명예학회 3기
커리큘럼 개론

AI명예학회

SKHU

머신러닝의 정의와 분류

AI명예학회
SKHU

1 **전공 강의에서 소개되지 않는 기반 지식**

전공 강의와 연계되어 있지만 강의에서는 간단히 넘어가거나 다루지 않는 중요한 지식들을 소개해 이해를 돕겠습니다.

2 **코드보다 구조적인 이해를 우선시**

전공 강의에서 많이 다루는 코드보다는 모델 구조, 수학적 원리 등 이론에 더 많은 시간을 할당했습니다.

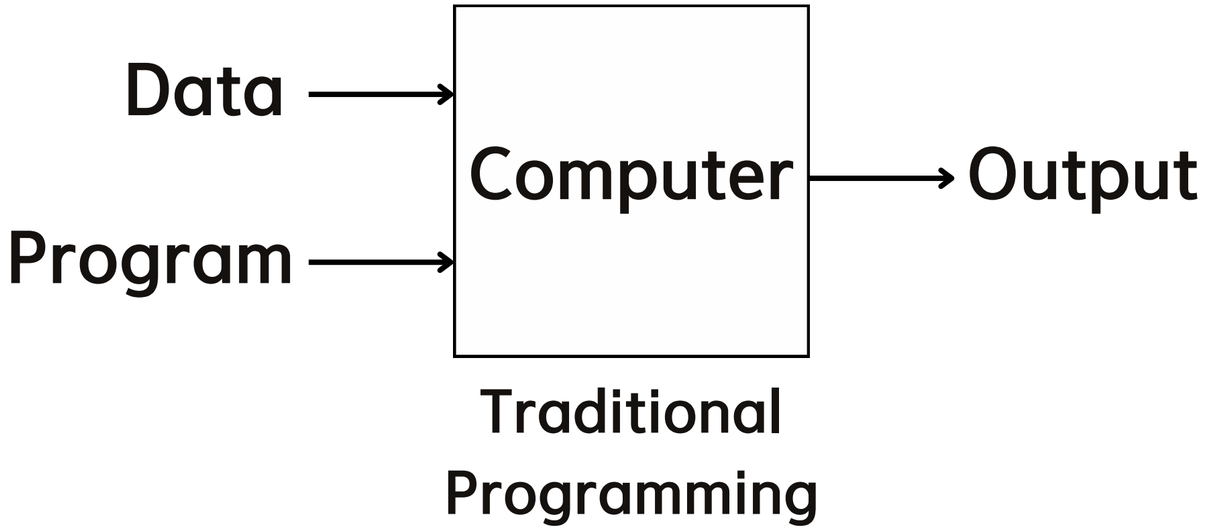
3 **인공지능 문제해결 과정**

머신러닝 개념만이 아니라 인공지능 문제해결 과정 전반에 대한 실습과 동반해 분석 방법, 모델 선택법 등을 소개해 이론과 더불어 실전에도 유용한 커리큘럼이 되도록 하겠습니다.

