

AI Security

01. Introduction to Machine Learning

Artificial Intelligence Research (AIR) LAB

<https://air.korea.ac.kr/>

School of Cybersecurity

Korea University

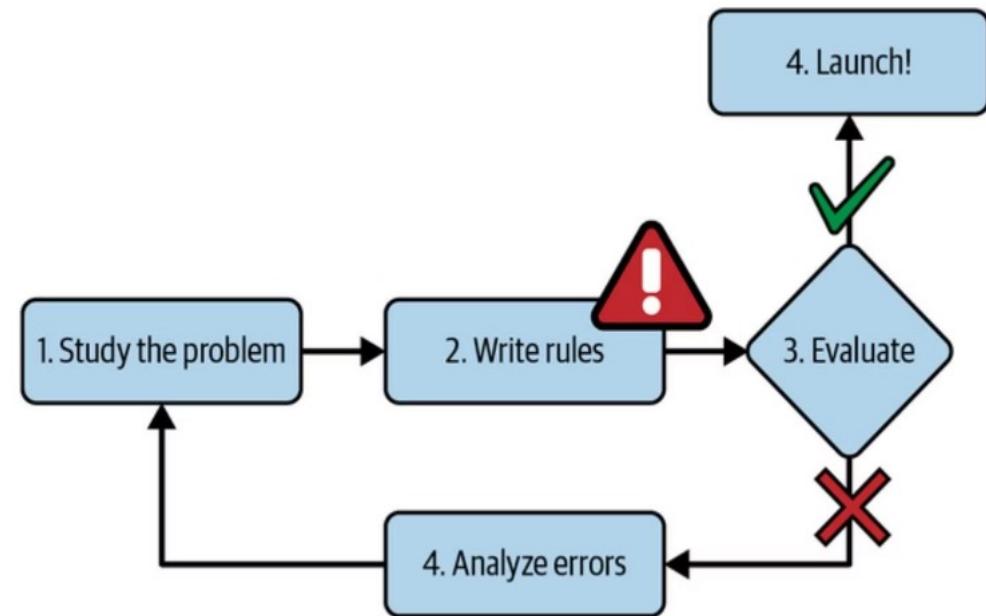
Sangkyun Lee (이상근)

2025. Spring Semester

기계 학습 개요

얼마 전까지만 해도 휴대폰에게 집으로 가는 길을 물어보면 무시당했을 것임. 하지만 이제 기계 학습은 공상 과학이 아닌 현실이 됨. 수십억 명이 매일 사용하고 있으며, 사실 광학 문자 인식(OCR)과 같은 특수 응용 프로그램에서는 수십 년 동안 사용되어 옴.

1990년대에 처음으로 주류가 된 기계 학습 응용 프로그램은 스팸 필터였음. 자의식 있는 로봇은 아니지만, 기술적으로는 기계 학습으로 분류됨. 이후 음성 프롬프트, 자동 번역, 이미지 검색, 제품 추천 등 수백 개의 기계 학습 응용 프로그램이 등장.



기계 학습이란 무엇인가?

기계 학습의 정의

기계 학습은 컴퓨터가 명시적으로 프로그래밍되지 않고 데이터로부터 학습할 수 있도록 하는 과학(및 예술).

아서 사무엘의 정의 (1959)

명시적으로 프로그래밍하지 않고도 컴퓨터에게 학습 능력을 부여하는 연구 분야.

톰 미첼의 정의 (1997)

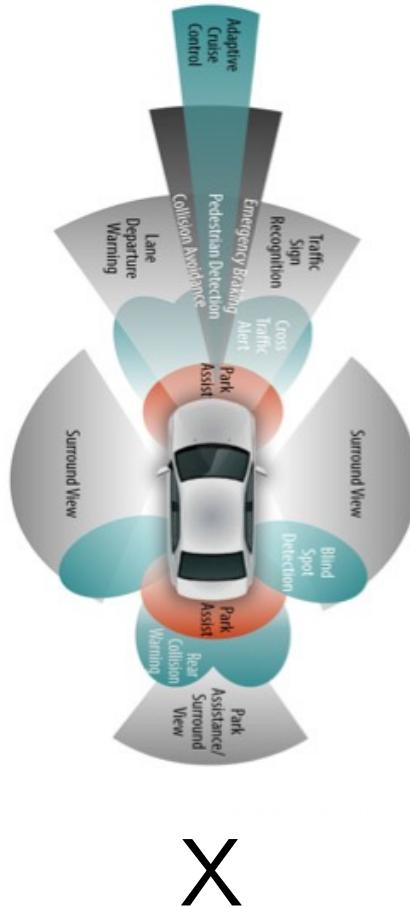
어떤 작업 T와 성능 측정 P에 대해, 경험 E로부터 학습하여 P로 측정된 T의 성능이 향상되는 컴퓨터 프로그램.



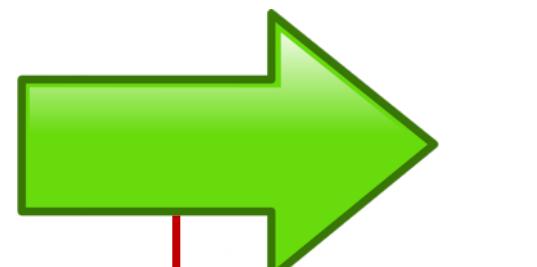
스팸 필터는 사용자가 표시한 스팸 이메일과 일반 이메일(햄)의 예를 통해 학습하는 기계 학습 프로그램. 시스템이 학습에 사용하는 예제를 훈련 세트라고 하며, 각 훈련 예제는 훈련 인스턴스(또는 샘플)라고 함. 학습하고 예측하는 기계 학습 시스템의 일부를 모델임.

기계학습 Machine Learning

입력 데이터



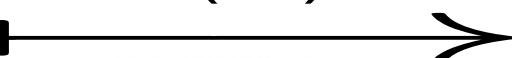
출력 데이터



w

parameters
(weights)

$$f_{\vec{w}}(X)$$



Y

기계 학습을 사용하는 이유



전통적 접근 방식의 한계

스팸 필터를 전통적인 프로그래밍 기법으로 작성하면 복잡한 규칙의 긴 목록이 되어 유지 관리가 어려움.



기계 학습 접근 방식의 장점

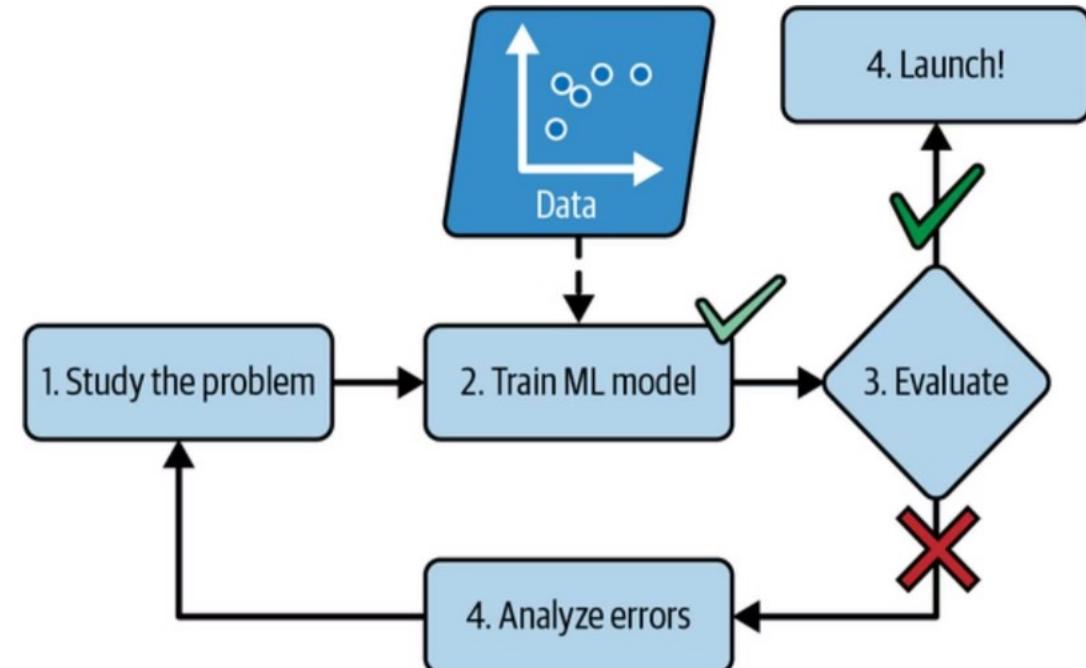
기계 학습 기반 스팸 필터는 햄 예제와 비교하여 스팸 예제에서 비정상적으로 자주 발생하는 단어 패턴을 자동으로 학습함

- **변화에 자동 적응**

스파머가 "4U"를 "For U"로 바꾸면, 기계 학습 기반 필터는 사용자의 개입 없이 이러한 변화를 자동으로 감지하고 적응함

- **인간의 학습 지원**

기계 학습 모델은 검사하여 학습한 내용을 확인할 수 있으며, 이를 통해 문제에 대한 더 나은 이해를 얻을 수 있음.



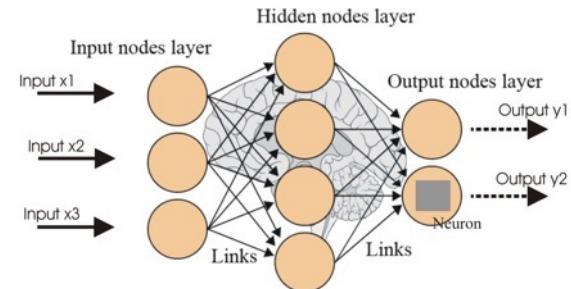
Why DL?

Increase of available data



DATA

Advances in AI and optimization techniques



AI + Optimization

Low-cost, high-perf
computers



HPC (High Performance Computing)

기계 학습 응용 사례



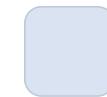
이미지 분석

생산 라인의 제품 이미지를 자동으로 분류하거나
뇌 스캔에서 종양을 감지하는데 CNN이나 트랜스포머를 사용함.



예측 및 추천

회사 수익 예측, 신용카드 사기 감지, 고객 세분화,
제품 추천 등 다양한 모델을 사용함.



자연어 처리

뉴스 기사 자동 분류, 공격적인 댓글 플래깅, 문서 요약, 챗봇 등에 RNN, CNN 또는 트랜스포머를 활용함.

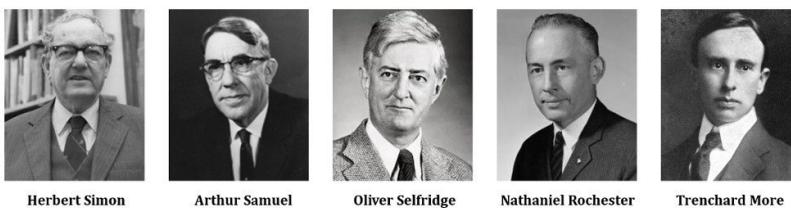
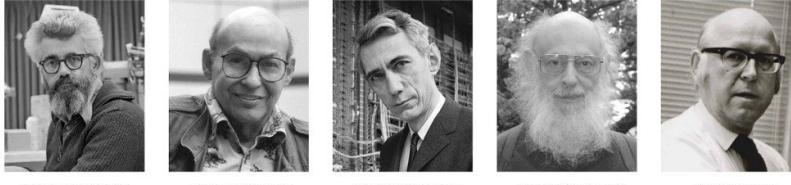


강화 학습

게임 봇과 같이 에이전트가 시간이 지남에 따라 보상을 최대화하는 행동을 선택하도록 훈련함.

