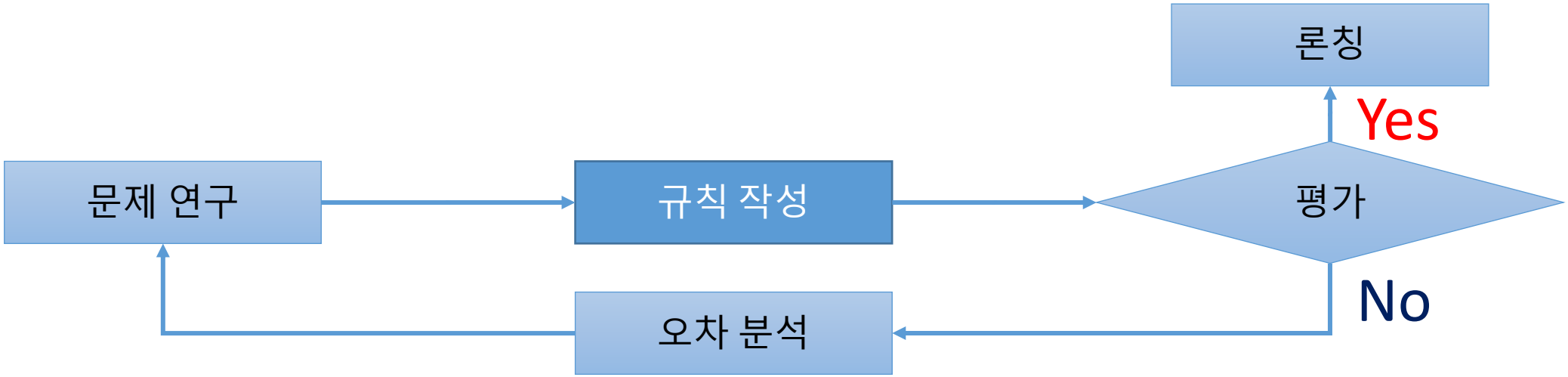
금융 사기거래 케이스

구분	내용
기존 프로그래밍 방식의 해결 방안	복잡한 금융 거래에서 발생한 수많은 변수에 대해 수십년간 발생한 다양한 시기 거래 조건을 감안하여 수천-수만 라인의 소스코드로 된 프로그램을 작성하여 사기거래 적발 프로그램을 구축
효율 저하	금융사기 전문가들은 경험적으로 프로그램 로직을 간파하고 풀어냄
환경 변화	수시로 변화하는 금융환경, 정부정책, 소비자 성향 등에 맞추어 기존 로직을 다시 수정하고 검증하는 작업을 거쳐야 함
머신러닝 해결방안	데이터를 기반으로 통계적인 신뢰도를 강화하고 예측 오류를 최소화하기 위한 다양한 수학적 기법들을 적용해 데이터 내의 패턴을 스스로 인지하고 신뢰도 있는 예측 결과를 도출

왜 머신 러닝을 사용하는가?

스팸필터 제작 재구성

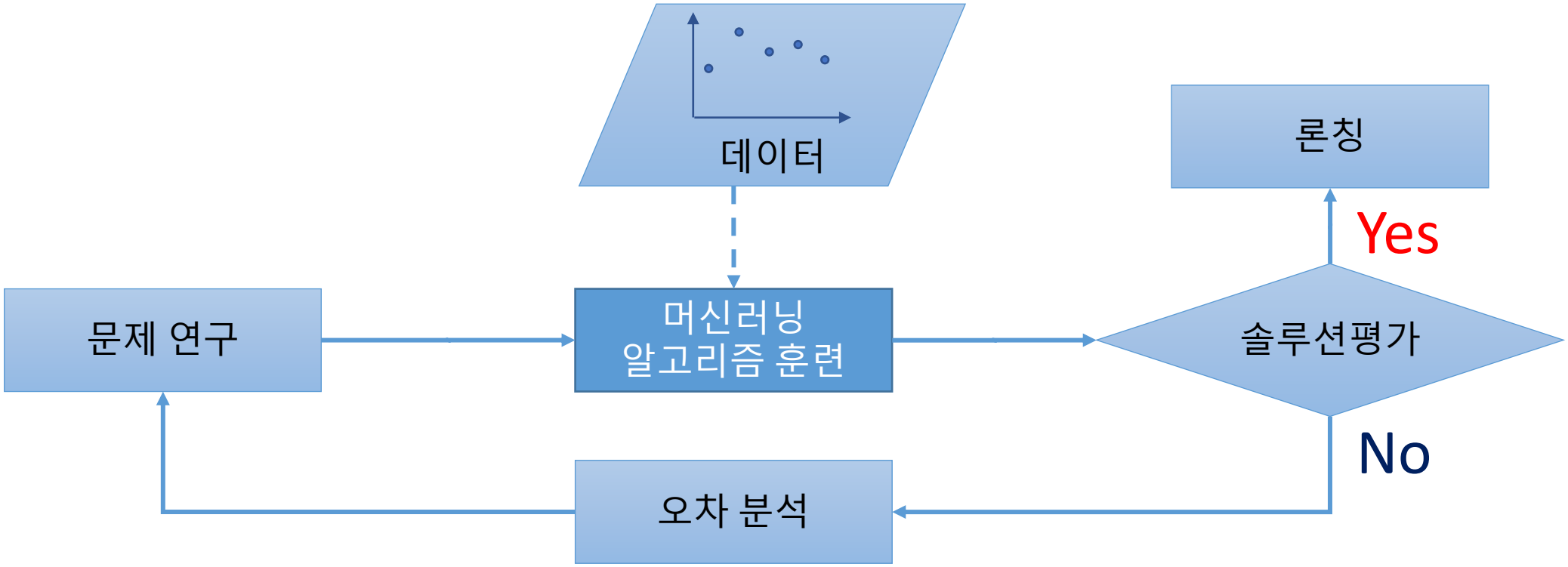
구분	내용
스팸 메일의 특성 탐색	스팸에 어떤 단어들이 나타나는 지 관찰 - 신용카드, 무료, 혜택, 상관없이 와 같은 단어들이 빈번하게 등장 - 발송자의 이름, 메일 주소, 본문이나 다른 요소들의 패턴도 감지 가능
패턴인식 결과에 따른 분류	발견된 각 패턴을 감지하는 규칙을 작성하여 프로그램이 이런 패턴을 발견할 때 그 메일을 스팸으로 분류
탐색과 분류 작업 반복	프로그램을 테스트하고 론칭할 만큼 충분한 성능이 나올 때까지 탐색과 분류작업 반복



왜 머신 러닝을 사용하는가?

머신 러닝 방식 스팸 필터

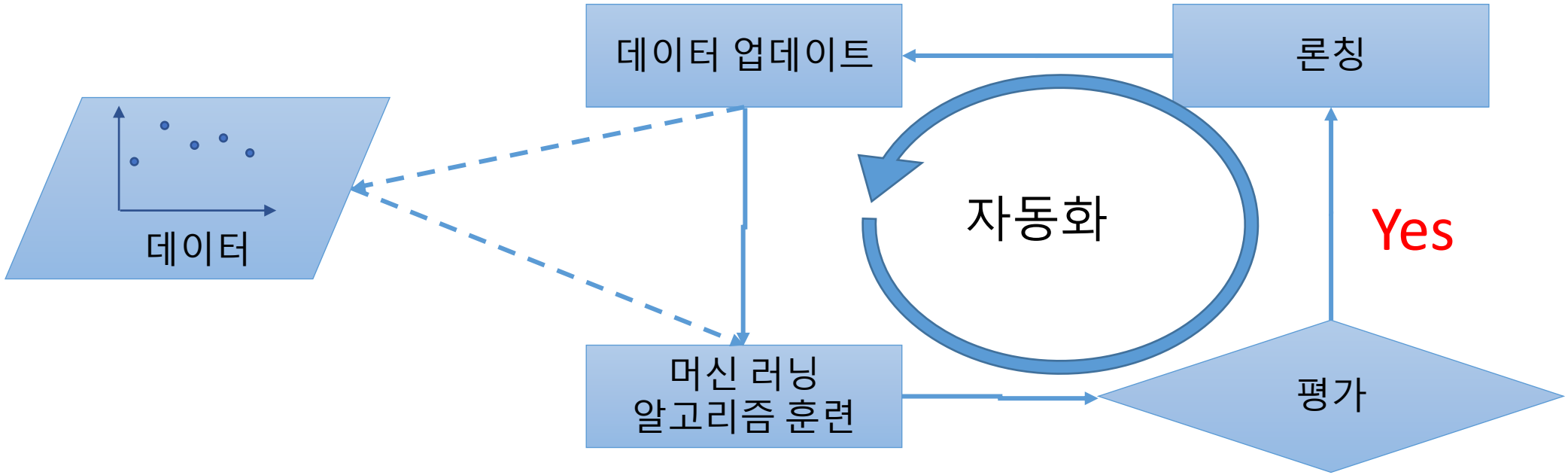
- 일반 메일과 스팸 메일을 비교하여 스팸 메일에 나타나는 패턴을 감지하여 어떤 단어와 구절이 스팸 메일을 판단하는데 좋은 기준인지 자동으로 학습하여 스팸 메일을 필터링함
- 패턴 발견이 자동화되어 프로그램이 짧아지고 보수하기 쉬우며 정확도가 더 높아짐



왜 머신 러닝을 사용하는가?

자동화된 변화적응

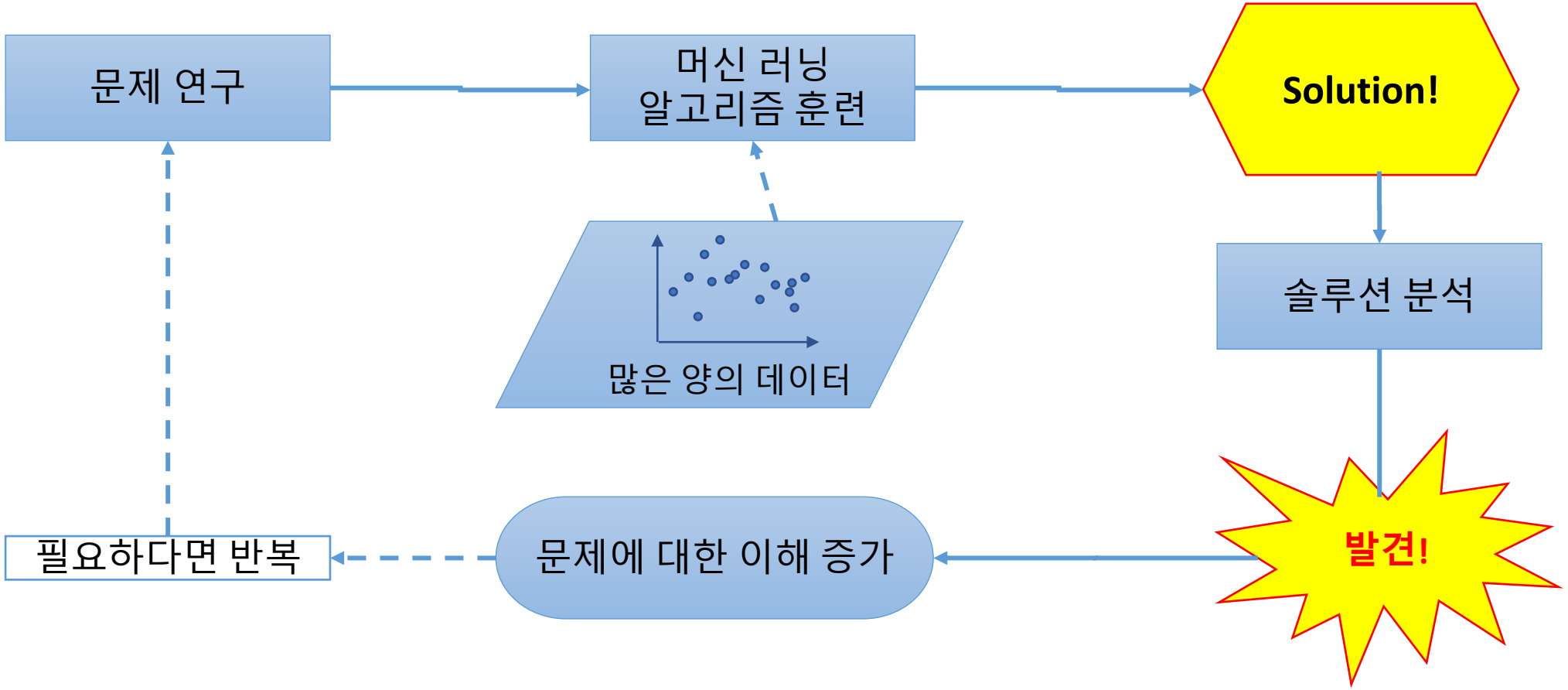
- 스팸메일 발송자가 단어의 사용패턴을 변화시키면 전통적인 프로그래밍 방식의 필터는새로운 데이터를 구분하기 위해 수정이 필요함→ 발송자가 패턴을 계속 변경하면 새론운 규칙을 계속 추가
- 사용자가 스팸으로 지정한 메일에 반복되는 패턴을 자동으로 인식하여 별도의 작업을 하지 않아도 이 메일을 스팸으로 인식하여 분류



왜 머신 러닝을 사용하는가?