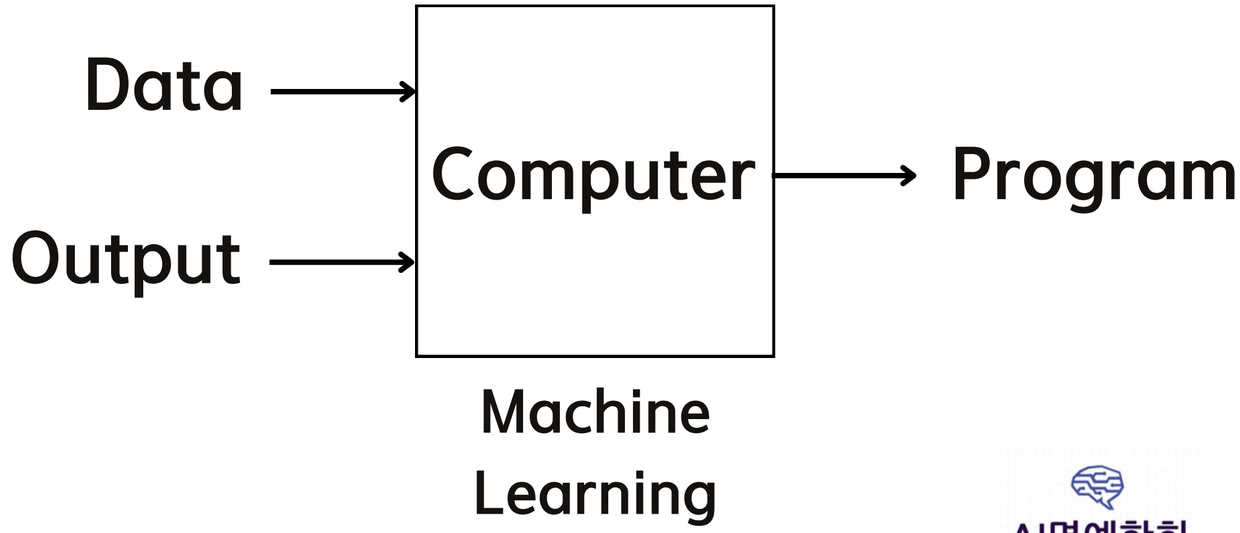
AI명예학회
SKHU

머신러닝이란?

전통적인 프로그래밍 기법과 다르게, 컴퓨터가 데이터에 기반해 학습하고 성장할 수 있는 알고리즘를 다루는 인공지능의 분야이다.
Improve on task T, with respect to performance metric P, based on experience E



Traditional Programming



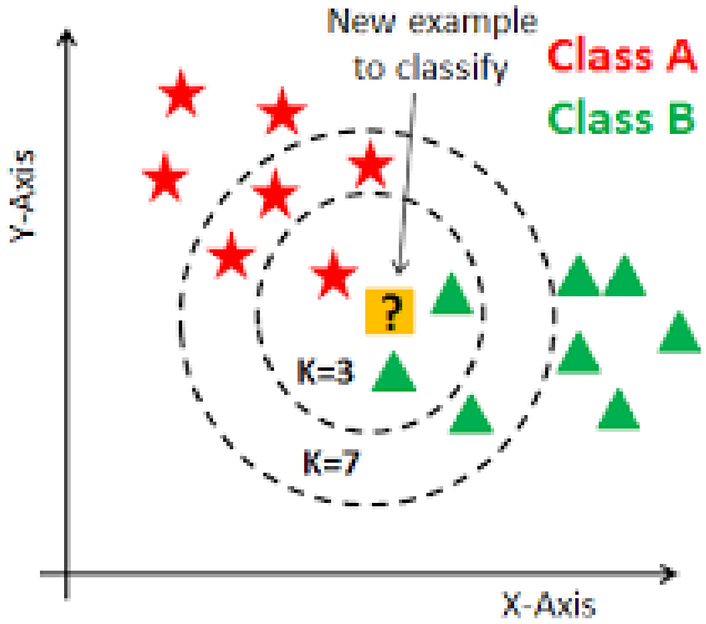
Machine Learning

AI명예학회
SKHU

사례 기반 학습 vs 모델 기반 학습

사례 기반 학습 (instance-based learning)

가장 간단한 형태의 학습은 단순히 기억하는 것이다.
단순히 훈련 데이터를 기억해 새로운 입력과 이전의 훈련 데이터가 같다면 동일한 레이블로 분류하는 것이다.
더 나아가, 단순히 동일함을 따지는 것이 아니라 유사도를 측정해서 구분하는 알고리즘을 가르친다.



학습하며 가중치를 업데이트하는 모델이 존재하는 것이 아니라 데이터 샘플과 고정되어있는 알고리즘만이 시스템을 이룬다.