Timeline of AI

1956 Dartmouth Conference: The Founding Fathers of AI



"AI" is coined!



1st wave:
Rule-based

GPU for DL
AI winter

"AGI" is coined!



2nd wave: Start
Big Data & ML,
Statistical



2nd wave: End
Generative AI,
Statistical

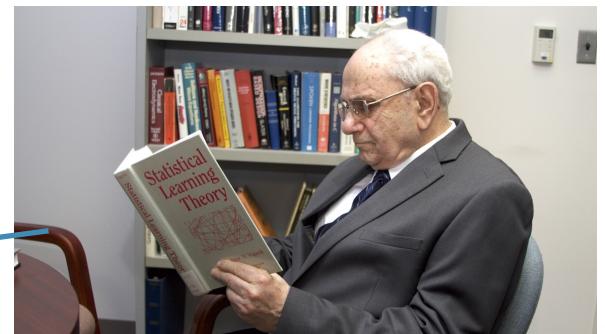


DALL-E(2)
AlphaFold
GPT/BERT
GPT-3
Transformers
AlphaGo

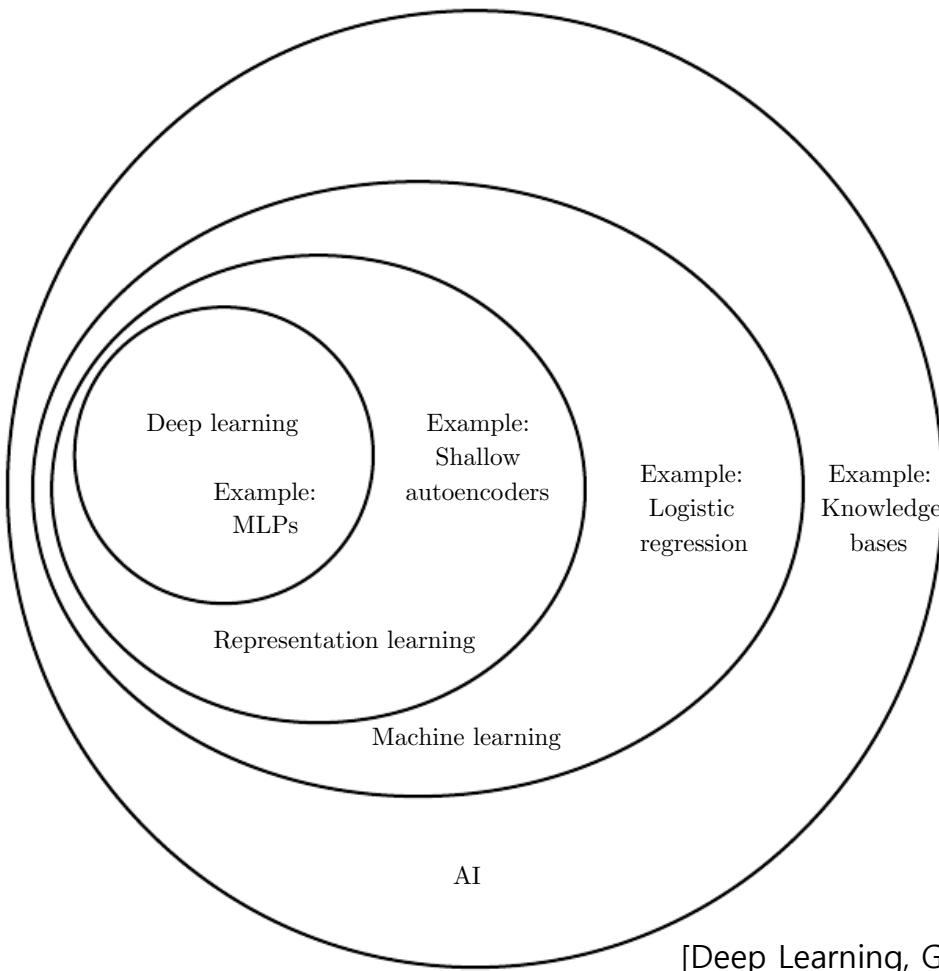


3rd wave:
Cognitive AI
to AGI

Vladimir Vapnik (1936~)



AI 인공지능



인공지능 AI

: Expert system, Cybernetics

기계학습 Machine Learning

: SVM, Logistic regression, decision trees

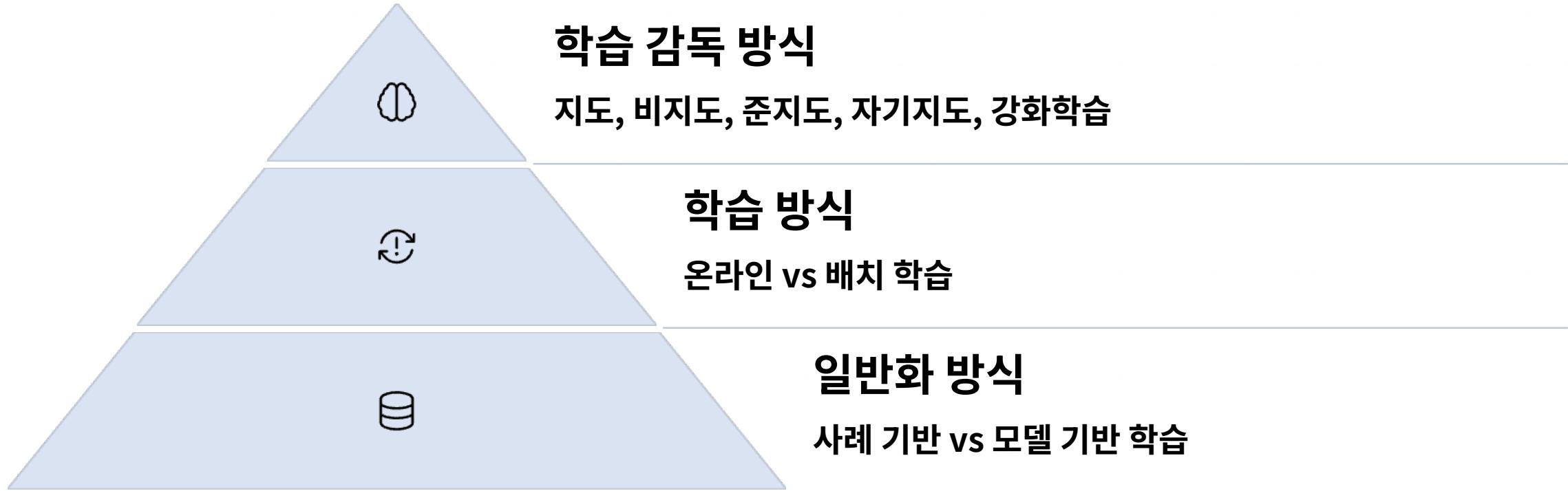
표현형 학습 Representation Learning

: Autoencoder

딥러닝 Deep Learning

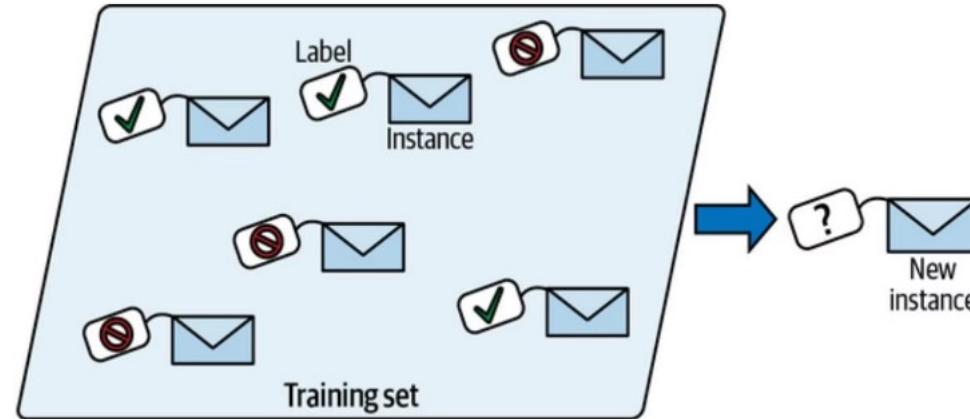
: NLP, computer vision, ...

기계 학습 시스템의 유형



이러한 기준은 배타적이지 않으며 원하는 방식으로 조합할 수 있음. 예를 들어, 최첨단 스팸 필터는 사람이 제공한 스팸과 햄 예제를 사용하여 훈련된 심층 신경망 모델을 사용하여 실시간으로 학습할 수 있음. 이는 온라인, 모델 기반, 지도 학습 시스템임.

훈련 감독 방식



지도 학습

지도 학습에서는 알고리즘에 공급하는 훈련 세트에 레이블이라고 하는 원하는 솔루션이 포함됨. 스팸 필터는 이에 대한 좋은 예임. 클래스(스팸 또는 햄)와 함께 많은 예제 이메일로 훈련되어 새 이메일을 분류하는 방법을 학습함.

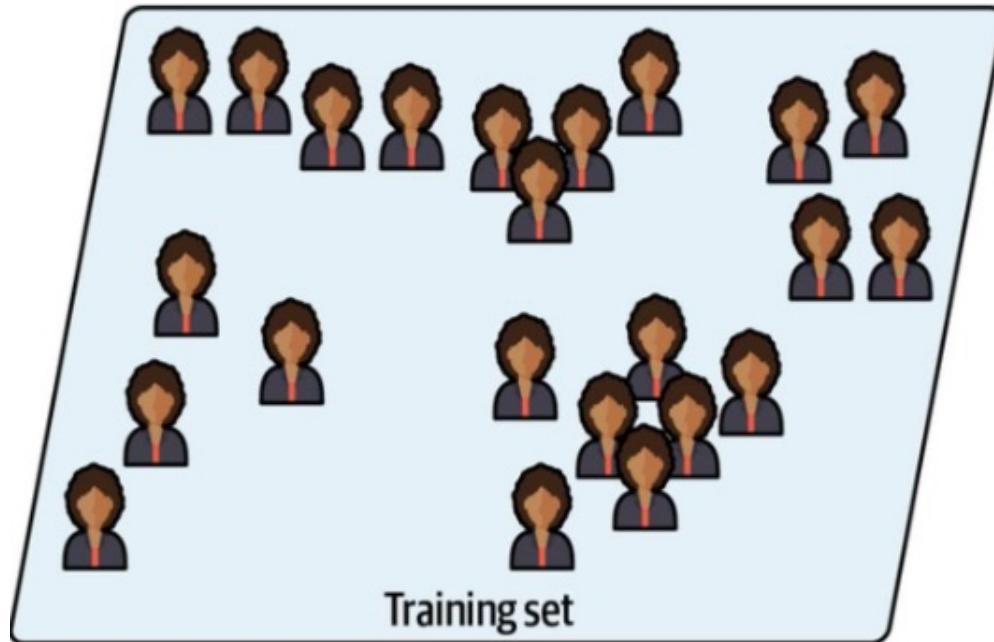
회귀와 분류

또 다른 일반적인 작업은 자동차 가격과 같은 대상 숫자 값을 예측하는 것임. 이러한 작업을 회귀라고 함. 시스템을 훈련하려면 특성과 대상(즉, 가격)을 모두 포함하는 많은 자동차 예제를 제공해야 함.