너무 복잡하거나 알려진 알고리즘이 없는 경우-음성인식 사례

- 음성의 특징을 기준으로 사운드의 강도를 측정하는 알고리즘을 통해 하드코딩으로 인식 프로그램 작성→소음이 있는 상황이나 여러 언어를 구분하는 환경에서는 사용 불가
- 단어의 발음을 녹음한 샘플을 통해 스스로 학습하게 하는 것이 최선의 방안



왜 머신 러닝을 사용하는가?

머신러닝이 필요한 분야

구분	내용
기존 솔루션으로는 많은 수동 조정과 규칙이 필요한 문제	하나의 머신 러닝 모델이 코드를 간단하게 만들고 전통적인 기법보다 더 효과적으로 수행됨
전통적인 방법으로는 해결방법이 없는 복잡한 문제	가장 뛰어난 머신 러닝 방법으로는 해결 방법을 찾을 수 있음
유동적인 환경	머신 러닝 시스템은 새로운 데이터에 적응이 가능
복잡한 문제와 대량의 데이터에서 통찰 얻기	문제가 정의 되지 않은 경우에도 데이터의 분석을 통한 문제의 발견과 통찰을 얻을 수 있음

애플리케이션 사례

대표적인 머신 러닝 작업의 사례와 이를 위한 기술

활용 사례	분석 기밥
생산라인에서 제품 이미지를 분석해 자동으로 분류	• 이미지 분류 작업 • 합성곱 신경망 Convolutional Neural Network(CNN) 사용
뇌를 스캔하여 종양 진단	• 시맨틱 분할 작업 • 합성곱 신경망을 사용하여 이미지의 각 픽셀을 분류
자동으로 뉴스 기사를 분류	• 자연어 처리 Natural Language Processing(NLP) 작업 • 텍스트 분류작업 • 순환 신경망 Recurrent Neural Network(RNN), 합성곱 신경망, 트랜스포머 Transformer 사용
토론 포럼에서 부정적인 코멘트를 자동으로 구분	• 텍스트 분류 작업 • 자연어 분류 활용
긴 문서를 자동으로 요약	• 텍스트 요약 • 자연언어처리의 한 분야
챗봇 또는 개인 비서 만들기	• 자연어 이해 Natural Language Understanding (NLU)와 질문-대답 Question-Answer 모듈을 포함한 자연어 처리 활용

애플리케이션 사례

대표적인 머신 러닝 작업의 사례와 이를 위한 기술

활용 사례	분석 기밥
다양한 성능 지표를 기반으로 회사의 내년도 수익을 예상	• 회귀 Regression 작업(결과 값 예측이 숫자로 표시됨) • 선형회귀 Linear Regression, 다항회귀 Polynomial Regression, 회귀SVM, 회귀 랜덤 포레스트 Random Forrest, 인공 신경망 Artificial Neural Network • 지난 성능 지표를 사용할 경우 순환 신경망, 합성곱 신경망, 트랜스포머 사용
음성 명령에 작동하는 앱 제작	• 음성인식 작업(오디오 샘플 처리) • 순환 신경망, 합성곱 신경망, 트랜스포머 사용
신용카드 부정거래 감지	• 이상치 탐지 작업
구매이력을 기반으로 고객을 나누고 각 집합마다 다른 마케팅 전략 기획	• 군집 Clustering 작업
고차원의 데이터셋을 명확하고 의미있는 그래프로 표현	• 데이터 시각화 작업 • 차원 축소 Dimensionality Reduction 사용

애플리케이션 사례

대표적인 머신 러닝 작업의 사례와 이를 위한 기술

활용 사례	분석 기밥
과거 구매 이력을 기반으로 고객이 관심을 가질 수 있는 상품 추천	• 추천시스템 • 과거 구매이력 및 고객 정보를 인공신경망에 학습시키고 구매 가능성이 가장 높은 상품을 출력하는 것 • 모든 고객의 구매 이력을 기반으로 훈련
지능형 게임 만들기	• 강화 학습으로 해결 • 시간이 지남에 따라 주어진 환경에서 보상이 최대가 되는 행동을 선택하는 봇과 같은 에이전트를 훈련하는 머신 러닝의 한 분야 • 알파고가 강화 학습을 통하여 구축됨

머신 러닝의 종류

머신 러닝의 여러가지 분류 기준과 이에 따른 분류

구분 기준	분석 기밥
사람의 감독하에 훈련을 하는 것인지 감독 없이 훈련하는 것인지	• 지도학습 • 비지도학습 • 준지도학습 • 강화학습
실시간으로 점진적인 학습을 하는지 아닌지	• 온라인 학습 • 배치 학습
단순하게 알고 있는 데이터 포인트와 새 데이터 포인트를 비교하는 것인지 아니면 과학자들이 하는 것처럼 훈련 데이터셋에서 패턴을 발견하여 예측 모델을 만드는지	• 사례기반 학습 • 모델기반 학습

머신 러닝의 종류

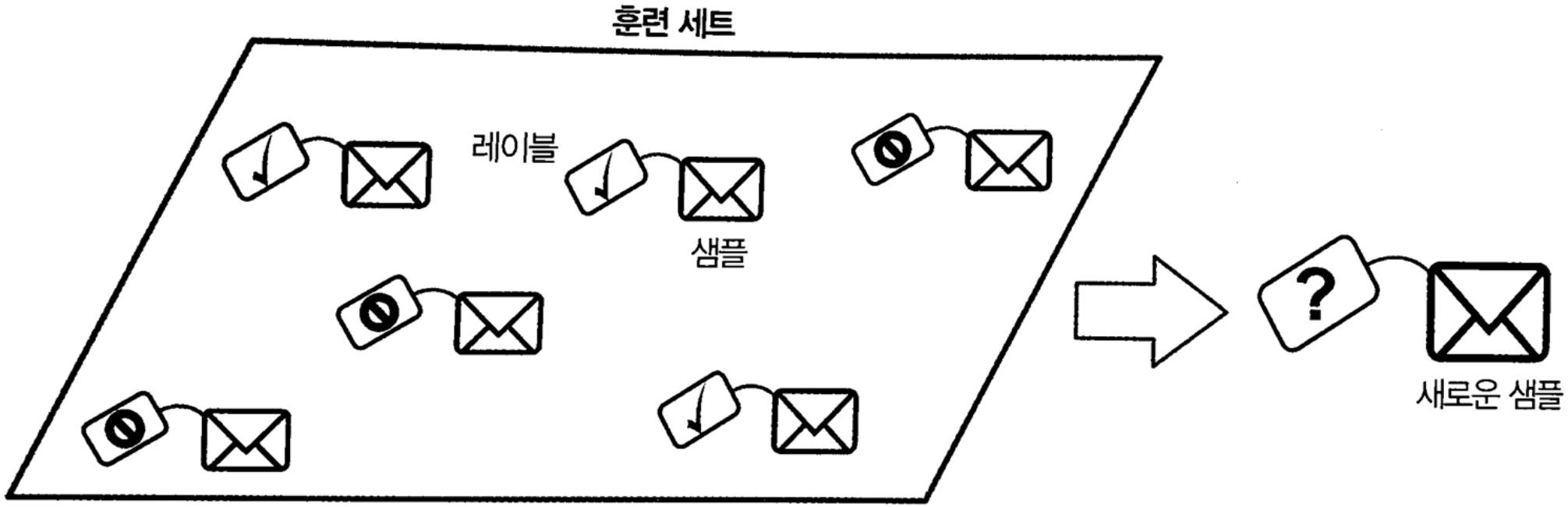
지도학습과 비지도학습

	비지도학습 Unsupervised Lerning	지도 학습 Supervised Learning
정의	• 데이터 분석의 목적이 명확히 정의된 형태의 특정 필드 값을 구하는 것이 아니라 데이터 자체의 결합, 연관성, 유사성 등을 중심으로 데이터의 상태를 표현하는 것 • 데이터 마이닝의 중심이 되는 방법 중 자료의 출력변수가 없이 입력 변수만 주어지는 경우 • 입력 변수간의 상호관계나 입력 자료값 간의 관계를 탐색적으로 분석할 때 사용되는 학습 방법	• 명확한 목적 하에 데이터 분석을 실시하는 것 • 분류, 예측, 최적화를 통해 사용자의 주도하에 분석을 실시하고 지식을 도출하는 것이 목적 • 데이터 마이닝의 중심이 되는 학습방법 중 자료가 입력 변수와 출력 변수로 주어지며 입력 변수와 출력 변수의 함수적 의존 관계를 자료로부터 추정함으로써 예측 모형을 얻을 때 사용되는 학습방법
예	• 장바구니분석 • 군집 분석 • 기술 통계 및 프로파일링	• 분류 • 시계열 분석 • 선형회귀

머신 러닝의 종류

지도학습

- 알고리즘에 주입하는 훈련 데이터에 레이블Label이라는 원하는 답이 포함됨



머신 러닝의 종류

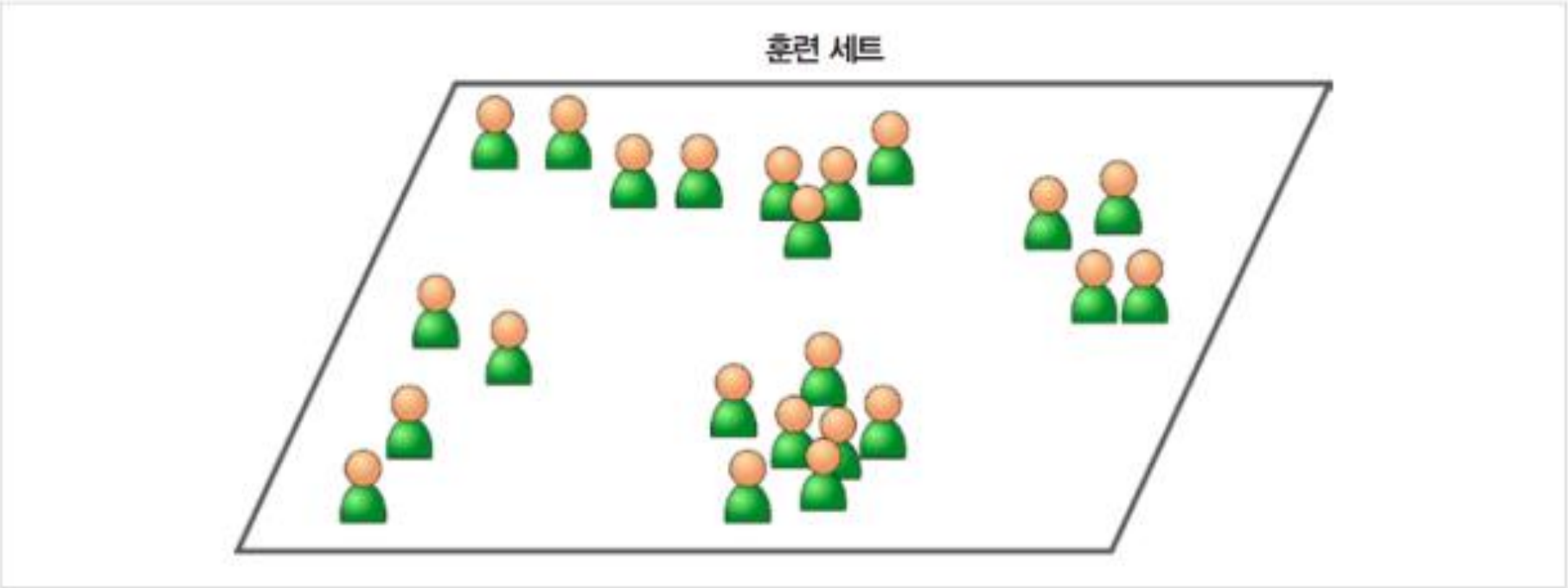
지도학습

구분	내용
분류 Classification	• 전형적인 지도학습 • 스팸 필터가 대표적인 예로 많은 메일 샘플과 소속 정보(스팸 메일인지 아닌지)호 훈련하며 새로운 메일이 들어왔을 때 훈련된 모델에 따라 스팸 메일 혹은 일반 메일로 분류
회귀 Regression	• 예측변수 Prediction Variable 이라 부르는 특성 Feature (주행거리, 연식, 브랜드 등)을 사용해 중고차 가격과 같은 타깃 Target 수치를 예측하는 것 • 시스템을 훈련하기 위해서 주행거리, 연식, 브랜드 등과 같은 예측변수와 레이블(중고차 가격)이 포함된 중고차 훈련 데이터가 필요
그밖의 중요한 알고리즘들	• K 최근접 이웃 K-nearest Neighbors • 선형 회귀 Linear Regression • 로지스틱 회귀 Logistic Regression • 서포트 벡터 머신 Support Vector Machine (SVM) • 결정 트리Decision Tree와 랜덤 포레스트 Random Forrest • 신경망 Neural Network