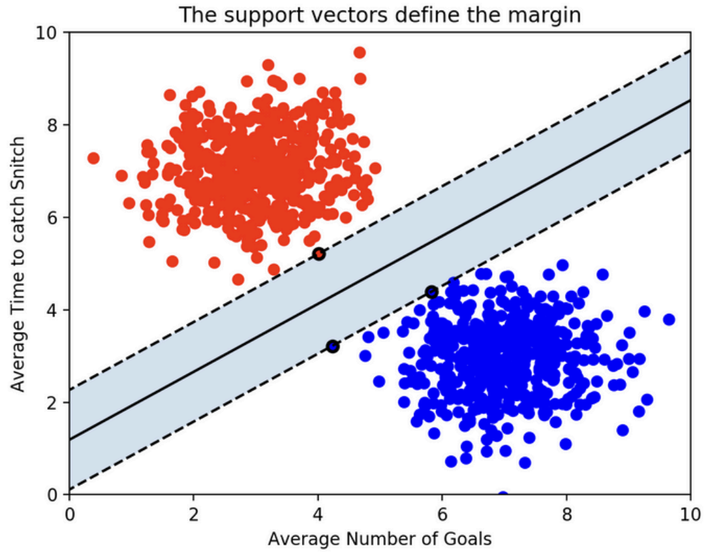
AI명예학회
SKHU

사례 기반 학습 vs 모델 기반 학습

모델 기반 학습 (model-based learning)

모델 기반 학습은 기존 샘플(훈련 데이터)로부터 모델을 만들어 예측에 사용하는 것이다.
모델이란 사례 기반 학습에서의 고정적인 알고리즘이 아닌 학습하면서 그 가중치가 예측에 최적화되도록 변화하는 시스템이다.

훈련데이터를 전부 기억하는 것이 아니라, 실제 데이터를 표현하는 함수와 근사한 가중치 값을 찾으며 학습한다.



AI명예학회
SKHU

사례 기반 학습 vs 모델 기반 학습

데이터에만 의존해 추정하기 때문에 잘못 추정된 모델을 사용하면 결과가 수렴하지 못할 수 있는 모델 기반 학습과 같은 위험성이 없다.

그러나 정확한 예측을 위해서는 많은 수의 데이터를 필요로 한다.

모델 기반 학습에서는 결국 학습할 수 있는 가중치들로 이루어진 일종의 함수를 만든다고 볼 수 있는데,
이 학습된 함수가 실제 함수와는 다를 위험성이 있다.
사용하는 모델에 따라 다양한 task에 적용할 수 있고, 사례기반 학습에 비해 복잡한 작업도 수행 가능하다.

AI명예학회
SKHU

학습의 종류

지도 학습 : 훈련데이터에 원하는 출력이 포함되어 있음

비지도 학습 : 훈련데이터에 원하는 출력이 포함되어있지 않음

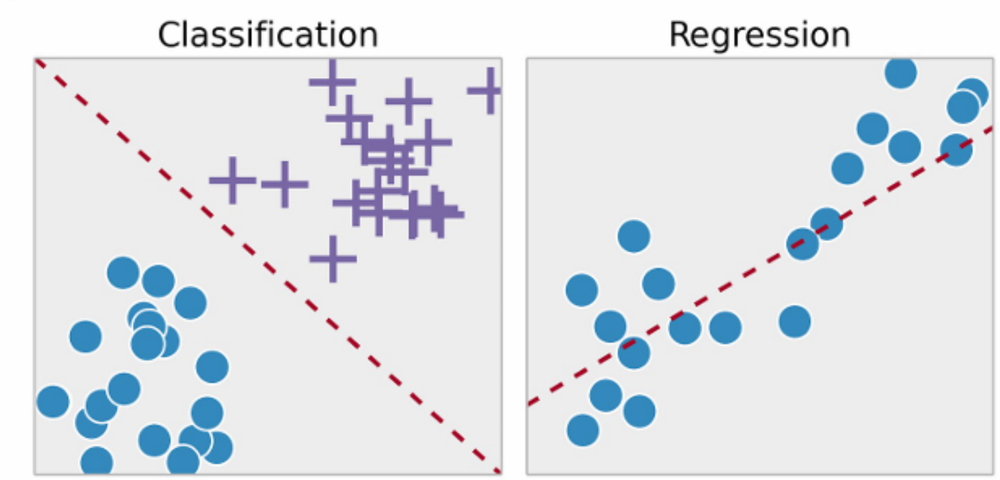
준지도 학습 : 일부 훈련데이터에만 원하는 출력이 포함되어 있음

강화 학습 : 데이터셋이 아니라 환경이 주어지고,
에이전트의 행동에 따른 보상으로 학습

AI명예학회
SKHU

지도 학습(Supervised Learning)

지도 학습은 훈련 데이터가 변수와 정답 쌍으로 구성되어 모델이 변수에 대한 정답을 맞추는 정확도를 높이는 방식으로 학습하는 머신러닝 기법이다.