타겟과 레이블은 일반적으로 지도 학습에서 동의어로 취급되지만, 타겟은 회귀 작업에서 더 일반적이고 레이블은 분류 작업에서 더 일반적임.

훈련 감독 방식

비지도 학습



클러스터링

블로그 방문자에 대한 많은 데이터가 있다고 가정해 보자. 클러스터링 알고리즘을 실행하여 유사한 방문자 그룹을 감지할 수 있음.

시각화

시각화 알고리즘은 복잡하고 레이블이 없는 데이터를 공급하면 쉽게 그릴 수 있는 2D 또는 3D 표현을 출력함.

차원 축소

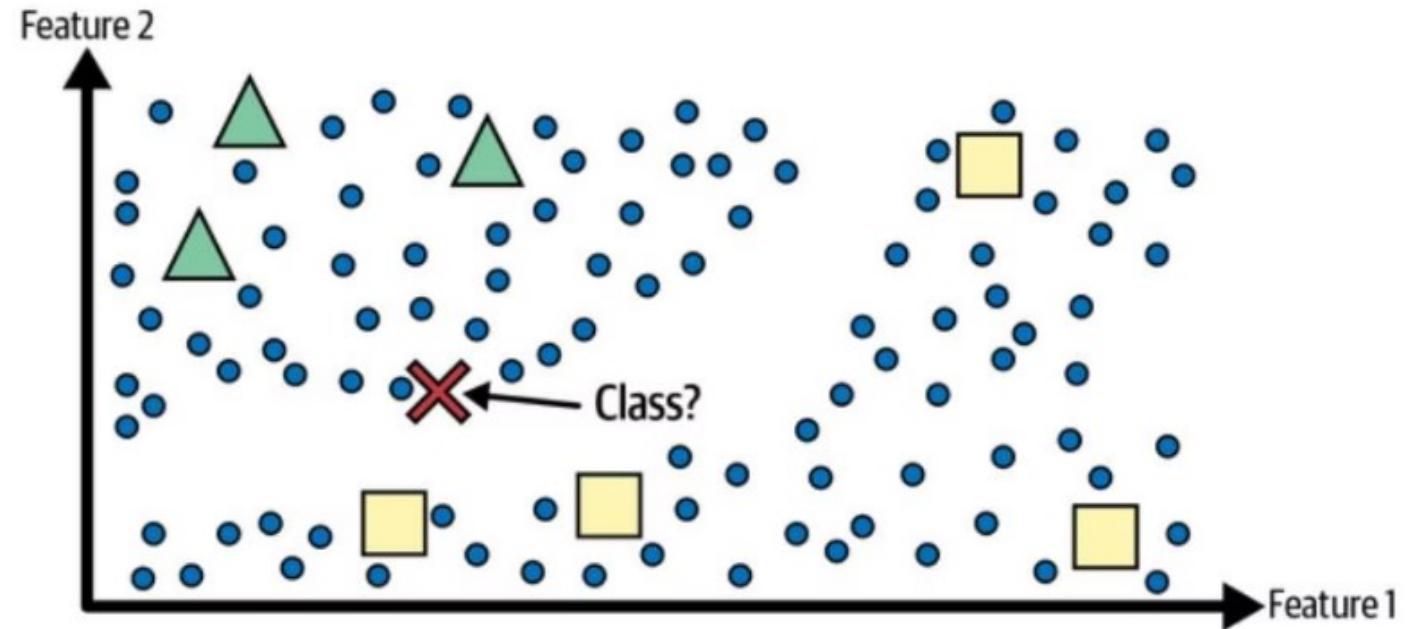
너무 많은 정보를 잃지 않고 데이터를 단순화하는 것이 목표임.
상관관계가 있는 여러 특성을 하나로 병합할 수 있음.

이상 탐지

사기 방지를 위한 비정상적인 신용 카드 거래 감지, 제조 결함 포착 또는 데이터셋에서 이상치 자동 제거에 사용됨.

훈련 감독 방식

준지도 학습



레이블링의 어려움

데이터 레이블링은 시간이 많이 걸리고 비용이 많이 들어 레이블이 없는 인스턴스는 많지만 레이블이 있는 인스턴스는 적은 경우가 많음.

실제 응용 사례

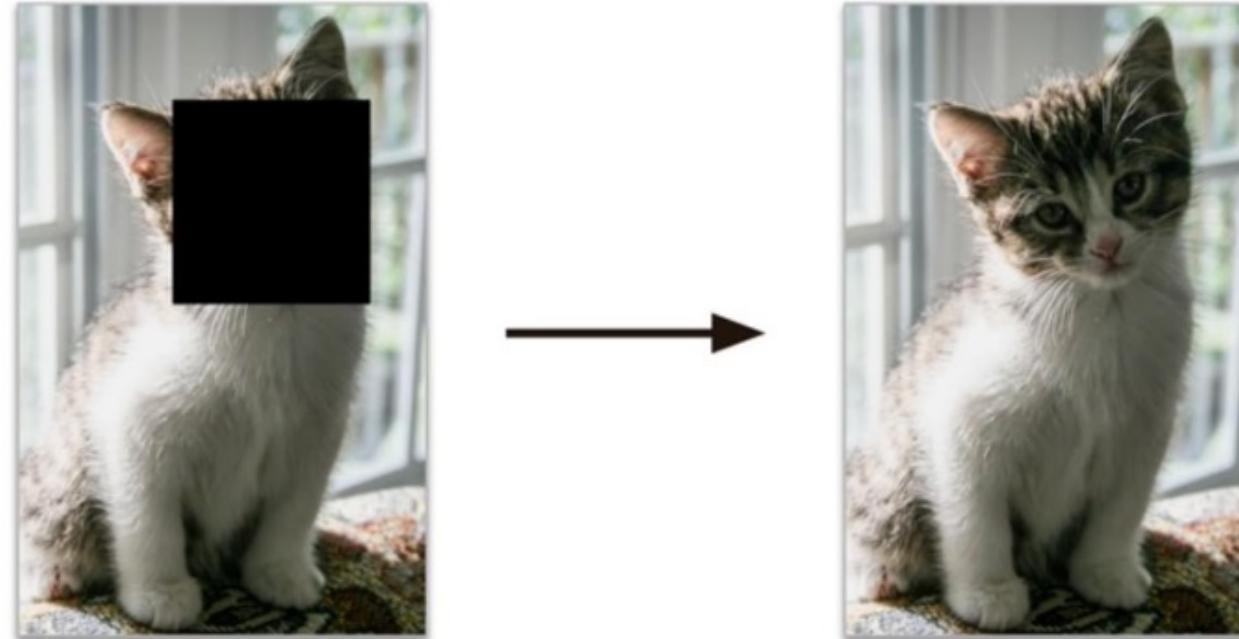
Google 포토와 같은 사진 호스팅 서비스가 좋은 예임. 사진을 업로드하면 동일한 사람이 여러 사진에 나타나는 것을 자동으로 인식함.

알고리즘 조합

대부분의 준지도 학습 알고리즘은 비지도 및지도 알고리즘의 조합임. 클러스터링 알고리즘을 사용해 유사한 인스턴스를 그룹화한 다음 레이블이 지정되지 않은 인스턴스에 클러스터의 가장 일반적인 레이블을 지정할 수 있음.

훈련 감독 방식

자기지도 학습



레이블 생성

완전히 레이블이 지정되지 않은 데이터셋에서 완전히 레이블이 지정된 데이터셋을 생성하는 접근 방식임. 이후 지도 학습 알고리즘을 사용할 수 있음.

이미지 복원 예제

레이블이 없는 이미지 데이터셋이 있는 경우, 각 이미지의 일부를 무작위로 마스킹한 다음 원본 이미지를 복구하도록 모델을 훈련할 수 있음. 훈련 중에 마스킹된 이미지는 모델의 입력으로 사용되고 원본 이미지는 레이블로 사용됨.

전이 학습

한 작업에서 다른 작업으로 지식을 전이하는 것을 전이 학습이라고 하며, 특히 심층 신경망을 사용할 때 오늘날 기계 학습에서 가장 중요한 기술 중 하나임.

훈련 감독 방식

강화 학습



에이전트

강화 학습에서 학습 시스템은 에이전트라고 불림. 환경을 관찰하고, 행동을 선택하고 수행하며, 그 대가로 보상을 받음.



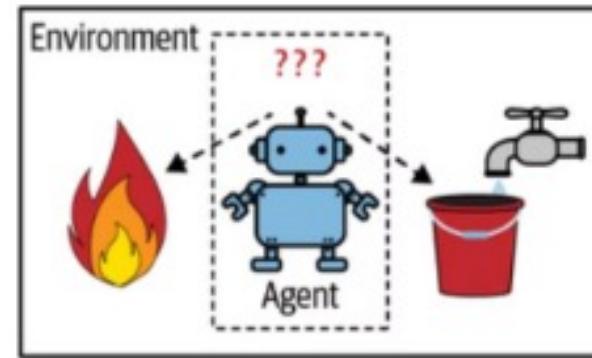
정책

에이전트는 시간이 지남에 따라 가장 많은 보상을 얻기 위한 최상의 전략(정책이라고 함)을 스스로 학습해야 함. 정책은 에이전트가 주어진 상황에서 어떤 행동을 선택해야 하는지 정의함.

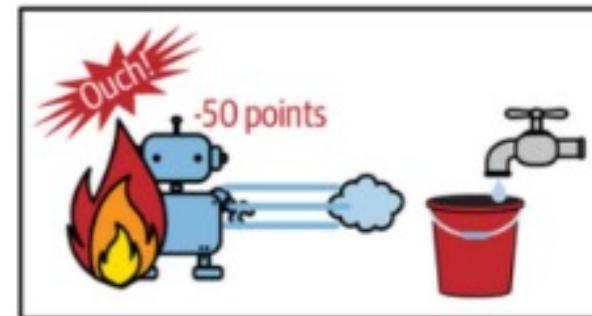


실제 응용 사례

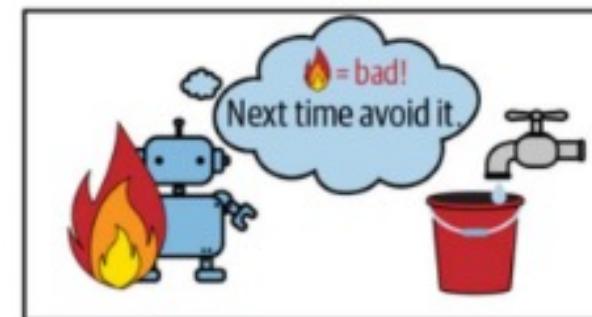
많은 로봇이 걷는 방법을 학습하기 위해 강화 학습 알고리즘을 구현함. DeepMind의 AlphaGo 프로그램도 강화 학습의 좋은 예임. 수백만 개의 게임을 분석한 다음 자신과 많은 게임을 함으로써 승리 정책을 학습함.