머신 러닝의 종류

비지도학습

- 훈련 데이터에 정답(레이블)이 없어서 시스템이 아무런 도움 없이 스스로 학습을 진행



머신 러닝의 종류

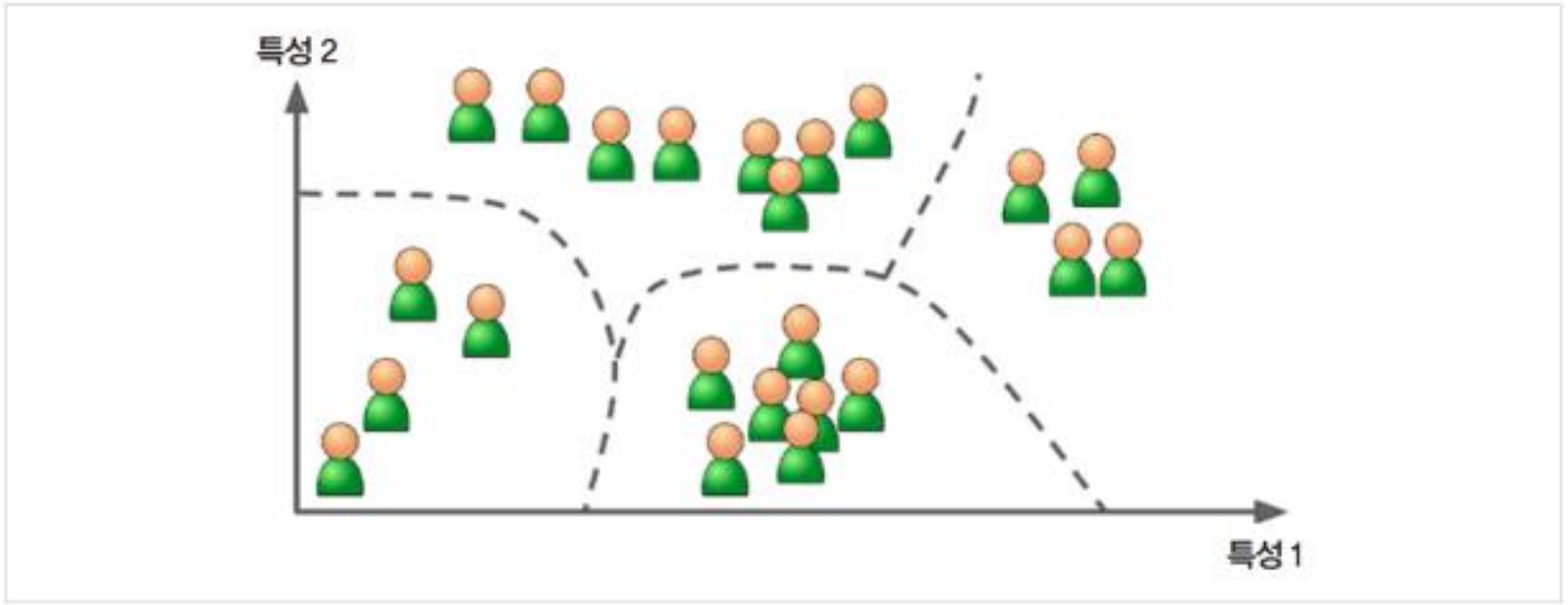
비지도학습

구분	내용
군집 Clustering	• K 평균 K-means • DBSCAN (Density-based Spatial Clustering of Application with Noise) • 계층 군집분석 Hierachical Cluster Analysis(HCA) • 이상치 탐사 Outlier Detection 과 특이치 탐색Novelty Detection • 원클래스 One-class SVM • 아이솔레이션 포레스트 Isolation Forrest
시각화 Visualization 차원 축소 Dimensionality Reduction	• 주성분 분석Principal Component Analysis (PCA) • 커널kernel PCA • 지역적 선형 임베딩Locally-Linear Embedding (LLE) • t-SNE(t-distributed Stochastic Neighbor Embedding)
연관 규칙 학습 Association Rule Learning	• 어프라이어리Apriori • 이클렛Eclat

머신 러닝의 종류

분류

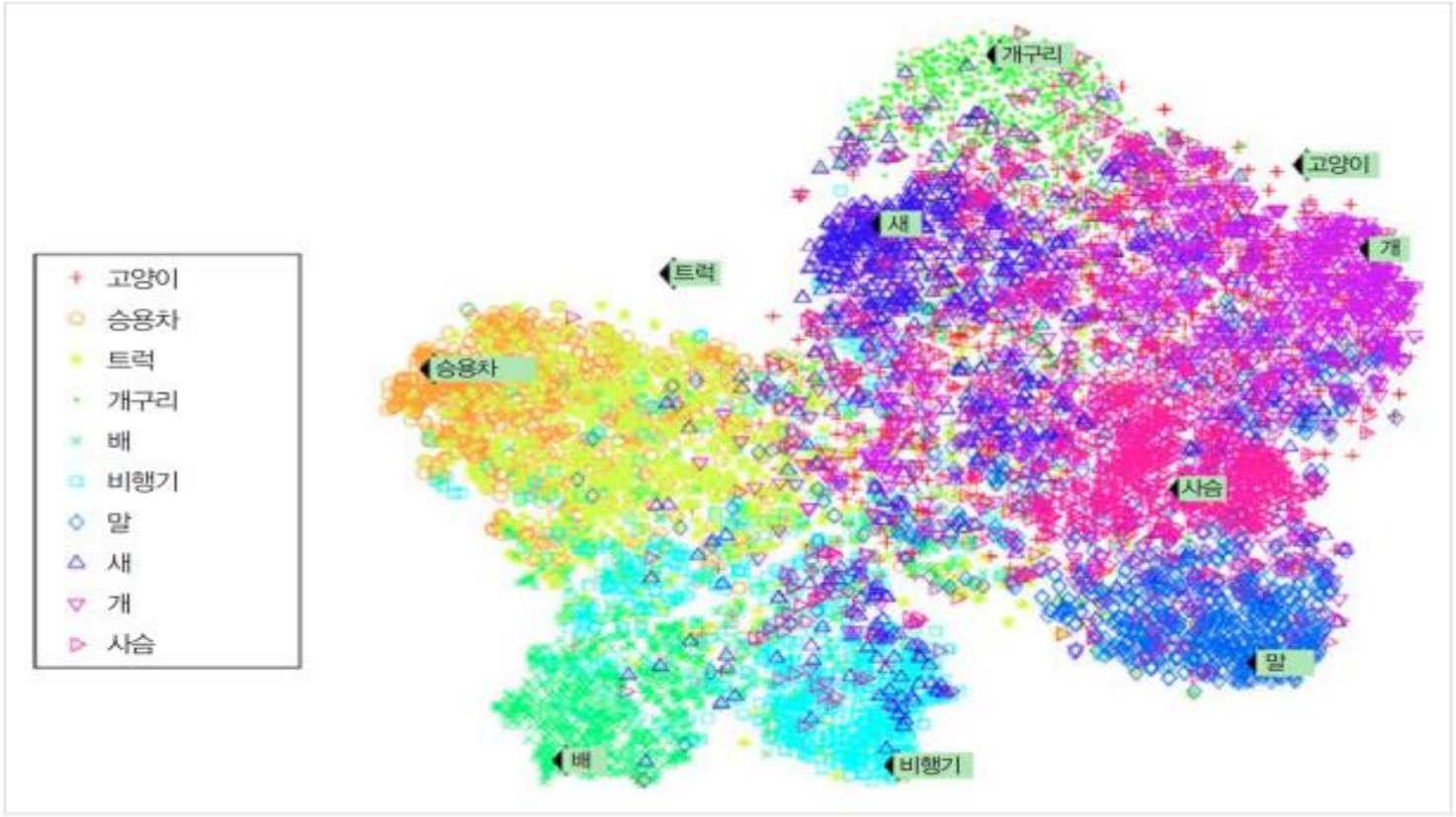
- 블로그 방문자를 그룹별로 분류하려고 할 때 방문자들이 어떤 그룹에 속하는 지를 알려줄 만한 데이터가 존재하지 않음
- 방문자의 다양한 특성 데이터를 사용하여 알고리즘이 스스로 방문자 사이의 연결고리를 찾아 분류를 수행함



머신 러닝의 종류

시각화와 차원 축소

- 정답(레이블)이 없는 대규모의 고차원 데이터를 넣어서 2D나 3D 이미지를 구성해 줌
- 데이터의 구조가 원형대로 잘 유지되므로 데이터의 조직화 내용이나 새로운 패턴을 발견하는 것이 가능
- 데이터가 가지고 있는 정보를 최대한 보존하면서 데이터를 간소화 하는 것을 **차원 축소**라고 하며 서로 상관관계가 있는 특성을 합치는 것을 특성추출이라고 함

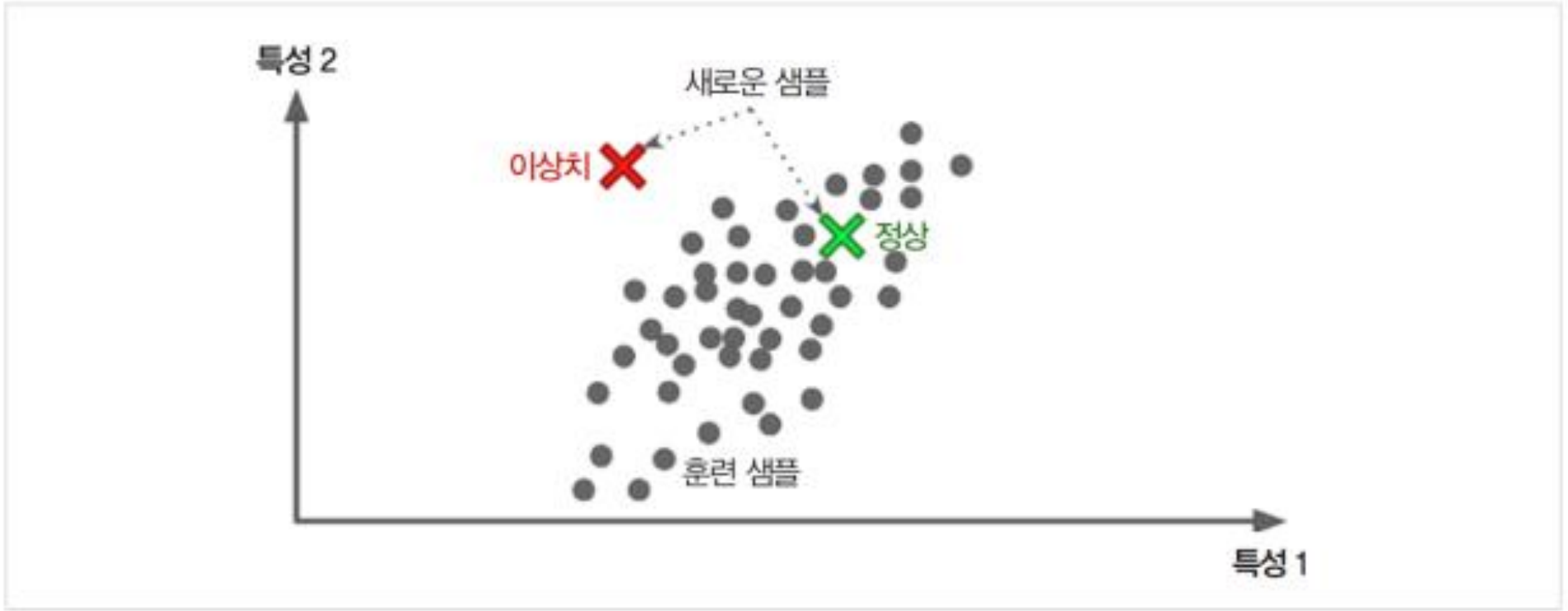


소셔(Socher), 간주(Ganjoo), 매닝(Manning), 앤드류 응(Andrew Ng)의 「T-SNE visualization of the semantic word space」(2013)