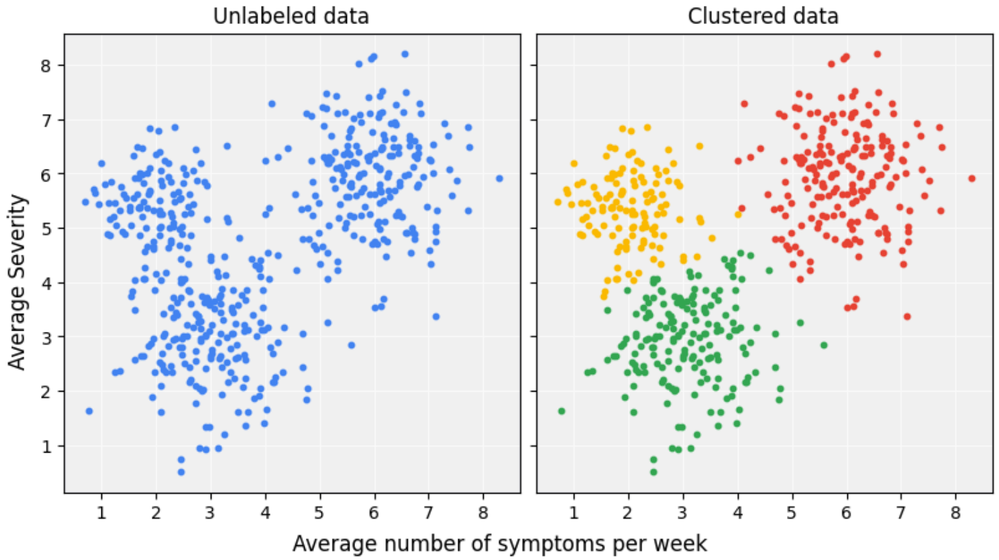
y is categorical y is continous

Given $(\boldsymbol{x}1, y1)$, $(\boldsymbol{x}2, y3)$, ⋯, $(\boldsymbol{x}i, yi)$,
learn a function $f(\boldsymbol{x})$ to predict y
given $\boldsymbol{x}$

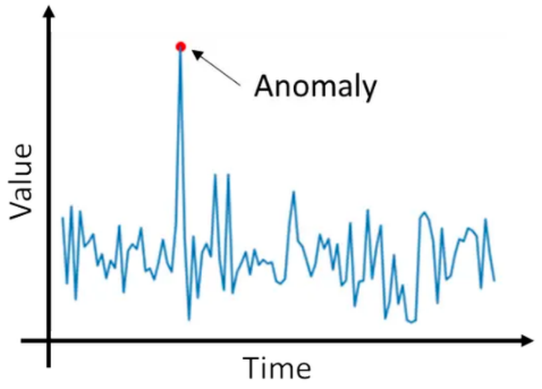
AI명예학회
SKHU

비지도 학습(Unsupervised Learning)

비지도 학습은 훈련 데이터에 원하는 결과를 포함하고 있지 않아 모델이 스스로 학습하며 데이터에 숨겨진 구조를 찾도록 하는 머신러닝 기법이다.



Given $\boldsymbol{x}1$, $\boldsymbol{x}2$, ⋯, $\boldsymbol{x}i$, find a hidden structures