- 1 Observe
- 2 Select action using policy



- 3 Action!
- 4 Get reward or penalty



- 5 Update policy (learning step)
- 6 Iterate until an optimal policy is found

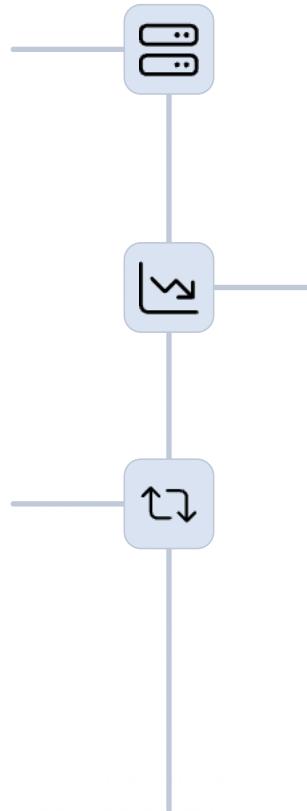
배치 학습

점진적 학습 불가

배치 학습에서 시스템은 점진적으로 학습할 수 없음.
사용 가능한 모든 데이터를 사용하여 훈련해야 함. 이는 일반적으로 많은 시간과 컴퓨팅 리소스가 필요하므로 일반적으로 오프라인에서 수행됨.

정기적 재훈련

해결책은 최신 데이터로 모델을 정기적으로 재훈련하는 것임. 얼마나 자주 해야 하는지는 사용 사례에 따라 다름. 모델이 빠르게 진화하는 시스템을 다루는 경우 상당히 빠르게 저하될 수 있음.



모델 성능 저하

불행히도 모델의 성능은 시간이 지남에 따라 천천히 저하되는 경향이 있음. 이는 모델이 변경되지 않은 상태로 유지되는 동안 세계가 계속 발전하기 때문임. 이 현상을 모델 부패 또는 데이터 드리프트라고 함.

온라인 학습



점진적 학습

온라인 학습에서는 데이터 인스턴스를 순차적으로 공급하여 시스템을 점진적으로 훈련함.



빠른 적응

각 학습 단계는 빠르고 저렴하므로 시스템은 데이터가 도착할 때 즉시 새로운 데이터에 대해 학습할 수 있음.



대용량 데이터셋

온라인 학습 알고리즘은 한 머신의 메인 메모리에 맞지 않는 거대한 데이터셋에서 모델을 훈련하는 데 사용할 수 있음(이를 코어 외 학습이라고 함).



주의사항

온라인 학습의 큰 과제는 잘못된 데이터가 시스템에 공급되면 시스템의 성능이 저하될 수 있다는 것임. 이 위험을 줄이려면 시스템을 면밀히 모니터링해야 함.

사례 기반 vs 모델 기반 학습

일반화 방법

대부분의 기계 학습 작업은 예측에 관한 것임. 이는 시스템이 이전에 본 적이 없는 예제에 대해 좋은 예측을 할 수 있어야 함을 의미함. 일반화에는 사례 기반 학습과 모델 기반 학습이라는 두 가지 주요 접근 방식이 있음.

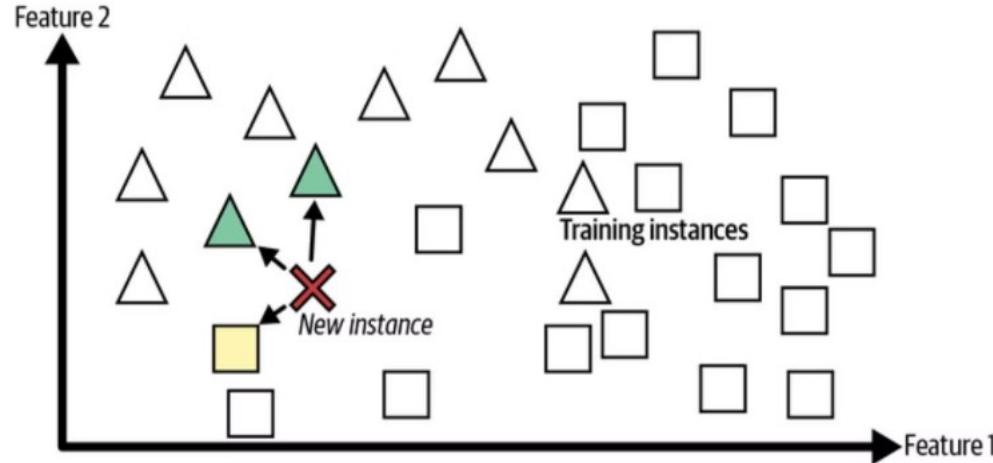
사례 기반 학습

가장 사소한 형태의 학습은 단순히 암기하는 것임. 이 방식으로 스팸 필터를 만들면 사용자가 이미 플래그를 지정한 이메일과 동일한 모든 이메일에 플래그를 지정함. 이 시스템은 유사성 측정을 사용하여 학습된 예제와 비교함으로써 새로운 사례로 일반화함.

모델 기반 학습

데이터를 기반으로 데이터의 특성/패턴을 학습하여, 미래 데이터에 대해 목적하는 스코어 값을 계산할 수 있는 인공지능 모델을 생성.

사례 기반 학습의 예



유사성 측정

동일한 스팸 이메일에 플래그를 지정하는 대신 스팸 필터는 알려진 스팸 이메일과 매우 유사한 이메일에도 플래그를 지정하도록 프로그래밍될 수 있음. 이를 위해서는 두 이메일 간의 유사성 측정이 필요함.

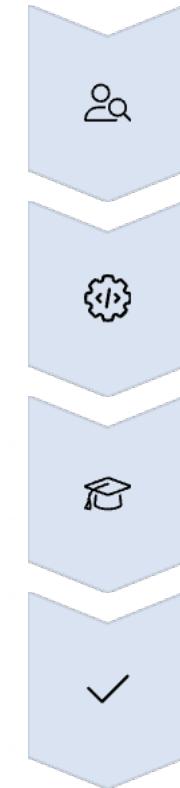
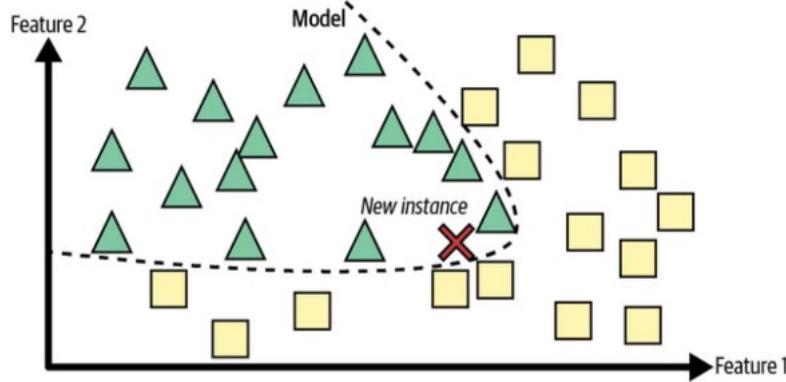
단어 공통성

두 이메일 간의 (매우 기본적인) 유사성 측정은 공통으로 가지고 있는 단어 수를 계산하는 것일 수 있음. 시스템은 알려진 스팸 이메일과 많은 단어를 공유하는 경우 이메일을 스팸으로 플래그 지정함.

다수결 원칙

그림에서 볼 수 있듯이 새 인스턴스는 가장 유사한 인스턴스의 대다수가 해당 클래스에 속하기 때문에 삼각형으로 분류됨.

모델 기반 학습



데이터 연구

데이터를 분석하고 패턴을 찾음.

모델 선택

데이터에 적합한 모델 유형을 선택함.

모델 훈련

학습 알고리즘이 모델 매개변수 값을 찾음.

예측 적용

새로운 사례에 대한 예측을 위해 모델을 사용함.

예를 들어, 돈이 사람들을 행복하게 만드는지 알고 싶다고 가정해 보자. OECD 웹사이트에서 Better Life Index 데이터와 World Bank에서 1인당 국내총생산(GDP) 통계를 다운로드함. 데이터를 분석하면 1인당 GDP가 증가함에 따라 생활 만족도가 대략 선형적으로 증가하는 것으로 보임.

기계 학습의 주요 챌린지

나쁜 모델

모델이 데이터에 적합하지 않거나, 너무 단순하거나 복잡할 수 있음. 모델 선택과 하이퍼파라미터 튜닝이 중요함.

나쁜 데이터

훈련 데이터가 충분하지 않거나, 대표성이 없거나, 품질이 낮거나, 관련 없는 특성이 많을 수 있음. 데이터 품질과 특성 엔지니어링이 중요함.

과적합과 과소적합

모델이 훈련 데이터에 너무 잘 맞거나(과적합) 충분히 맞지 않는(과소적합) 문제가 발생할 수 있음. 정규화와 모델 복잡성 조정이 필요함.

불충분한 훈련 데이터

1000+

간단한 문제

가장 간단한 문제에도 일반적으로 수천 개의 예제가 필요함.

1M+

복잡한 문제

이미지나 음성 인식과 같은 복잡한 문제의 경우 수백만 개의 예제가 필요할 수 있음.

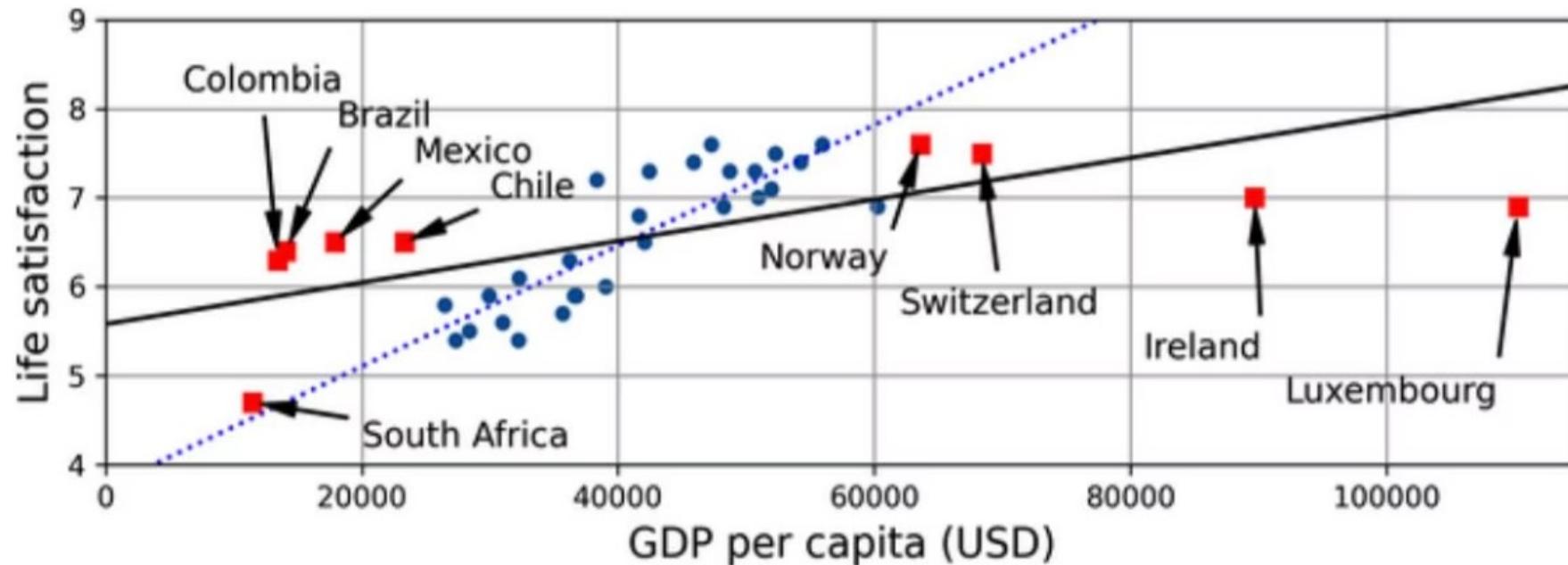
68%

데이터 효과

Microsoft 연구원들은 충분한 데이터가 주어지면 다양한 기계 학습 알고리즘이 복잡한 자연어 모호성 문제에서 거의 동일하게 잘 수행 된다는 것을 보여줌.