머신 러닝의 종류

이상치 탐지

- 시스템은 정상적인 데이터를 학습하고 새로운 데이터가 주어졌을 때 정상 데이터인지 이상 데이터인지 판별
- 부정 금융거래 감지. 불량품 탐지, 학습 데이터 입력 전에 데이터의 이상한 값 탐지 등에 사용
- 특이치 탐지 : 훈련 세트에 있는 데이터와 달라 보이는 새로운 샘플을 탐지하는 것이 목적
- 연관규칙학습: 대량의 데이터에서 특성간의 새로운 관계를 찾아내는 것



머신 러닝의 종류

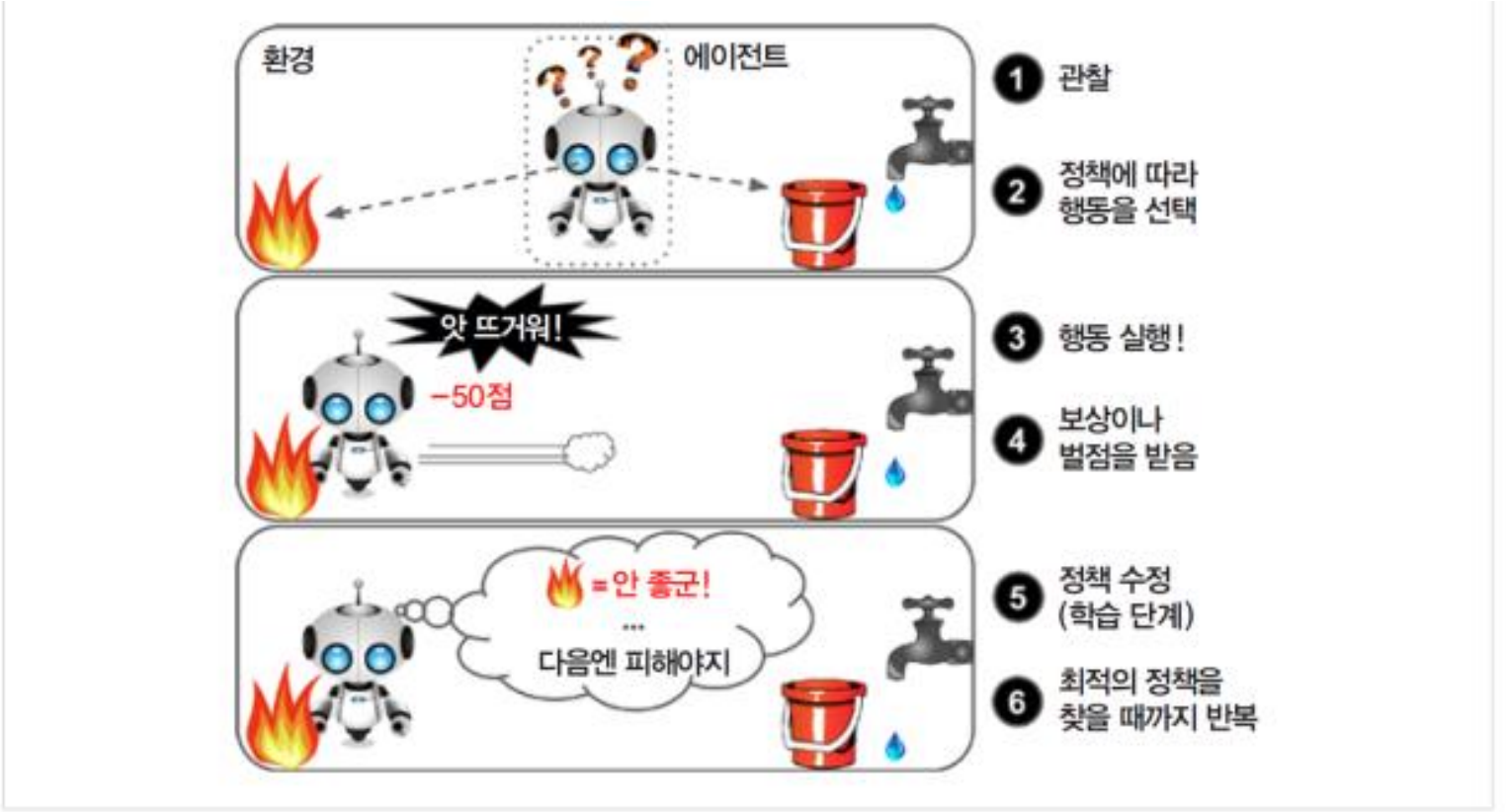
준지도학습 Semisupervised Learning

- 데이터에 정답(레이블)을 다는 작업은 시간과 비용이 많이 소요됨→레이블이 없는 샘플이 대부분
- 일부만 정답(레이블)이 붙어있는 데이터를 다루는 것이 준 지도 학습
- 구글 포토의 경우 사진 속의 인물 데이터의 분류는 비지도 학습으로 수행하지만 인물 하나에 정답(레이블)을 붙이면 같은 사람으로 분류된 모든 사진들에 이름표가 붙는다.

머신 러닝의 종류

강화학습 Reinforcement Learning

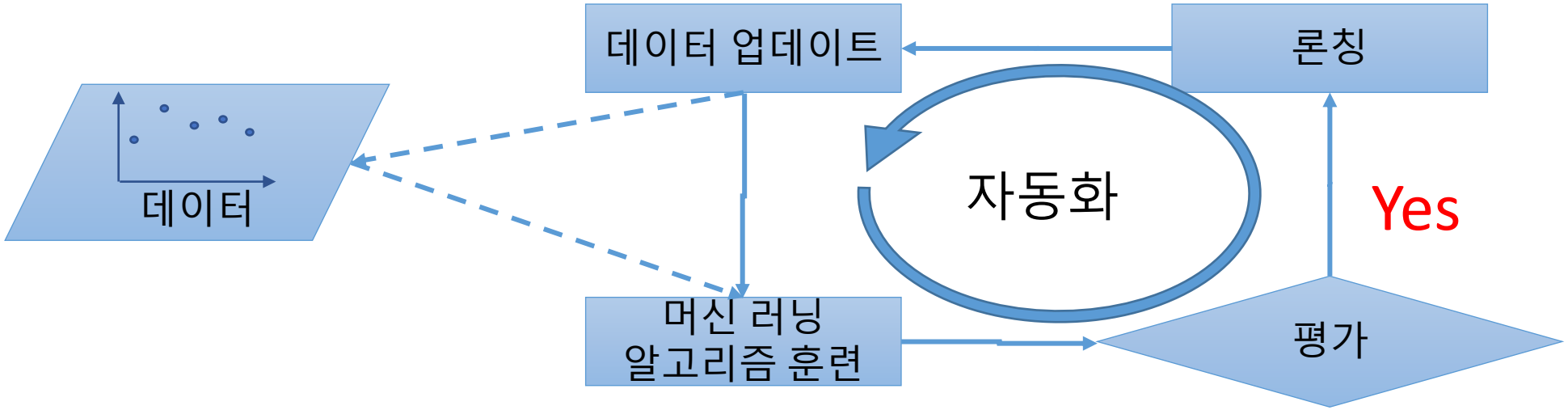
- 에이전트: 학습하는 시스템
- 보상과 벌점: 에이전트가 상황을 관찰해서 실행하고 나서 받는 긍정적, 부정적 피드백
- 정책: 지속적인 학습의 결과로 에이전트가 최상의 보상을 얻기 위해 만들어내는 전략
- 알파고나 보행로봇의 훈련이 대표적인 사례



머신 러닝의 종류

배치학습 batch Learning

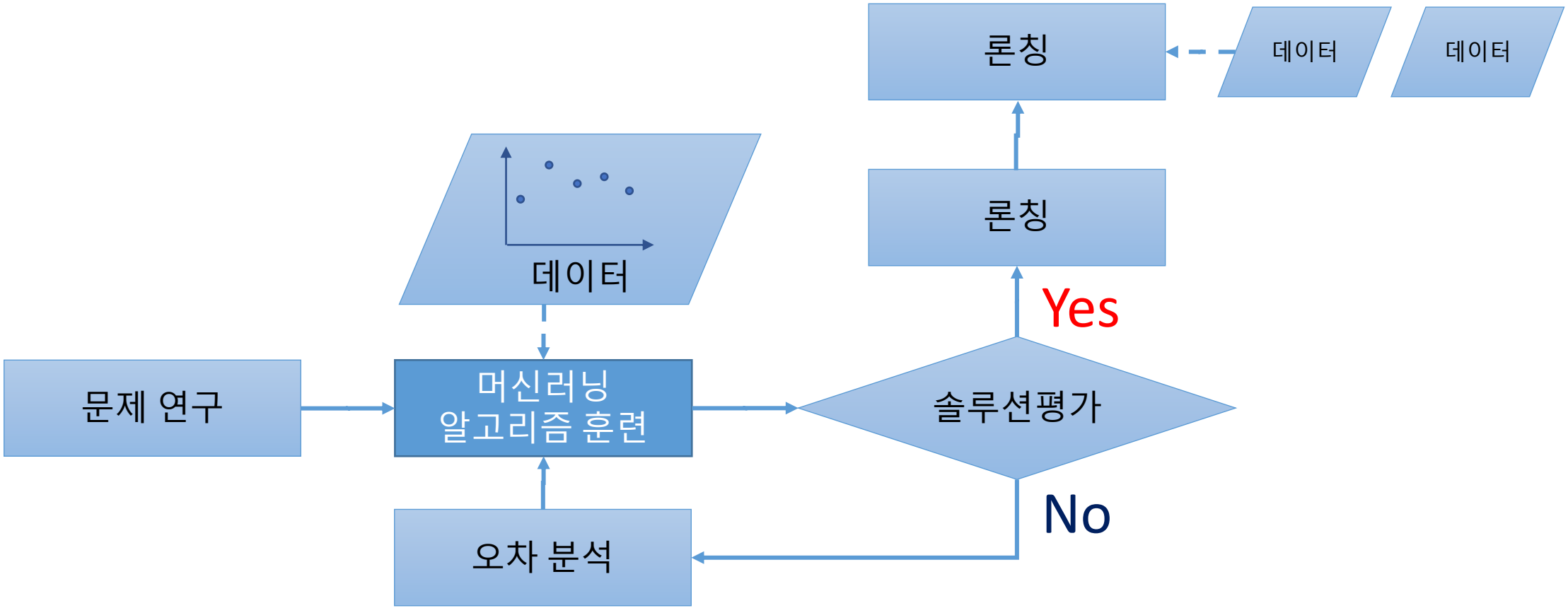
- 가용한 모든 데이터를 사용하여 모델을 훈련 시키는 방법
- 시간과 비용이 많이 소요되므로 오프라인 환경에서 훈련을 실행하고 학습이 종료된 후 시스템을 적용함
- 새로운 데이터 훈련을 위해서는 새로운 데이터 셋을 처음부터 다시 학습시키고 이전 모델과 교체 실행
- 머신 러닝 시스템의 자동화를 통해 주기적인 학습을 하는 형태로 적용하는 경우가 많음
- 전체 데이터셋을 매번 사용하기 때문에 전산 자원이 많이 필요하기 때문에 한정된 자원을 가진 시스템에서는 사용이 제한 될 수 있음