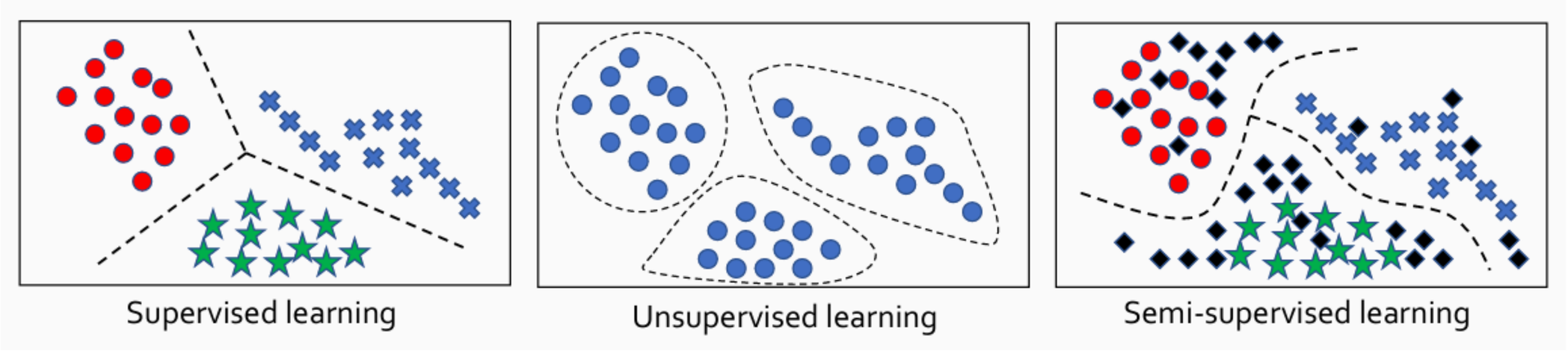
ex) Clustering, Anomaly detection, density estimation

AI명예학회
SKHU

준지도 학습(Semi-supervised Learning)

지도학습에 필요한 데이터는 만들기 어렵다. 작은 숫자의 라벨링 데이터와 큰 숫자의 언라벨링 데이터를 이용해 지도학습 task를 수행한다.

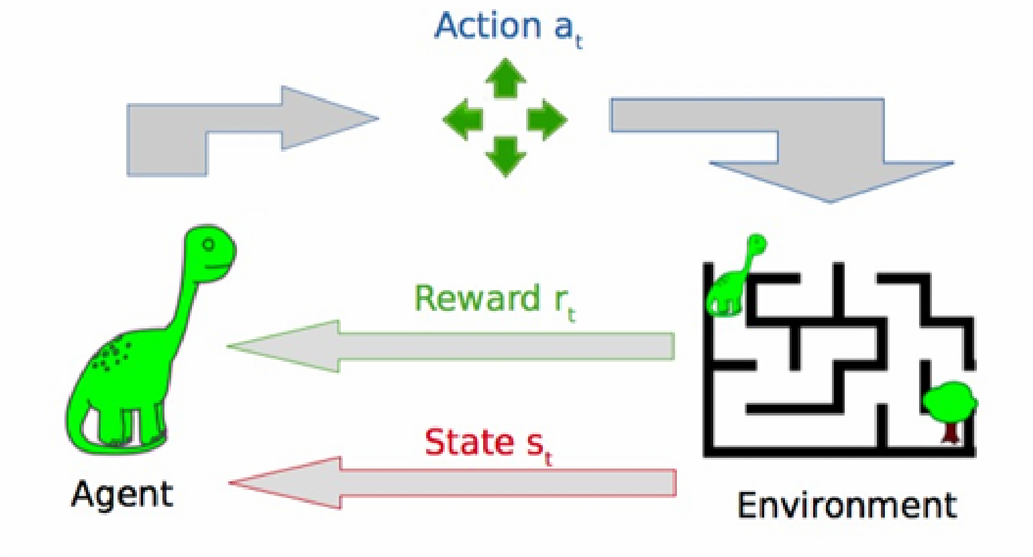
비슷한 데이터는 비슷한 라벨을 가지고 있기 때문에, 결정경계를 만드는데 도움을 준다.