"데이터가 알고리즘보다 복잡한 문제에 더 중요하다"는 아이디어는 2009년 Peter Norvig 등이 발표한 "데이터의 비합리적 효과성"이라는 논문에서 더욱 대중화됨. 그러나 작은 규모와 중간 규모의 데이터셋은 여전히 매우 일반적이며, 추가 훈련 데이터를 얻는 것이 항상 쉽거나 저렴한 것은 아님

훈련 데이터의 대표성 이슈



잘 일반화하기 위해서는 훈련 데이터가 일반화하려는 새로운 사례를 대표하는 것이 중요함. 이는 사례 기반 학습이든 모델 기반 학습이든 마찬가지임.

예를 들어, 선형 모델 훈련에 사용한 국가 집합은 완벽하게 대표적이지 않았음. 1인당 GDP가 \$23,500 미만이거나 \$62,500를 초과하는 국가는 포함되지 않았음. 이러한 국가를 추가하면 모델이 크게 변경될 뿐만 아니라 이러한 단순한 선형 모델이 잘 작동하지 않을 것임이 분명해짐.

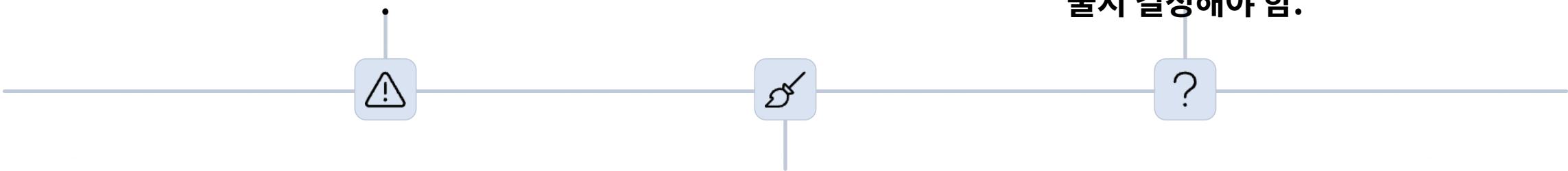
데이터의 품질 이슈

오류와 이상치

훈련 데이터에 오류, 이상치 및 노이즈(예: 품질이 낮은 측정으로 인한)가 많으면 시스템이 기본 패턴을 감지하기 어려워져 시스템이 잘 수행될 가능성이 낮아짐

누락된 값 처리

일부 인스턴스에 몇 가지 특성이 누락된 경우(예: 고객의 5%가 나이를 지정하지 않음), 이 속성을 완전히 무시할지, 이러한 인스턴스를 무시할지, 누락된 값을 채울지 결정해야 함.



데이터 정리

훈련 데이터를 정리하는 데 시간을 투자하는 것이 종종 가치가 있음. 사실, 대부분의 데이터 과학자는 상당한 시간을 이 작업에 할애함.

데이터 특징 (피처) 이슈

특성 선택

기존 특성 중에서 훈련에 가장 유용한 특성을 선택함. 이는 모델의 성능을 향상시키고 과적합 위험을 줄이는 데 도움이 됨.

특성 추출

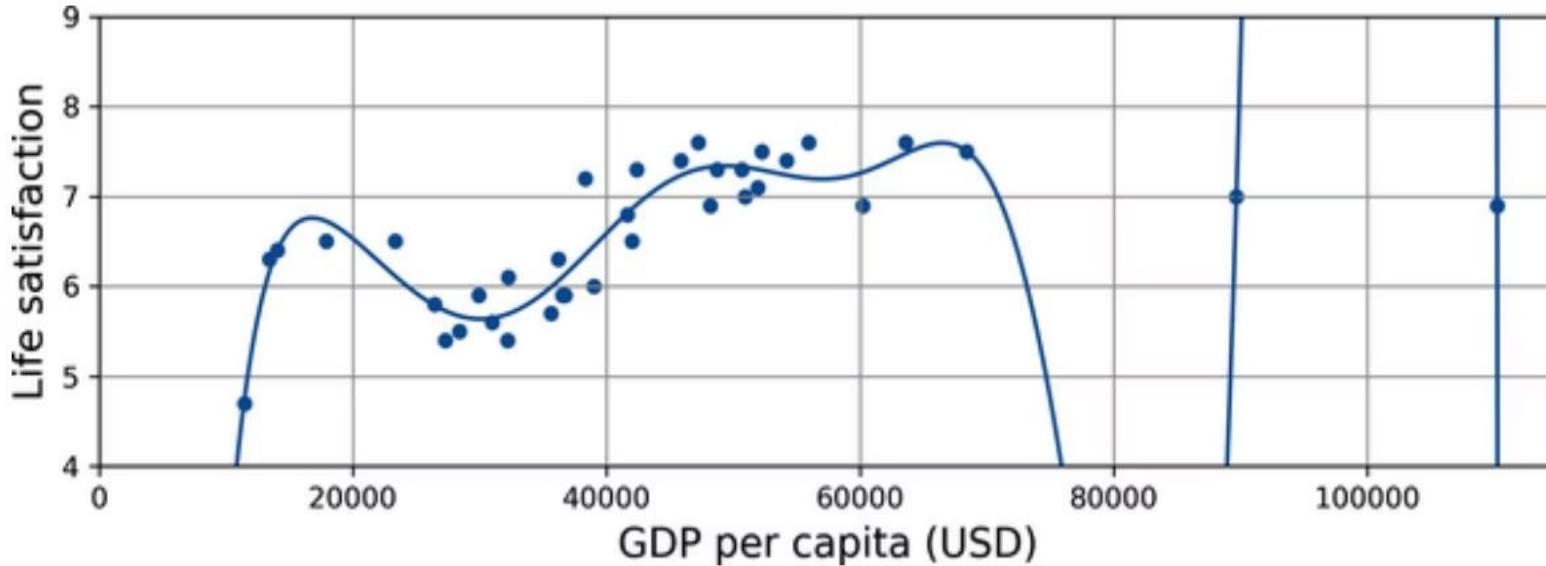
기존 특성을 결합하여 더 유용한 특성을 생성함. 차원 축소 알고리즘이 이 과정에서 도움이 될 수 있음.

새로운 특성 생성

새로운 데이터를 수집하여 모델에 더 유용한 정보를 제공할 수 있는 새로운 특성을 만듦.

속담에서 말하듯이: 쓰레기를 넣으면 쓰레기가 나옴. 시스템은 훈련 데이터에 충분한 관련 특성이 포함되어 있고 관련 없는 특성이 너무 많지 않은 경우에만 학습할 수 있음. 기계 학습 프로젝트 성공의 중요한 부분은 훈련할 좋은 특성 세트를 만드는 것임.

훈련 데이터 과적합



과적합이란?

과적합은 모델이 훈련 데이터에서는 잘 수행되지만 일반화가 잘 되지 않는 것을 의미함. 이는 모델이 데이터의 실제 패턴이 아닌 노이즈 패턴을 감지할 때 발생함.

해결책

과적합은 모델이 훈련 데이터의 양과 노이즈에 비해 너무 복잡할 때 발생함. 가능한 해결책은 다음과 같음:

- 매개변수가 적은 모델 선택, 훈련 데이터의 속성 수 줄이기, 모델 제약 등을 통해 모델 단순화
- 더 많은 훈련 데이터 수집
- 훈련 데이터의 노이즈 감소(데이터 오류 수정 및 이상치 제거)

훈련 데이터 과소적합



과소적합이란?

과소적합은 과적합의 반대임. 모델이 데이터의 기본 구조를 학습하기에 너무 단순할 때 발생함. 예를 들어, 생활 만족도의 선형 모델은 과소적합되기 쉬움



더 강력한 모델

매개변수가 더 많은 더 강력한 모델을 선택함. 복잡한 패턴을 포착 하려면 더 표현력이 풍부한 모델이 필요할 수 있음.



특성 엔지니어링

학습 알고리즘에 더 나은 특성을 제공함. 좋은 특성은 모델이 데이터의 패턴을 더 쉽게 학습할 수 있게 함.



제약 완화

모델에 대한 제약을 줄임(예: 정규화 하이퍼파라미터 감소). 이를 통해 모델이 더 복잡한 패턴을 학습할 수 있음.

중간 정리

기계 학습의 본질

기계 학습은 규칙을 명시적으로 코딩하는 대신 데이터에서 학습함으로써 기계가 일부 작업에서 더 나아지도록 하는 것임.

ML 프로젝트 워크플로

ML 프로젝트에서는 훈련 세트에서 데이터를 수집하고 이를 학습 알고리즘에 공급함. 알고리즘이 모델 기반인 경우 모델을 훈련 세트에 맞추기 위해 일부 매개변수를 조정함.

다양한 ML 시스템

지도 여부, 배치 또는 온라인, 사례 기반 또는 모델 기반 등 다양한 유형의 ML 시스템이 있음.

성능 영향 요소

훈련 세트가 너무 작거나, 데이터가 대표적이지 않거나, 노이즈가 있거나, 관련 없는 특성으로 오염된 경우 시스템이 제대로 수행되지 않음. 또한 모델이 너무 단순하거나(과소적합) 너무 복잡해서는(과적합) 안 됨.