머신 러닝의 종류

온라인 학습 Online Learning

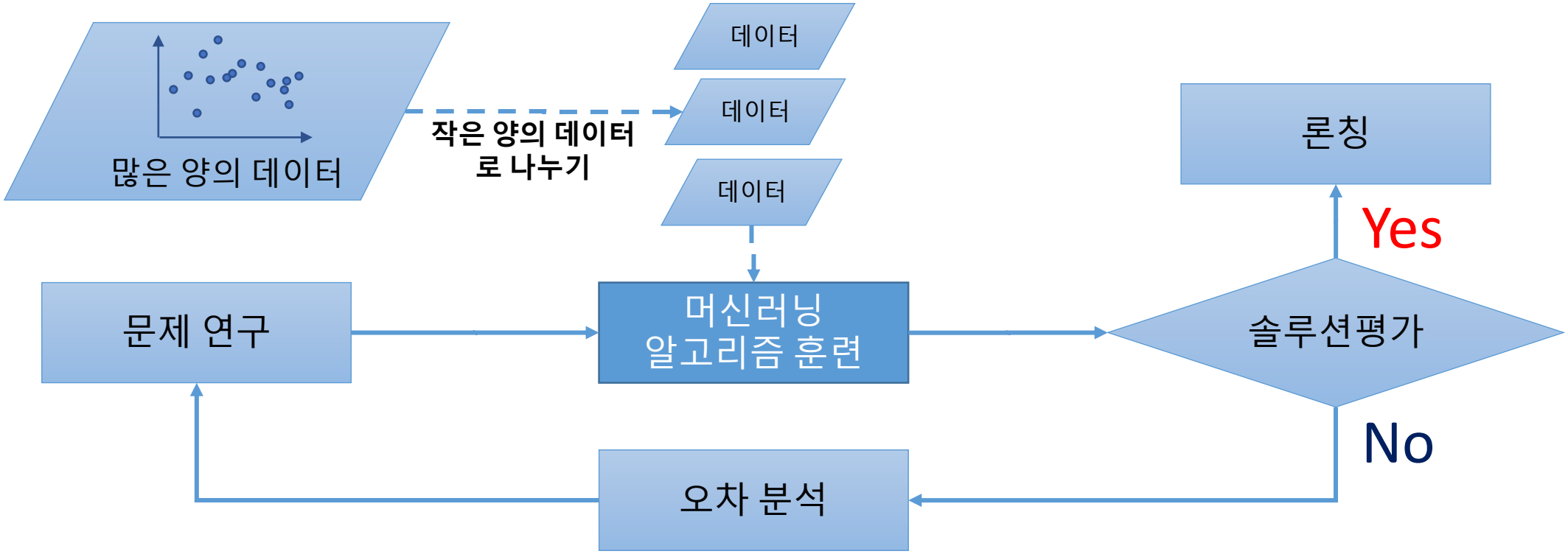
- 데이터를 순차적으로 한 개씩 혹은 미니배치(작은 단위의 데이터 묶음)단위로 사용하여 시스템을 훔련
- 데이터가 생성되는 즉시 시스템 학습에 적용이 가능
- 학습을 마친 데이터는 폐기하면 되기 때문에 전산 자원이 제한된 경우에도 사용이 가능



머신 러닝의 종류

온라인 학습 Online Learning

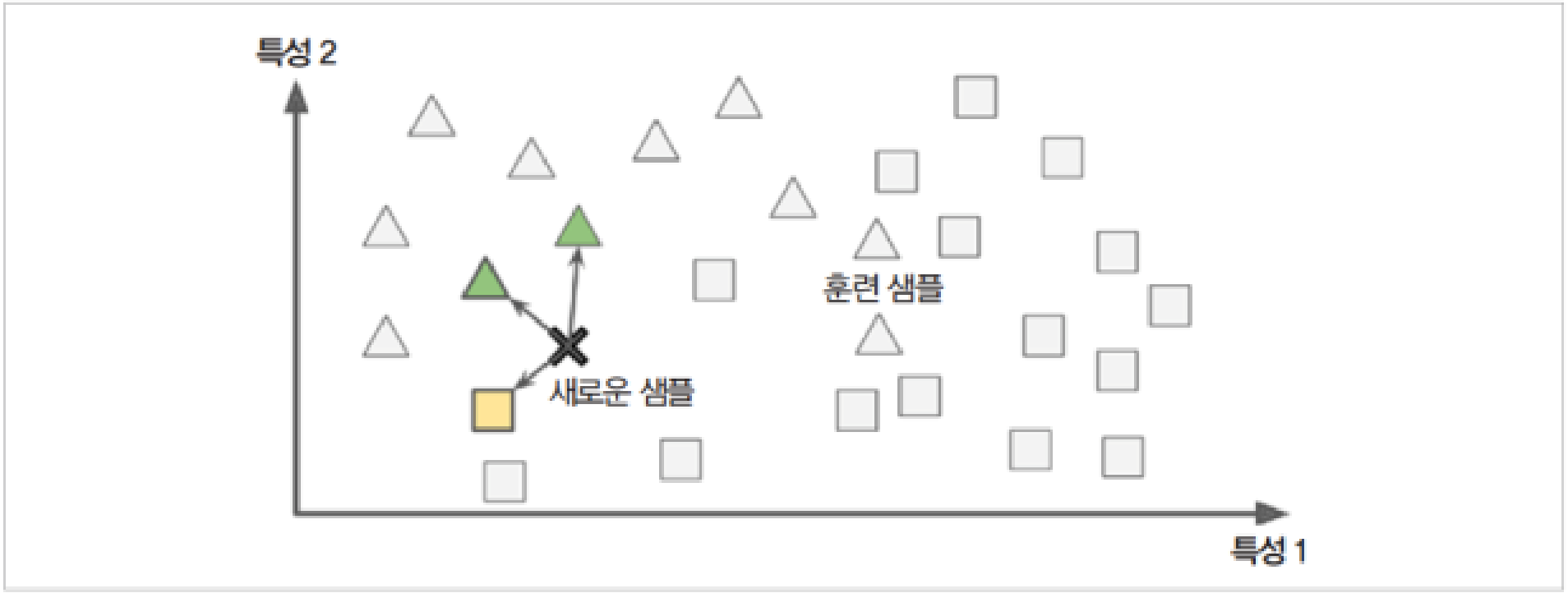
- 학습률: 변화하는 데이터를 얼마나 빠르게 적용시키는 지의 정도.
- 학습률이 높은 시스템은 최근의 데이터에 대해 민감하게 반응하며 과거 데이터를 무시하는 경향이 나타남
- 온라인 학습의 문제는 악성 데이터에 의해 모델이 잘못된 학습을 할 수 있는 점



머신 러닝의 종류

사례 기반 학습 instance-based Learning

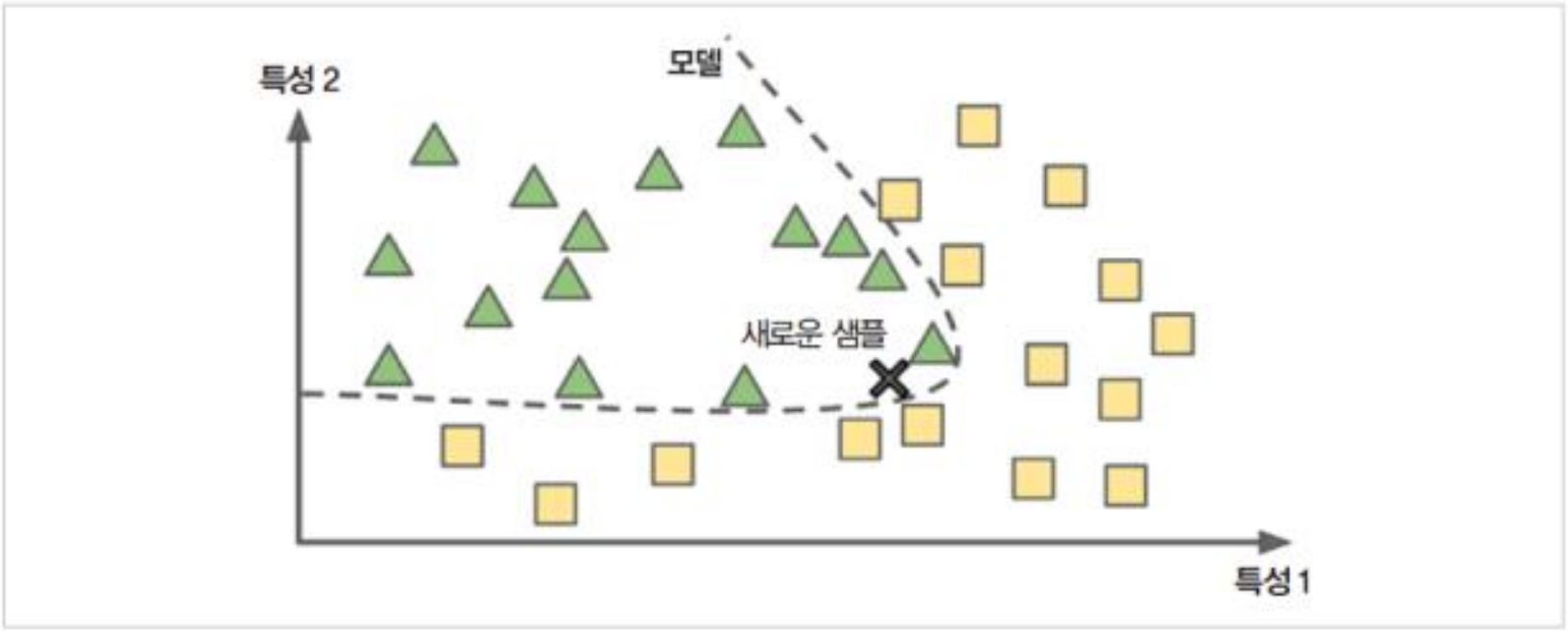
- 시스템이 훈련 샘플을 기억함으로써 학습하는 방식.
- 유사도 측정 Similarity Measurement : 시스템이 새로운 데이터와 학습한 데이터를 비교함으로써 일반화
- 스팸 메일 필터의 사례에서 새로운 메일을 스팸메일로 분류할 때 유사성의 기준으로 사용되는 것은 스팸메일과 새 메일에 공통으로 포함된 단어의 수