테스트 및 검증

모델 평가의 중요성

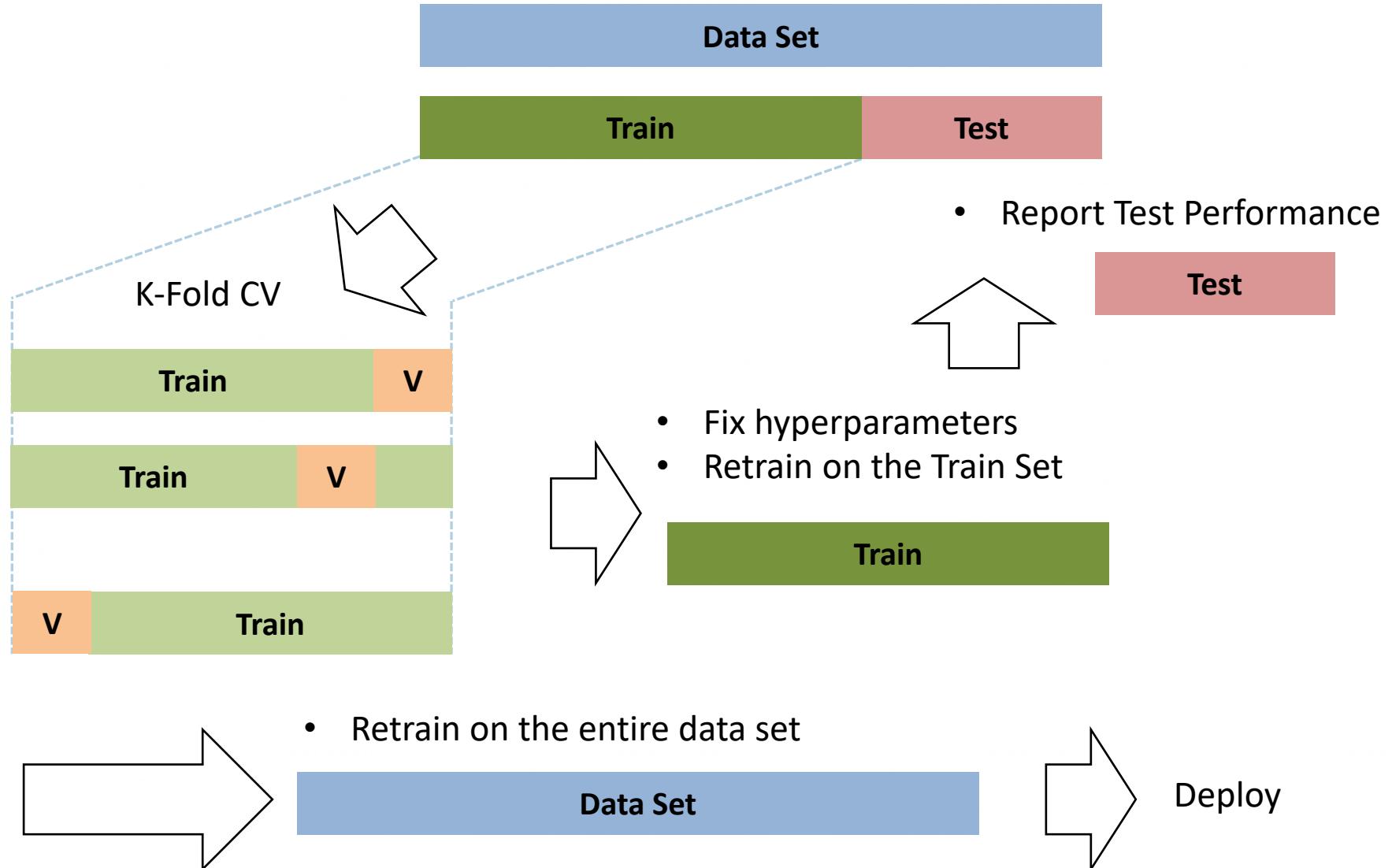
모델이 새로운 사례에 얼마나 잘 일반화될지 아는 유일한 방법은 실제로 새로운 사례에서 시도해 보는 것임. 한 가지 방법은 모델을 프로덕션에 배포하고 얼마나 잘 수행되는지 모니터링하는 것임. 이 방법은 잘 작동하지만 모델이 매우 나쁘면 사용자가 불평할 것임.

데이터 분할

더 나은 옵션은 데이터를 훈련 세트와 테스트 세트의 두 세트로 분할하는 것임. 이름에서 알 수 있듯이 훈련 세트를 사용하여 모델을 훈련하고 테스트 세트를 사용하여 테스트함.

새로운 사례에 대한 오류율을 일반화 오류(또는 샘플 외 오류)라고 하며, 테스트 세트에서 모델을 평가하여 이 오류의 추정치를 얻을 수 있음. 이 값은 모델이 이전에 본 적이 없는 인스턴스에서 얼마나 잘 수행될지 알려줌.

ML Development Cycle



Data

- Supervised Learning: X (input), Y (output)
- Unsupervised Learning: X (input), no Y
- Semi-supervised Learning: (X₁,Y₁) and X₂
- Self-supervised Learning: X → (X', Y')

In Statistical Learning...

Data Set $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$

A set consists of m samples

Note:

- There can be many possible data sets
- We're dealing with just one set of samples

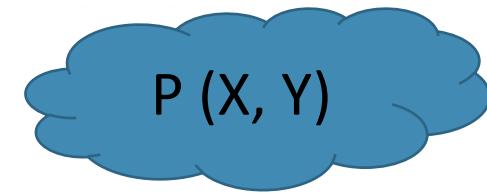
Sampling (i.i.d)

i : independent
i.d.: identical distribution

P (X, Y)

Probability distribution
(usually, unknown)

Performance



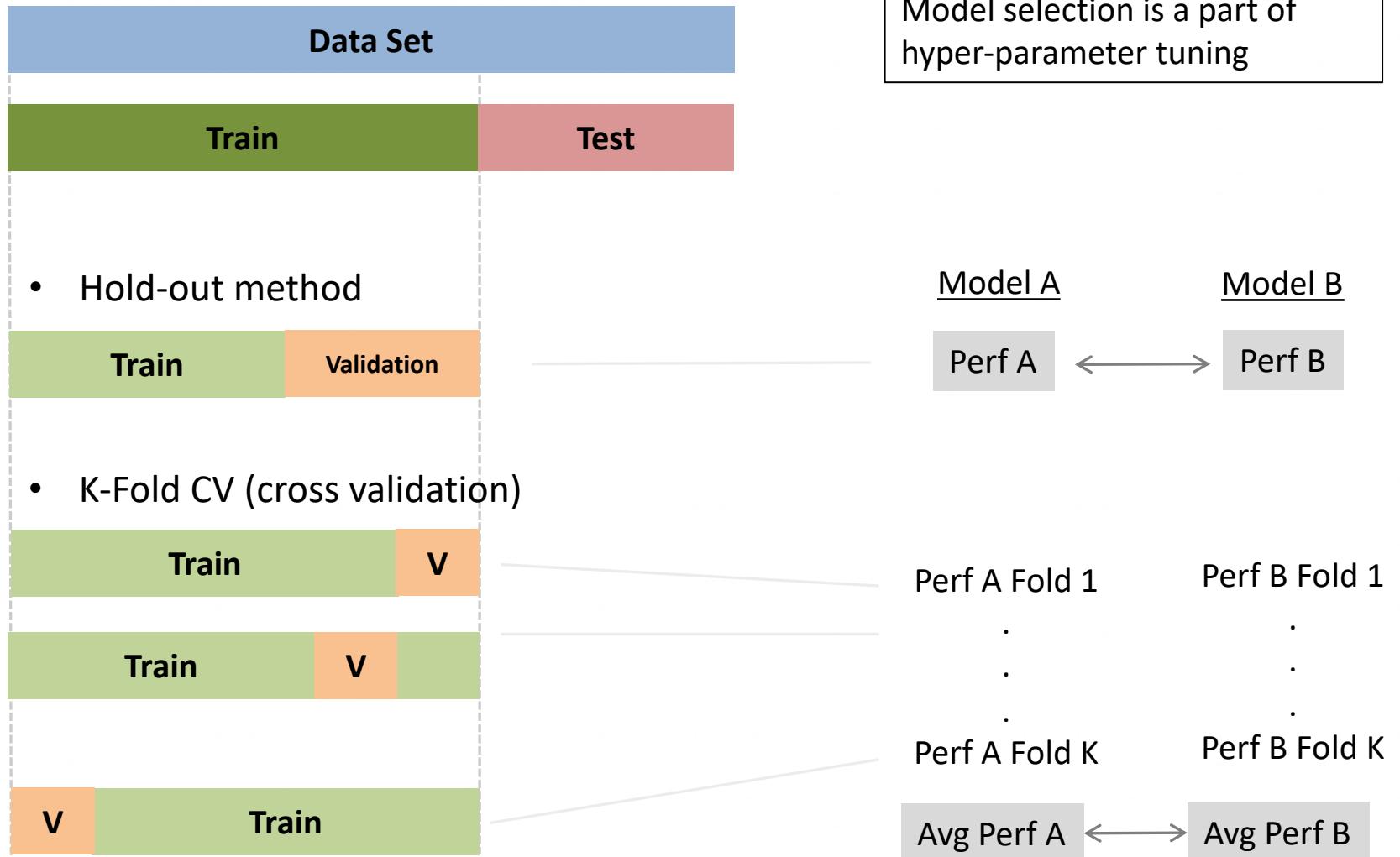
- **Training Error (Rate)** : error on the training set
- **Test Error (Rate)** : error on the test set
- **Generalization Error**: error on the **all possible** data

$$\frac{1}{|\text{tr}|} \sum_{i \in \text{tr}} \mathbf{1}[y^{(i)} \neq f_w(x^{(i)})]$$

$$\frac{1}{|\text{tt}|} \sum_{i \in \text{tt}} \mathbf{1}[y^{(i)} \neq f_w(x^{(i)})]$$

$$\mathbb{E}_{(X, Y)} [\mathbf{1}[Y \neq f_w(X)]]$$

Model Selection



MNIST

MNIST 데이터셋 소개

이 장에서는 MNIST 데이터셋을 사용할 것임. 이 데이터셋은 고등학생과 미국 인구조사국 직원들이 손으로 쓴 70,000개의 작은 숫자 이미지로 구성됨. 각 이미지에는 해당 숫자가 레이블로 지정되어 있음.

이 데이터셋은 너무 많이 연구되어 머신 러닝의 "헬로 월드"라고 불림. 새로운 분류 알고리즘을 개발할 때마다 MNIST에서 어떤 성능을 보이는지 확인하고, 머신 러닝을 배우는 사람이라면 누구나 이 데이터셋을 다루게 됨.

데이터 구조

Scikit-Learn은 인기 있는 데이터셋을 다운로드하는 여러 헬퍼 함수를 제공함. MNIST도 그 중 하나임. 각 이미지는 28×28 픽셀이며, 각 특성은 하나의 픽셀 강도를 나타냄(0은 흰색, 255는 검은색).

데이터셋은 70,000개의 이미지로 구성되어 있으며, 각 이미지는 784개의 특성을 가짐. 이미 훈련 세트(처음 60,000개 이미지)와 테스트 세트(마지막 10,000개 이미지)로 나뉘어 있음.



이진 분류기 훈련



문제 단순화

문제를 단순화하기 위해 하나의 숫자(예: 5)만 식별하는 것으로 시작함. 이 "5-감지기"는 "5"와 "5 아님"라는 두 클래스만 구별할 수 있는 이진 분류기의 예시임.