머신 러닝의 종류

모델 기반 학습 Model-based Learning

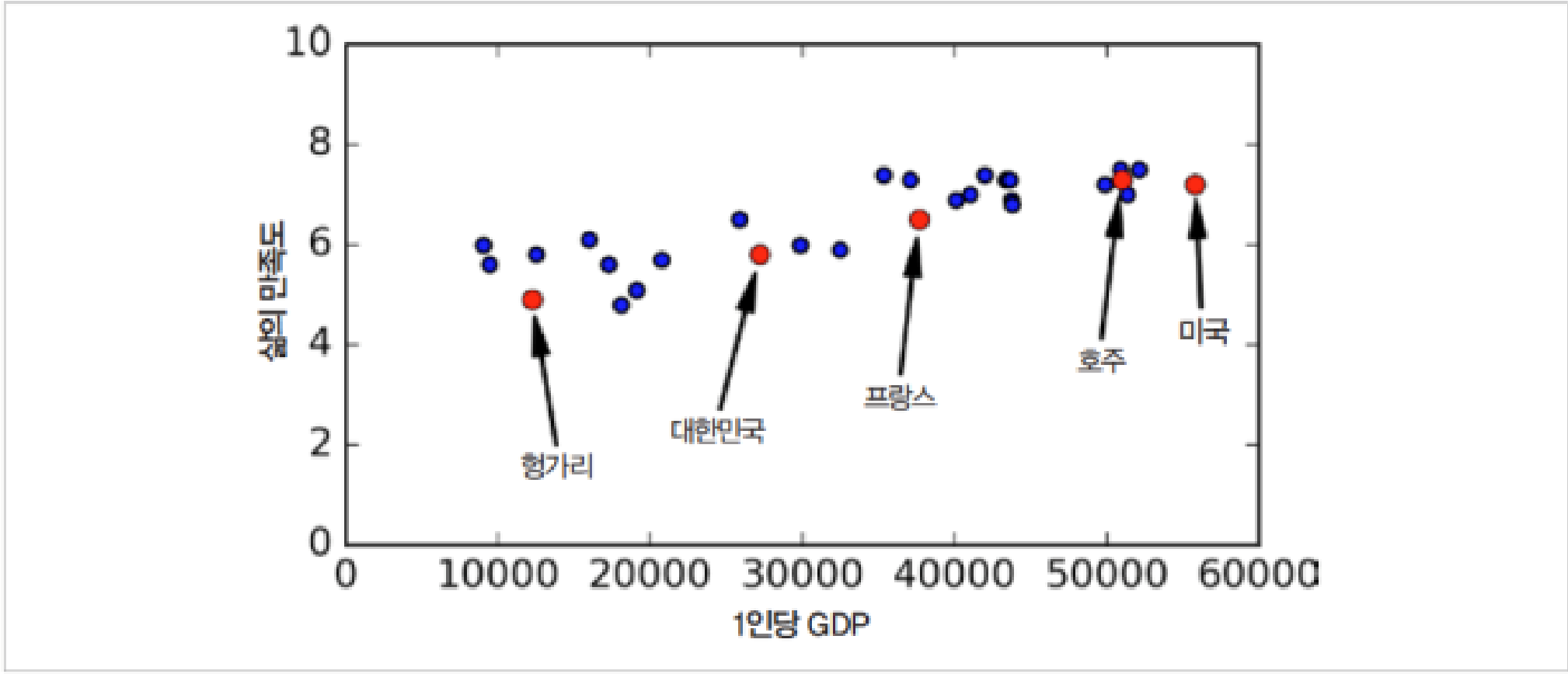
- 샘플들을 기반으로 알고리즘을 활용하여 모델을 만들어 예측에 활용하는 방법



머신 러닝의 종류

모델 기반 학습 Model-based Learning

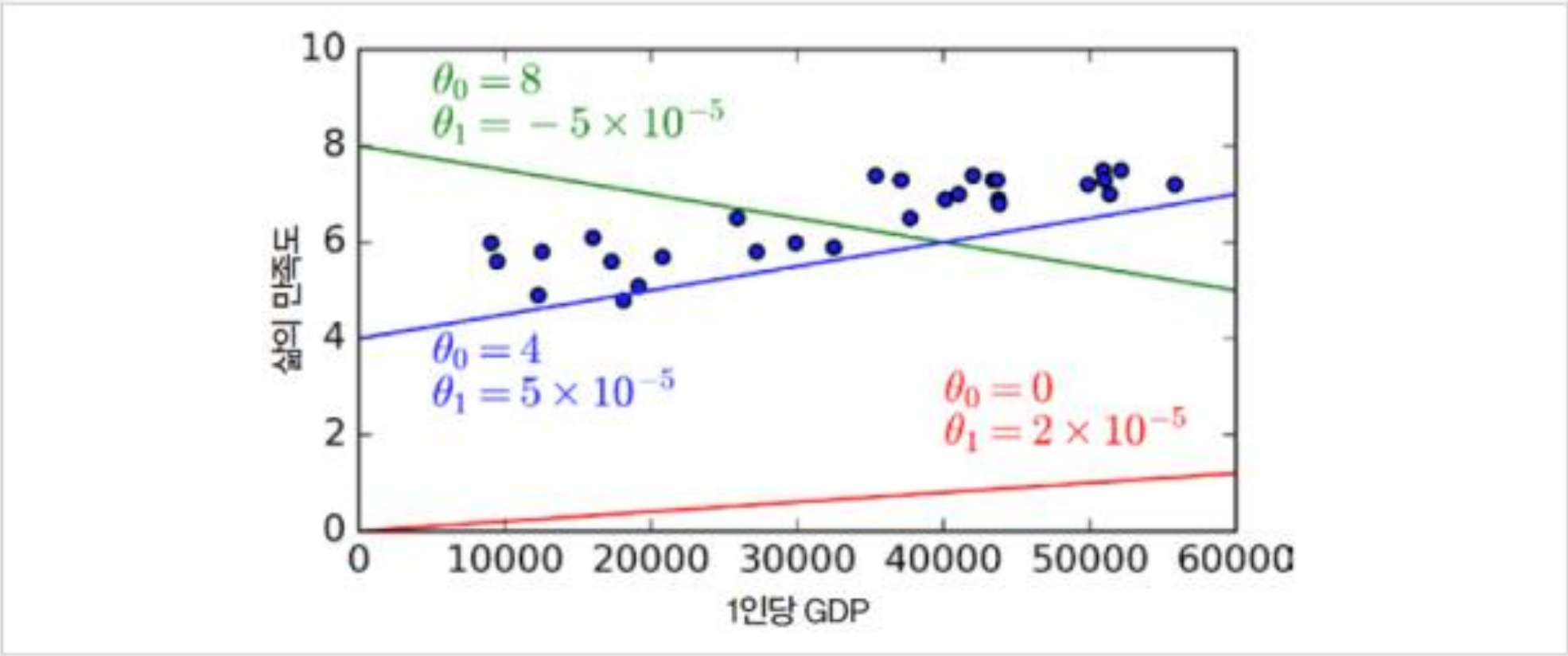
- 국민 소득과 삶의 만족도 사이의 상관 관계를 그림으로 표시



머신 러닝의 종류

모델 기반 학습 Model-based Learning

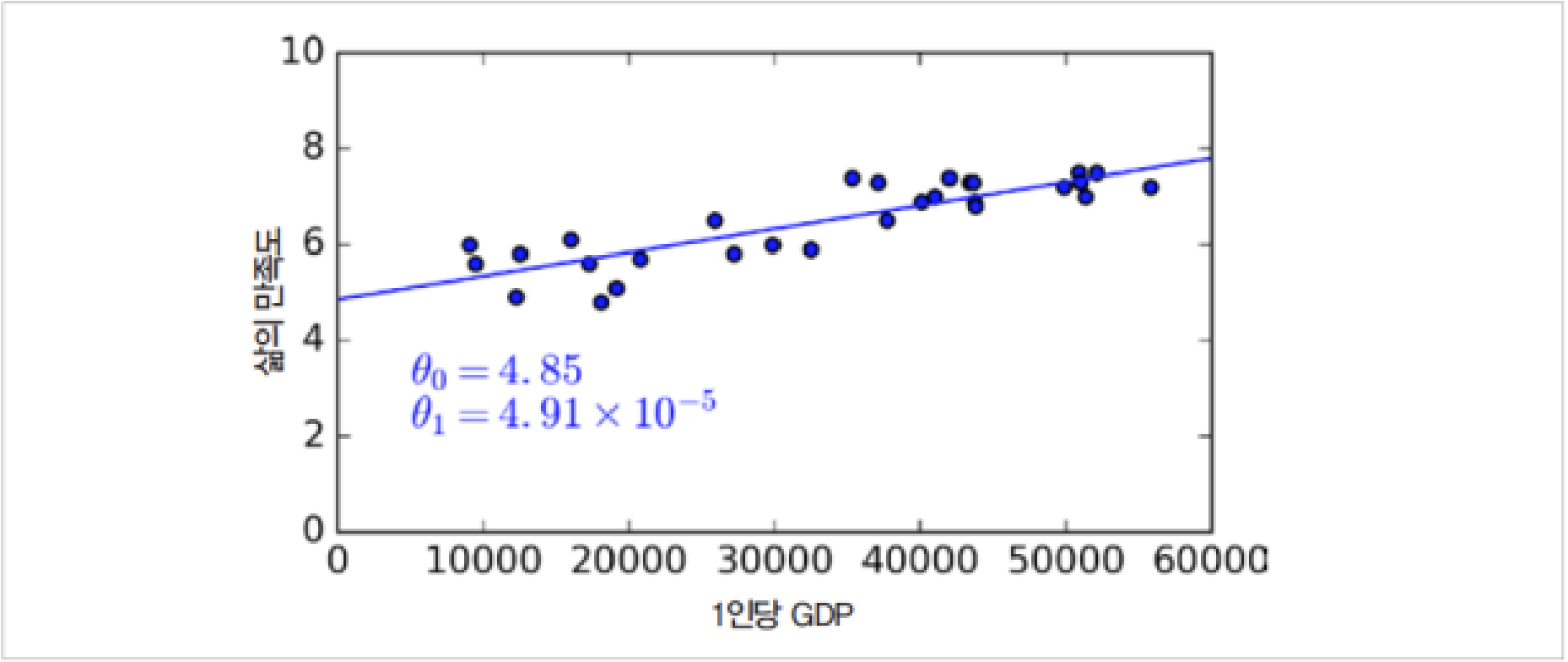
- 데이터가 분포한 모양에 따라 선형 모델을 구성함
- 삶의 만족도 = θ_0+θ_1 X 1인당 GDP
- 효용함수/적합도 함수 : 모델이 얼마나 현실 데이터에 잘 적용되는지 나타내 주는 함수
- 비용함수 Cost Function : 모델이 데이터에서 얼마나 떨어져 있는지 측정해주는 함수



머신 러닝의 종류

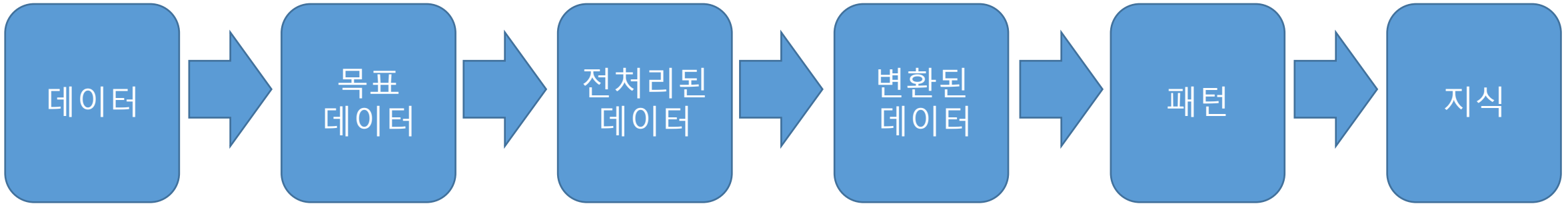
모델 기반 학습 Model-based Learning

- 알고리즘(선형회귀 모델)에 훈련 데이터(알고 있는 삶의 만족도와 1인당 GDP 데이터)를 공급하여 가장 잘 맞는 선형 모델의 파라미터(θ_0, θ_1 값)을 찾아내는 과정(훈련)
- 훈련 과정의 결과 값 : $\theta_0 = 4.85$, θ_1=4.91X 10^{-5}
- 새로운 데이터가 주어질 경우 이 모델에 대입하여 예상치를 예측할 수 있음



머신 러닝의 종류

데이터 분석 프로젝트는 어떤 순서로 수행되는가?



데이터 선택
Selection

데이터 전처리
Preprocessing

데이터 변환
Transformation

데이터 마이닝
Data Mining

결과해석 및 평가
interpretation/
Evaluation

삶의만족도 및 1인당 GDP 데이터 선택

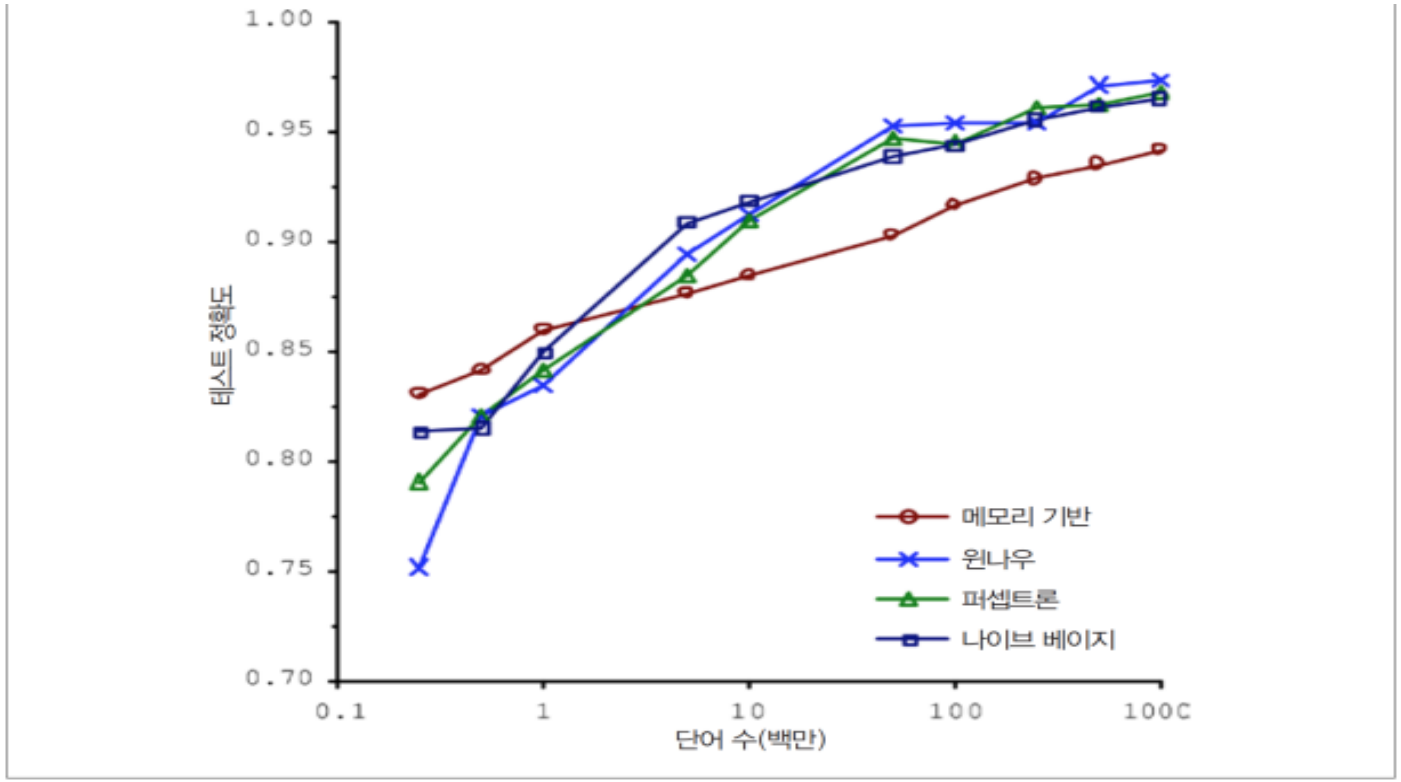
모델 선택
모델 훈련

새로운 데이터를 넣고 예측

머신 러닝의 주요 도전 과제

충분하지 않은 양의 훈련 데이터

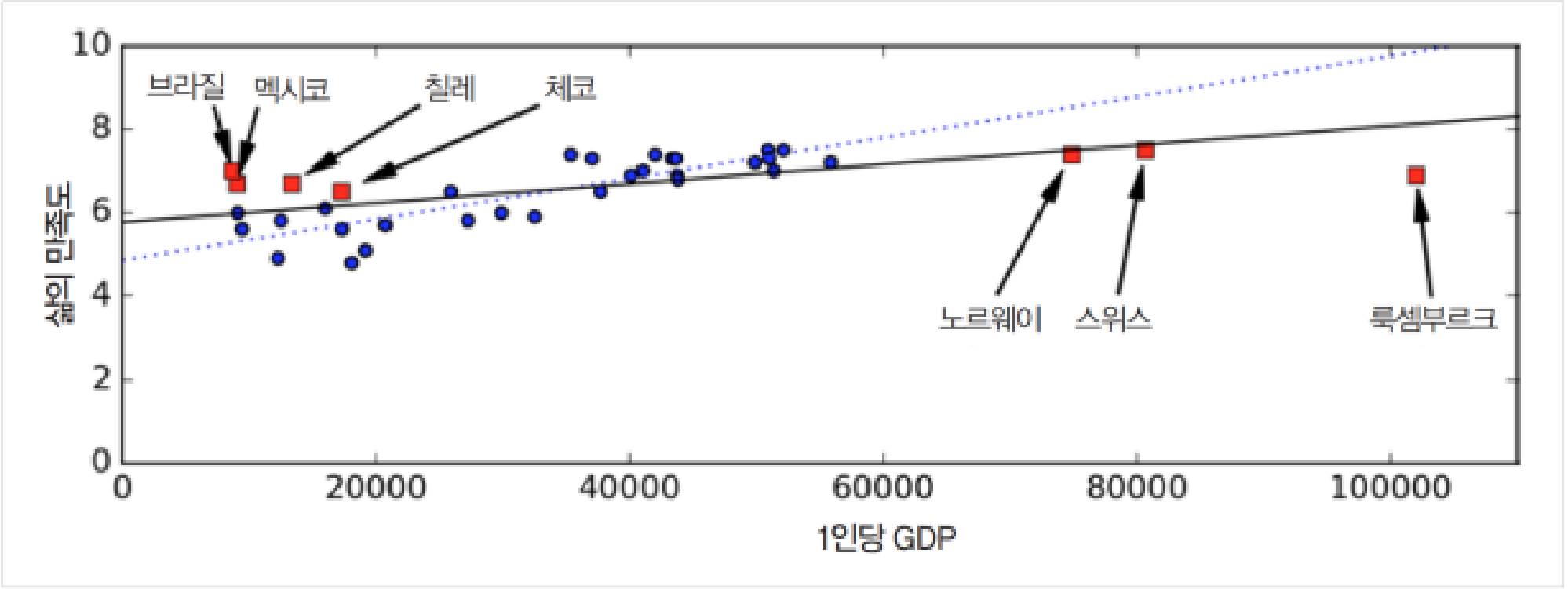
- 머신러닝의 학습에는 충분한 양의 데이터가 전제되어야 함
- 간단한 문제의 해결에도 수천개 이상의 데이터가 필요하며 음성인식과 같은 복잡한 문제의 해결을 위해서는 수백만개의 데이터가 필요한 경우도 있음
- 데이터의 개수가 충분할 경우 알고리즘의 종류에 관계없이 유효한 결과를 산출해내는 경우가 많음



머신 러닝의 주요 도전 과제

대표성 없는 훈련 데이터

- 모델이 현실에 적용 가능한 일반화가 되려면 훈련 데이터가 현실 데이터를 잘 대표하고 있어야 함
- 사례기반 학습과 모델 기반 학습에 공통적인 사항
- 데이터가 전체의 대표성을 충분히 갖지 못했을 경우 모델의 예측력이 크게 떨어지는 경우가 발생



머신 러닝의 주요 도전 과제

대표성 없는 훈련 데이터

- 샘플링 잡음: 우연에 의한 대표성 없는 데이터
- 샘플링 편향: 훈련 데이터의 추출 방법이 잘못되어 만들어진 현실성이 없는 훈련 데이터
- 1936년 미국 대통령 선거(루즈벨트 vs, 랜던)
  - The Literary Digest 잡지의 사전 여론 조사
  - 우편을 통해 1000만명을 대상으로 한 대규모 여론조사
  - 240만명의 응답자 중 랜던의 득표확률이 57%로 예측
  - 실제로는 62% 득표로 루즈벨트 당선
- 샘플링의 문제
  - 여론조사의 대상자 선정을 위해 전화번호부, 자사 고객 명부, 유명 클럽 회원 명부 등을 사용
  - 정치에 관심 없는 사람, 해당 잡지를 싫어하는 사람 등 76%의 사람들을 제외한 일부만이 해당 설문에 응답
  - 결과적으로 부유한 사람, 정치에 관심있는 사람들만의 데이터가 샘플 데이터로 활용됨

머신 러닝의 주요 도전 과제

낮은 품질의 데이터

- 훈련 데이터에 일정 비율 이상의 이상치, 잡음 등이 포함되어있을 경우 시스템이 데이터에 내재된 패턴을 잘 찾아내지 못하게 됨
- 훈련 데이터를 잘 정제하는 것이 필요
- 일부 샘플이 명확하게 이상치인 경우 이상치 데이터를 무시하거나 수동으로 조정하는 과정을 거쳐서 훈련 데이터로 사용하는 것이 효과적
- 일부 데이터가 결측치(Missing Value)인 경우 이 특성을 무시하고 훈련 데이터로 사용할 것인지 빠진 값을 특정한 값(평균, 중간값 등) 으로 변경하여 사용할 것인지 결정해야 함

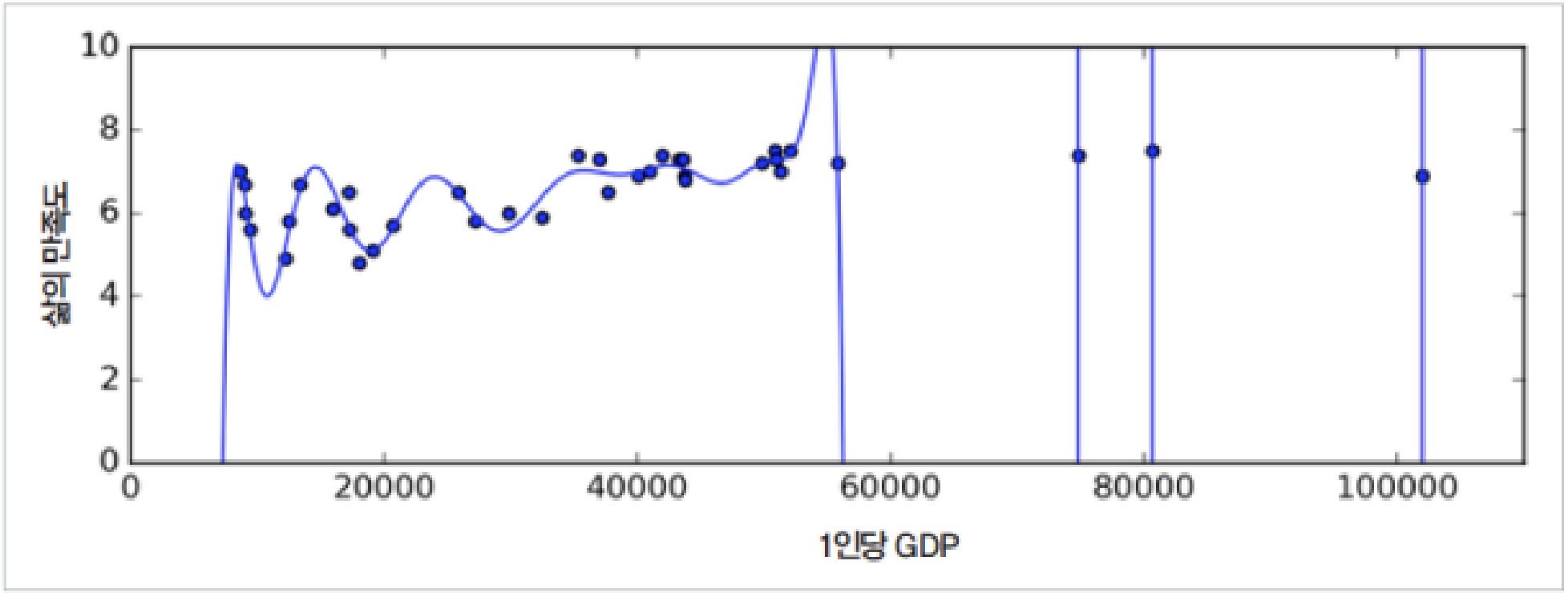
관련 없는 특성

- 훈련 데이터에 관련 없는 특성이 많이 포함되어 있을 경우 모델이 정확한 예측을 해낼 수 없음
- 삶의 만족도를 예측하는 모델에 나라이름 특성을 사용하는 경우
- 특성 선택 : 데이터가 가지고 있는 특성 중 훈련에 가장 도움이 되는 특성을 선택
- 특성 추출: 상관관계가 있는 특성들을 서로 결합하여 더 유용한 특성을 만들어냄-->차원 축소에 활용