분류기 선택

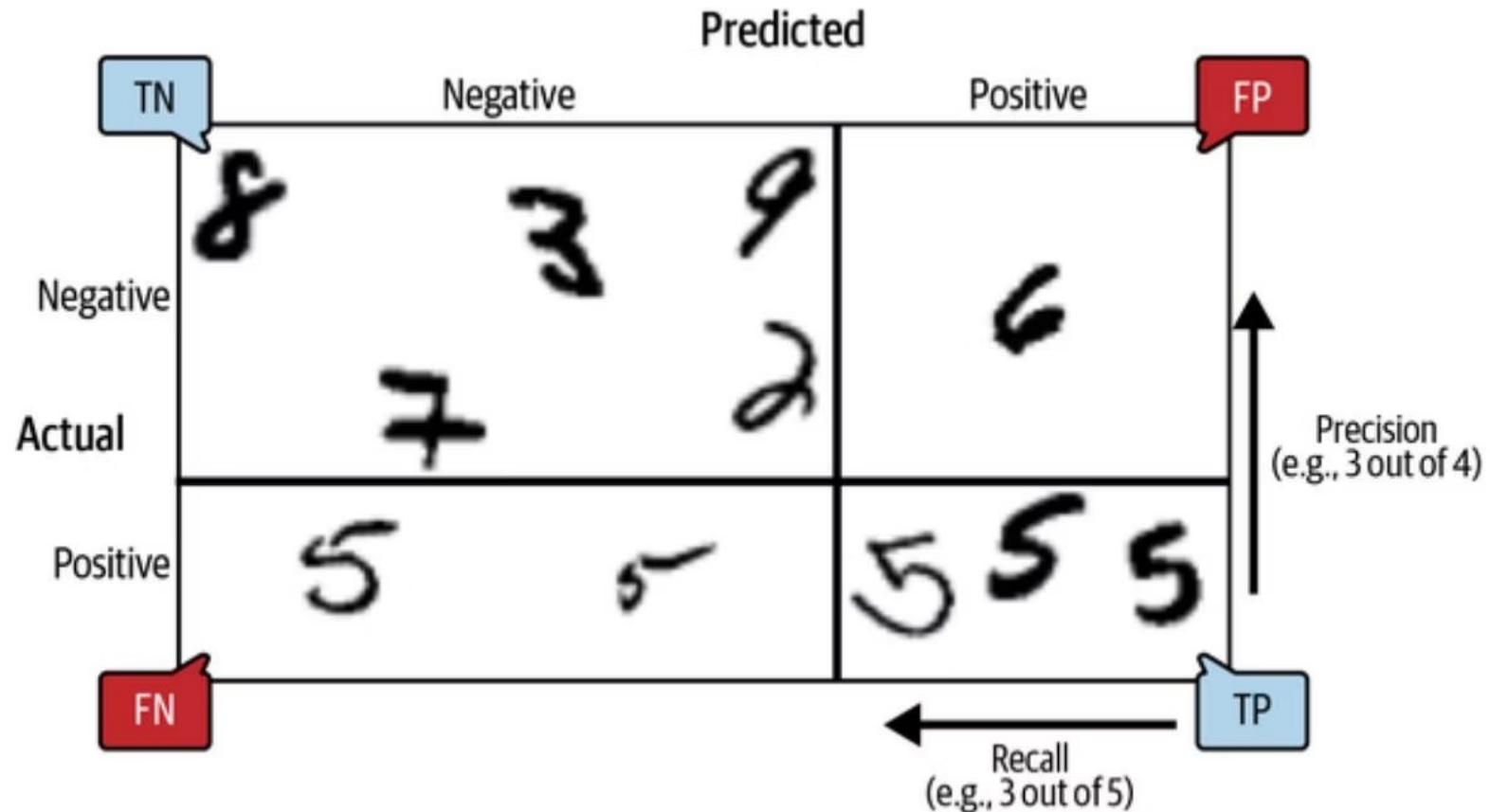
확률적 경사 하강법(SGD) 분류기를 사용함. Scikit-Learn의 SGDClassifier 클래스는 매우 큰 데이터셋을 효율적으로 처리할 수 있음. SGD는 훈련 인스턴스를 독립적으로 하나씩 처리하기 때문에 온라인 학습에도 적합함.



모델 평가

분류기를 훈련한 후에는 이를 사용하여 숫자 5의 이미지를 감지할 수 있음. 이제 이 모델의 성능을 평가해야 함. 분류기를 평가하는 것은 회귀 모델을 평가하는 것보다 훨씬 더 주의가 필요함

혼동 행렬



혼동 행렬의 개념

혼동 행렬의 일반적인 아이디어는 클래스 A의 인스턴스가 클래스 B로 분류되는 횟수를 모든 A/B 쌍에 대해 계산하는 것임. 예를 들어, 분류기가 8의 이미지를 0으로 혼동한 횟수를 알고 싶다면 혼동 행렬의 8행, 0열을 확인하면 됨.

Copyright © 2025 Sangkyun Lee

정밀도와 재현율

정밀도: 양성 예측의 정확도, $TP/(TP+FP)$ 로 계산됨.

재현율(민감도, TPR): 분류기가 올바르게 감지한 양성 인스턴스의 비율, $TP/(TP+FN)$ 로 계산됨.

F1 점수

정밀도와 재현율을 단일 지표로 결합하는 것이 편리할 때가 있음. F1 점수는 정밀도와 재현율의 조화 평균임. 일반 평균은 모든 값을 동등하게 취급하지만, 조화 평균은 낮은 값에 훨씬 더 많은 가중치를 줌.

Confusion Matrix

		Predicted class	
		P	N
		True Positives (TP)	False Negatives (FN)
Actual Class	P	True Positives (TP)	False Negatives (FN)
	N	False Positives (FP)	True Negatives (TN)

Accuracy Rate:

$$ACC = \frac{\text{Correct}}{\text{ALL}} = \frac{TP + TN}{ALL} = 1 - ERR$$

Precision:

$$PRE = \frac{TP}{\text{Predicted } P} = \frac{TP}{TP + FP}$$

Recall:

$$REC = TPR = \frac{TP}{\text{Actual } P} = \frac{TP}{TP + FN}$$

Q. A dataset with (Female, Male) = (50,50) students. A classifier always says an input as Female. What is the Precision and Recall of the classifier ?

정밀도와 재현율

1

정밀도와 재현율 계산

Scikit-Learn은 정밀도와 재현율을 포함한 분류기 메트릭을 계산하기 위한 여러 함수를 제공함. 우리의 5-감지기는 이미지가 5를 나타낸다고 주장할 때 83.7%만 정확하며, 5의 65.1%만 감지함.

2

F1 점수 활용

F1 점수는 두 분류기를 비교하기 위한 단일 메트릭이 필요할 때 특히 편리함. 분류기는 정밀도와 재현율이 모두 높은 경우에만 높은 F1 점수를 얻을 수 있음.

3

상황에 맞는 선택

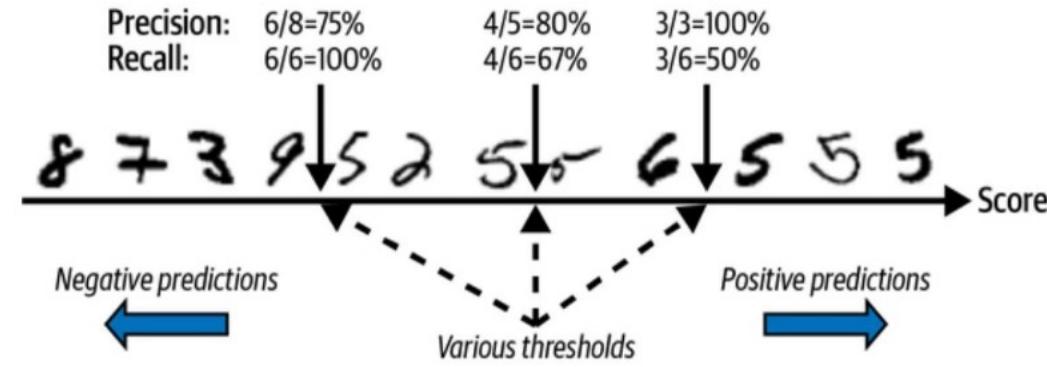
일부 상황에서는 주로 정밀도에 관심이 있고, 다른 상황에서는 재현율에 관심이 있을 수 있음. 예를 들어, 아이들에게 안전한 비디오를 감지하는 분류기를 훈련시킨다면, 많은 좋은 비디오를 거부하더라도(낮은 재현율) 안전한 비디오만 유지하는(높은 정밀도) 분류기를 선호할 수 있음.

F1 Score

- The harmonic mean of precision and recall

$$F1 = 2 \frac{PRE \times REC}{PRE + REC}$$

정밀도/재현율 트레이드오프



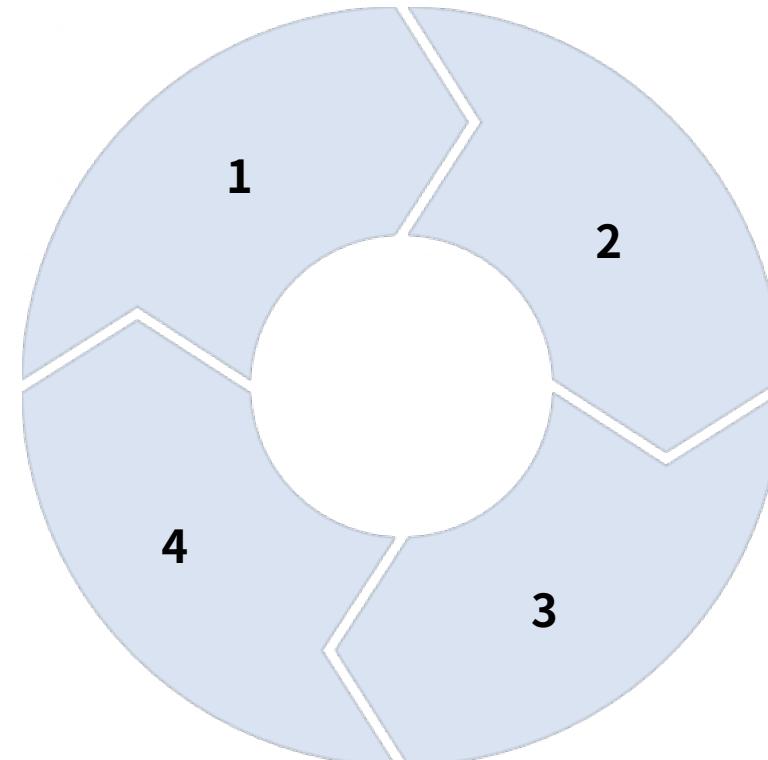
결정 함수 이해

대부분 분류기는 각 인스턴스에 대해 결정 함수를 기반으로 점수를 계산

그 점수가 임계값보다 크면 인스턴스를 양성으로, 그렇지 않으면 음성으로 분류

시각화 및 분석

정밀도와 재현율을 시각화하여 비교



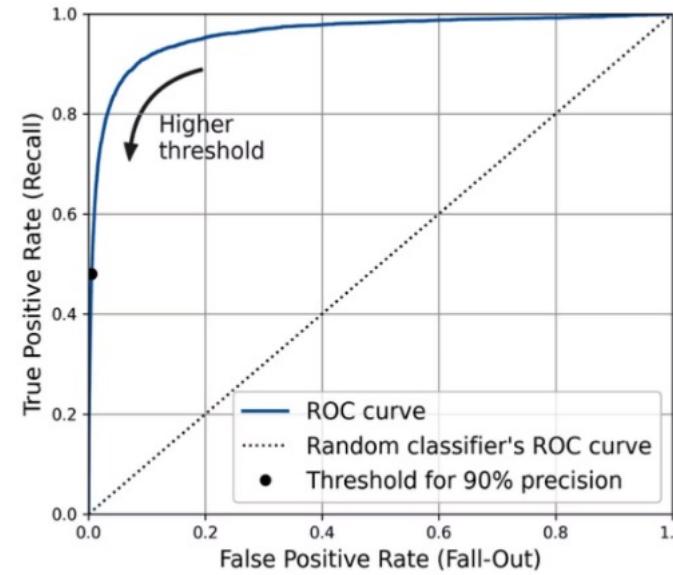
임계값 조정

위 그림: 임계값을 높이면 거짓 양성(6)이 참 음성이 되어 정밀도가 증가하지만, 하나의 참 양성이 거짓 음성이 되어 재현율이 감소함. 반대로 임계값을 낮추면 재현율이 증가하고 정밀도가 감소함.

최적 임계값 선택

어떤 임계값을 사용할지 결정하기 위해 훈련 세트의 모든 인스턴스에 대한 점수를 얻은 다음, 모든 가능한 임계값에 대한 정밀도와 재현율을 계산

ROC (Receiver Operator Characteristic) 곡선



ROC 곡선 이해

수신자 조작 특성(ROC) 곡선은 이진 분류기에 사용되는 또 다른 일반적인 도구임. 정밀도 대 재현율 곡선과 매우 유사하지만, 참 양성 비율(재현율의 다른 이름)을 거짓 양성 비율(FPR)에 대해 플롯함.

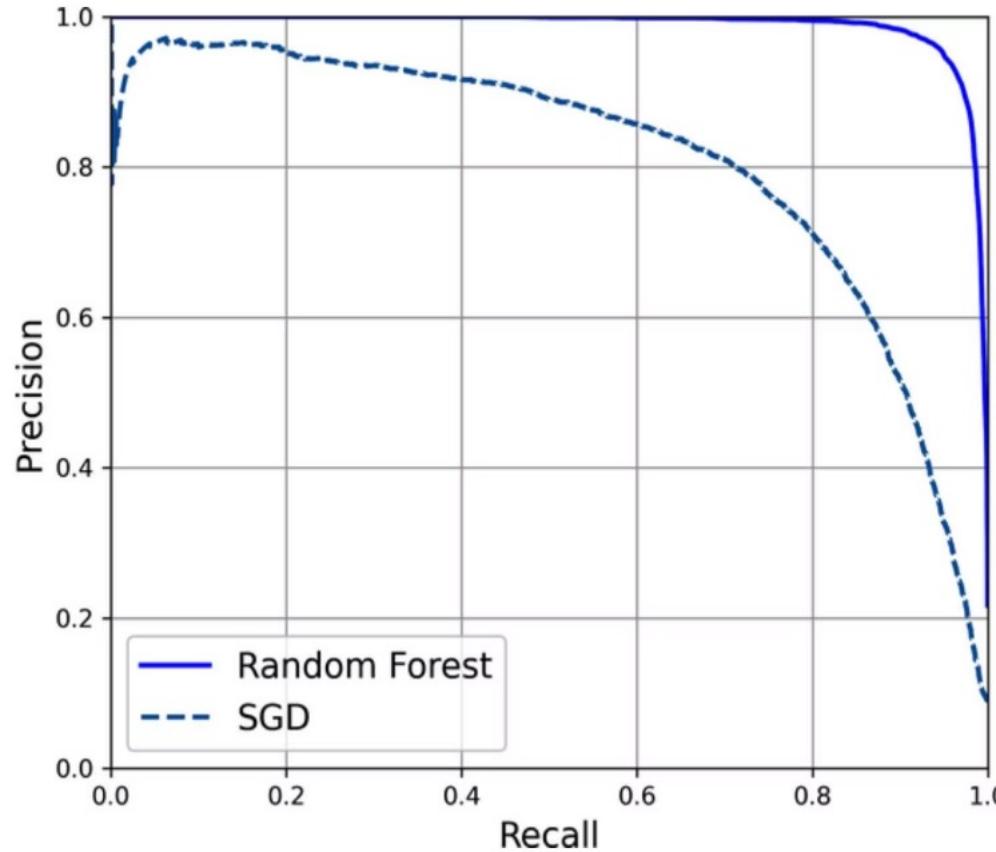
ROC 곡선 생성

ROC 곡선을 플롯하려면 먼저 `roc_curve()` 함수를 사용하여 다양한 임계값에 대한 TPR과 FPR을 계산한 다음 Matplotlib을 사용하여 FPR을 TPR에 대해 플롯함.

AUC 측정

분류기를 비교하는 한 가지 방법은 곡선 아래 영역(AUC)을 측정하는 것임. 완벽한 분류기는 ROC AUC가 1이고, 순전히 무작위 분류기는 ROC AUC가 0.5임.

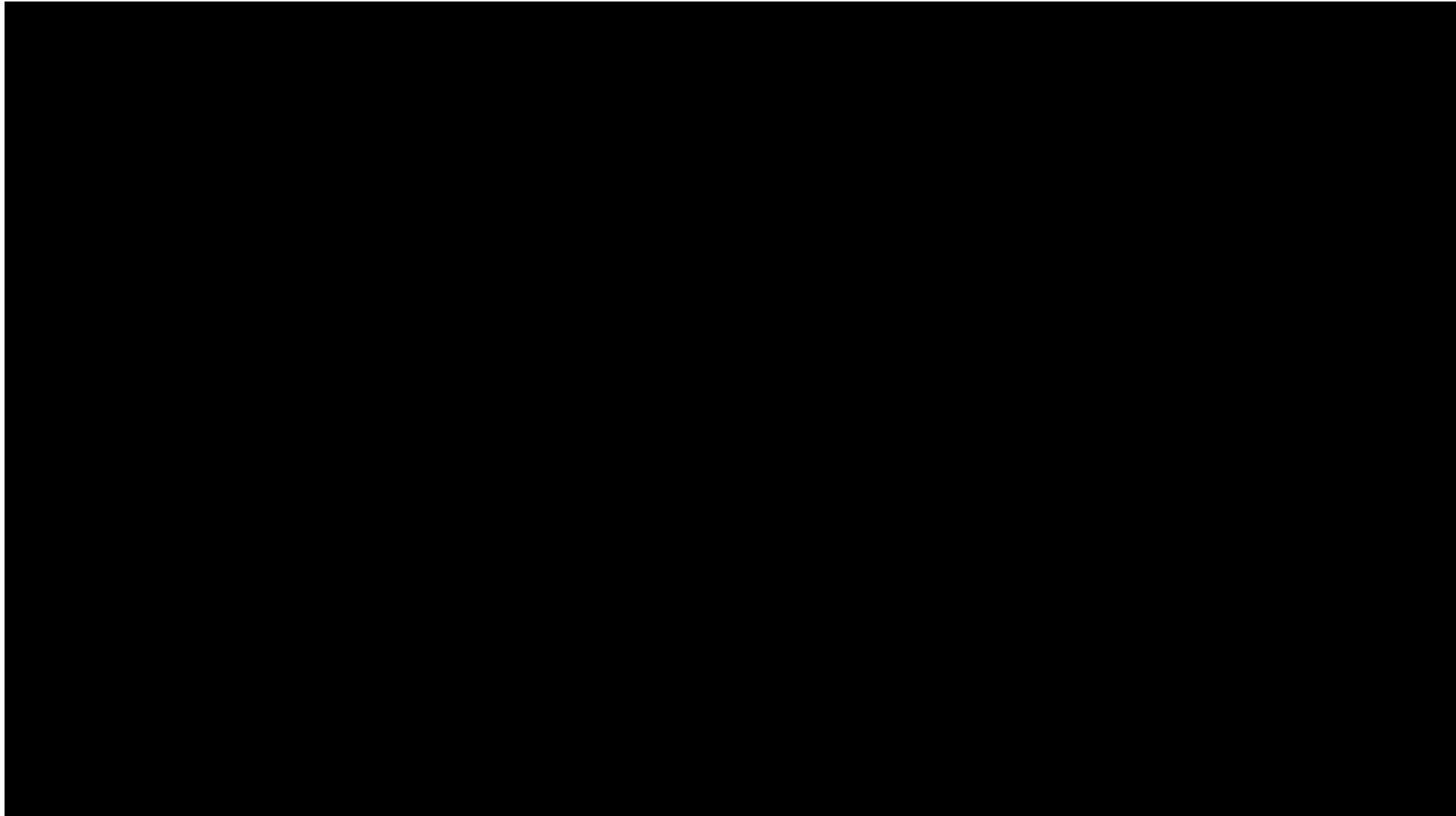
Precision-Recall 곡선



성능 비교

PR 곡선을 플롯하면 RandomForestClassifier의 PR 곡선이 SGDClassifier보다 훨씬 더 좋아 보임. 오른쪽 상단 모서리에 훨씬 더 가깝고 AUC가 더 큼. F1 점수와 ROC AUC 점수도 상당히 더 좋음.

Turing Award 2018

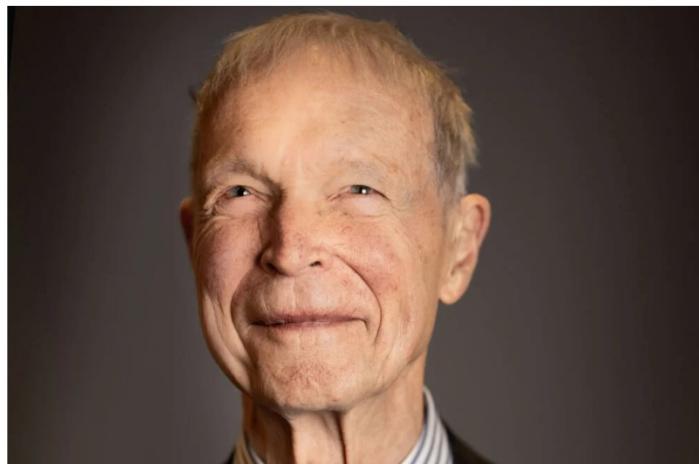


<https://www.youtube.com/watch?v=HzilDIhWhrE>

The Nobel Prize in Physics 2024

John J. Hopfield

“for foundational discoveries and inventions that enable machine learning with artificial neural networks”



© Nobel Prize Outreach. Photo: Nanaka Adachi

Hopfield network (1982)

Geoffrey Hinton

“for foundational discoveries and inventions that enable machine learning with artificial neural networks”



© Nobel Prize Outreach. Photo: Clément Morin

1984

Hinton, G.-E., Sejnowski, T. J., and Ackley, D. H.
Boltzmann Machines: Constraint satisfaction networks that learn.
Technical Report CMU-CS-84-119, Carnegie-Mellon University. [\[pdf\]](#)

They used physics to find patterns in information

This year’s laureates used tools from physics to construct methods that helped lay the foundation for today’s powerful machine learning. John Hopfield created a structure that can store and reconstruct information. Geoffrey Hinton invented a method that can independently discover properties in data and which has become important for the large artificial neural networks now in use.

Related articles

[Press release](#)

[Popular information: They used physics to find patterns in information](#)

[Scientific background: “for foundational discoveries and inventions that enable machine learning with artificial neural networks”](#)

The Nobel Prize in Chemistry 2024

David Baker

“for computational protein design”



© Nobel Prize Outreach. Photo: Clément Morin

Biochemist (U of Washington)

Demis Hassabis

“for protein structure prediction”



© Nobel Prize Outreach. Photo: Clément Morin

Co-founder of Deepmind
(76년생)

John Jumper

“for protein structure prediction”



© Nobel Prize Outreach. Photo: Clément Morin

Director of Deepmind
(85년생)

Thank You