머신 러닝의 주요 도전 과제

훈련 데이터 과대 적합

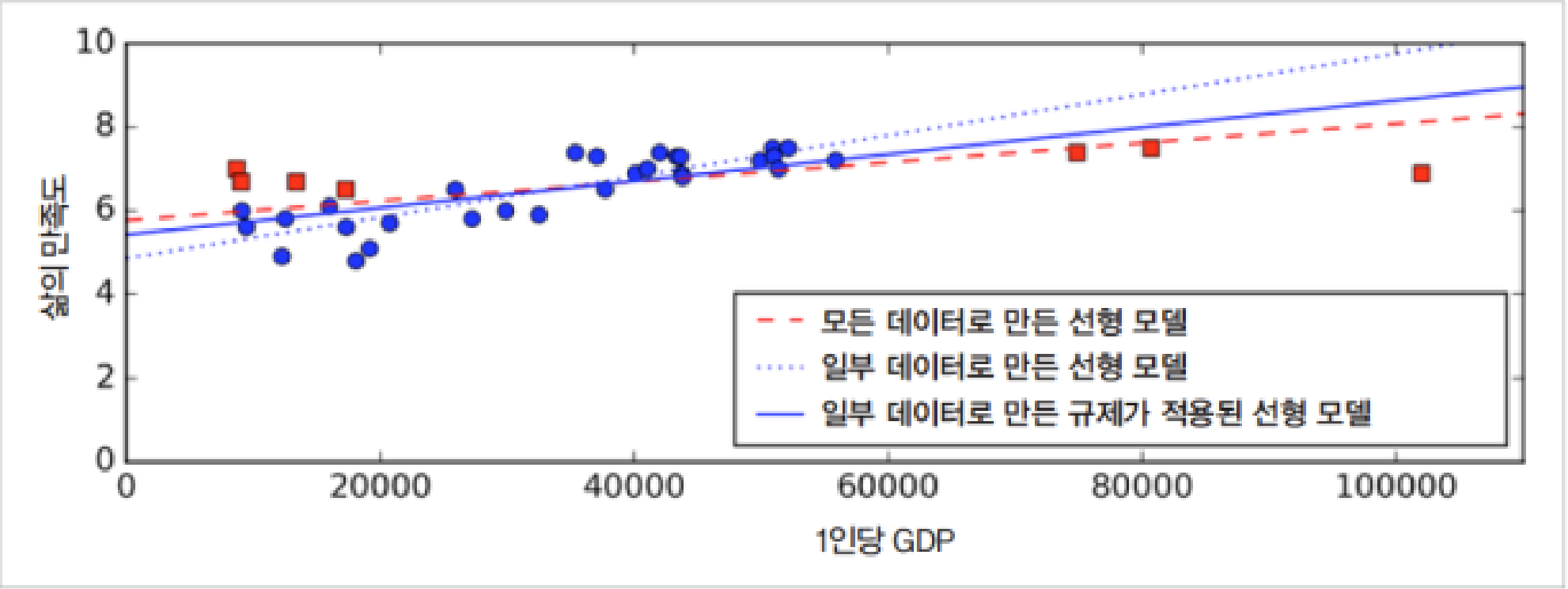
- 과대적합 : 성급한 일반화의 오류. 현실 데이터의 일부인 훈련 데이터에만 모델을 특화시키면 현실성이 없는 모델이 만들어짐
- 심층 신경망과 같은 복잡한 모델은 단숭한 선형회귀에 비해 훈련 데이터에서 시부적인 패턴을 더 잘 발견해 낼 수 있지만 데이터셋이 너무 작거나 이상한 데이터가 포함되어있을 경우 현실과 동떨어진 모델을 만들어 낼 수 있음



머신 러닝의 주요 도전 과제

훈련 데이터 과대 적합-규제

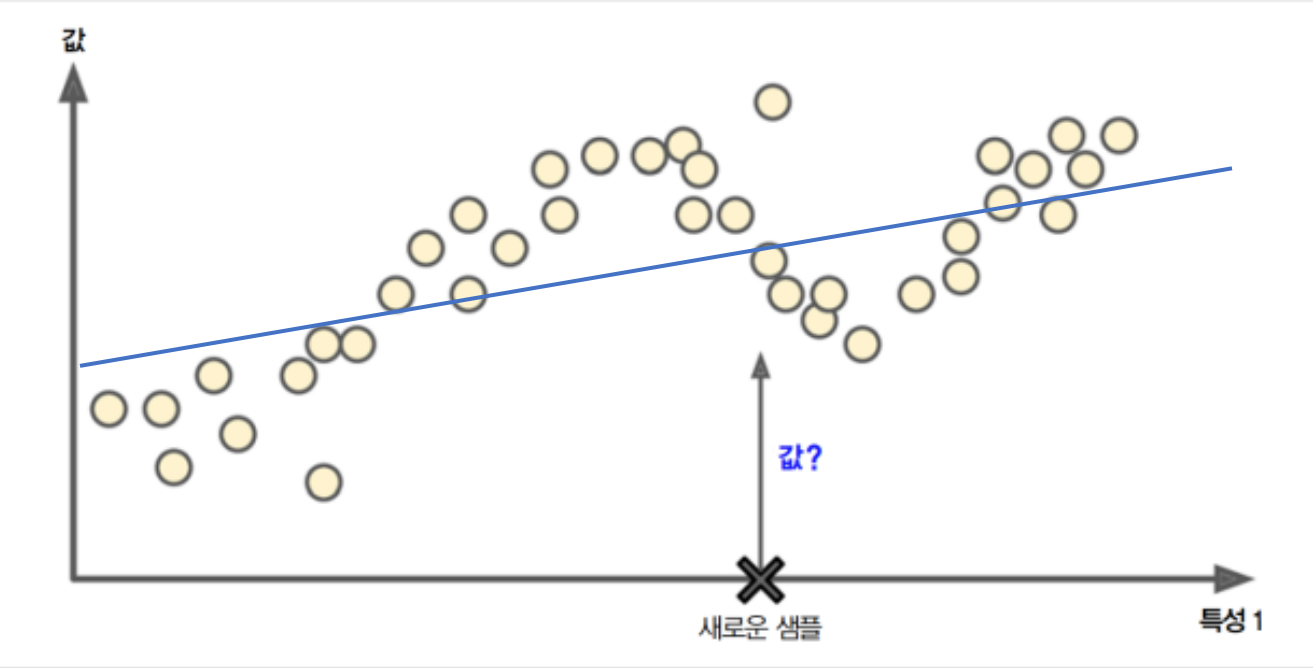
- 규제: 모델을 일반화하고 과대적합을 피하기 위해 모델에 제약을 가하는 것
- 모델의 파라미터(선형회귀에 있어서의 절편과 기울기)에 제약을 가하면 훈련 데이터에 대한 설명력을 떨어지지만 현실 데이터에 더 잘 맞는 모델을 만들 수 있음
- 모델의 규제를 위해 부여하는 값을 하이퍼파라미터라고 함-모델과 관련 없는 모델 형성에 필요한 숫자



머신 러닝의 주요 도전 과제

훈련 데이터 과소 적합

- 적용한 모델이 너무 단순해서 훈련 데이터 뿐 아니라 현실에도 적용이 어려운 경우
- 해결 방안
  - 파라미터가 더 많은(더 많은 특성을 사용하는) 모델을 적용
  - 학습 알고리즘을 위해 더 적합한 특성을 사용(특성공학)
  - 모델에 적용된 규제를 줄임



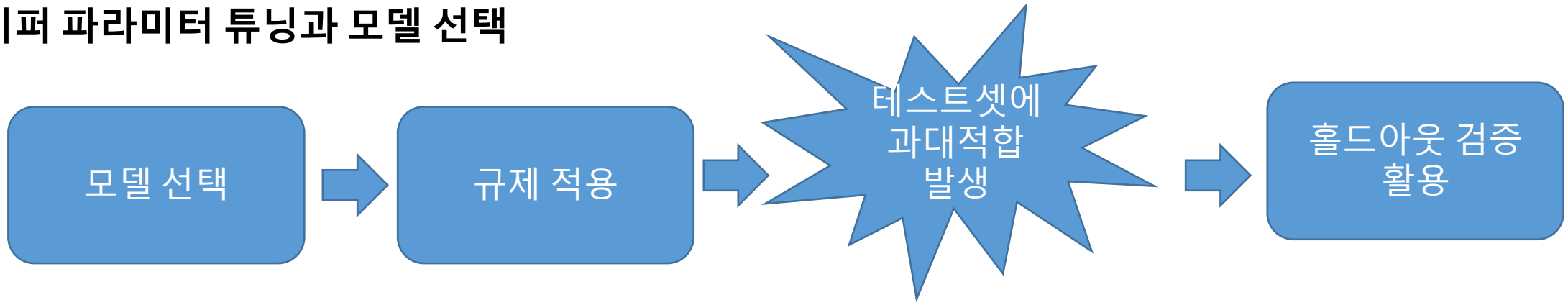
테스트와 검증

데이터 분할-훈련 세트와 테스트 세트

- 모델이 얼마나 현실에 잘 맞는지 보려면 실제 데이터로 적용하기 전에 테스트를 해보는 것이 필요
- 훈련 세트 : 시스템을 훈련하기 위해 사용하는 데이터의 집합
- 테스트 세트: 모델이 잘 만들어졌는지 사전에 테스트하기 위해 따로 준비해 둔 데이터
- 일반화 오차 : 새로운 데이터를 적용했을 때 모델이 얻은 값의 오류비율. 테스트 세트를 통해 오차에 대한 추정치를 얻을 수 있음
- 훈련 세트에 적합한 모델을 만들었지만 일반화 오차가 높다면 이는 모델이 훈련 데이터에 과대적합
- 과대적합을 피하기 위해 지나치게 단순화된 모델을 사용할 경우 과소적합의 우려가 있음

테스트와 검증

하이퍼 파라미터 튜닝과 모델 선택



훈련 데이터에 여러 개의 알고리즘을 훈련해보고 적합한 알고리즘을 선택

일반화 오류를 피하기 위해 테스트셋을 통해적합한 하이퍼 파라미터를 선택하여 적합한 규제를 적용

하이퍼 파라미터 튜닝에 테스트셋을 활용했기 때문에 테스트셋에 과대적합 발생

훈련 셋을 분할하여 검증세트를 만들고 다양한 파라메터중 적합한 하이퍼 파라메터 선택

여러 개의 작은검증 세트를 사용하여 모델 및 하이퍼 파라미터를 검증하고 평균하여 모델 평가

테스트와 검증

데이터 불일치 문제

- 훈련 데이터셋과 현실 데이터 셋 사이의 환경적 차이가 존재할 경우 모델의 문제인지 데이터의 기본적인 불일치가 원인지 불분명
- 검증 세트와 테스트 세트가 현실의 데이터를 가장 잘 대표해야 함
- 현실을 대표하는 데이터셋을 검증세트와 데이터셋에 나누어 넣고 검증과 테스트 실행
- 훈련 세트를 나누어 훈련-개발 세트를 만들고 훈련한 모델을 훈련-개발 세트에서 평가→모델이 잘 작동할 경우 과대적합문제가 아니라 데이터 불일치 문제
- 데이터 불일치의 경우 데이터 자체에 문제가 있을 수 있으므로 훈련 데이터를 전처리하여 현실 데이터와 유사하게 처린 한 후 모델 재 훈련

공짜 점심 없음 이론

- 모델은 현실에서 나타는 현상을 단순화하여 예측 가능하게 만든 일종의 모형
- 모델에는 일정한 가정(데이터가 직선을 따라 배열되어있다)을 해야 함
- 어떤 데이터에 어떤 모델이 적합할 지는 사전에 알 수 있는 방법이 없으므로 데이터에 대해 타당한 가정을 세우고 적용해 보는 방법 외에는 적합한 모델을 결정할 수 있는 방법이 없다.(돈 내고 먹어보기 전에는 어떤 음식이 괜찮은지 알 수가 없다)