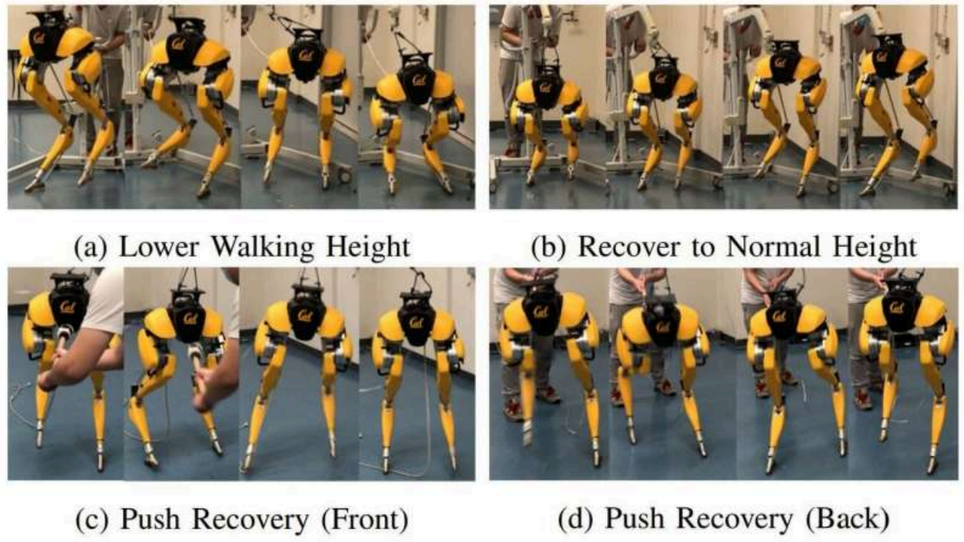


AI명예학회
SKHU

강화 학습(Reinforcement Learning)

에이전트와 환경 사이의 상호작용을 통해 나온 보상을 기반으로 학습한다. 일반적으로 보상의 최대화를 목적으로 하며, 에이전트가 선택하는 행동에 따라 환경이 변화한다 . 피드백이 즉각적이지 않기에 설계하기 까다롭다.




(a) Lower Walking Height (b) Recover to Normal Height

(c) Push Recovery (Front) (d) Push Recovery (Back)

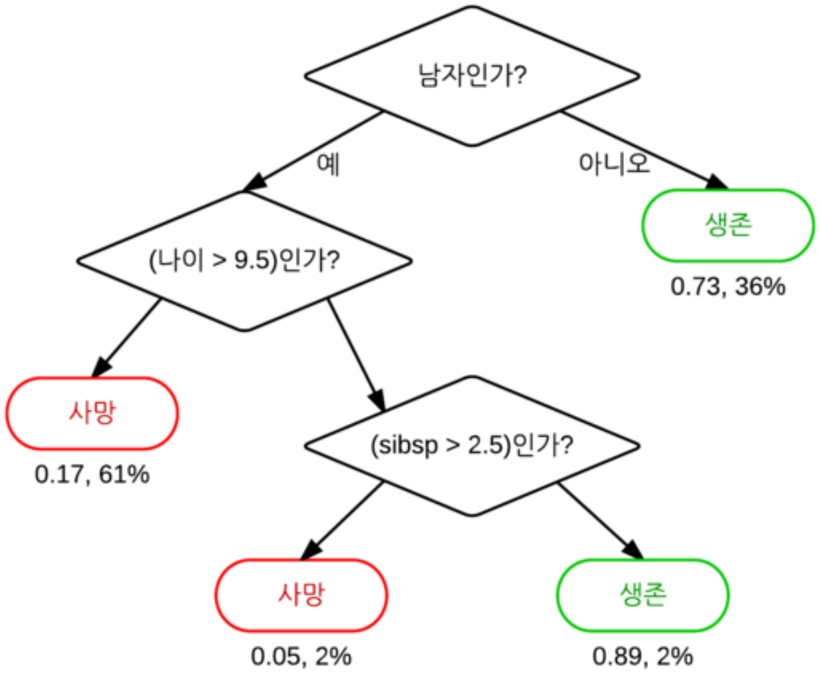
AI명예학회
SKHU

어떤 모델이 좋은 모델인가?

AI명예학회
SKHU

모델의 설명력

실업에서 인공지능을 사용할 때 중요한 요소로 보는 것 중 하나가 설명력이다.
아무리 예측이 정확한 모델이라도, 왜 그런 예측이 나오는지 설명하는 것이 불가능하다면 실제로 그 결정을 기반으로 판단할 때 신뢰도가 떨어져 실업에 사용하기 힘들다.
반면, 모델의 결정 과정은 중요하지 않고
결과만 중요한 작업이라면, 설명력이 떨어지는 모델을 사용해도 무방할 것이다.



ChatGPT

AI명예학회
SKHU

일반화 (Generalization)

학습이란 특정한 인스턴스들의 집합인 훈련 데이터를 이용해 일반적인 개념이나 주장으로 구성하는 추상화(abstraction)의 일종

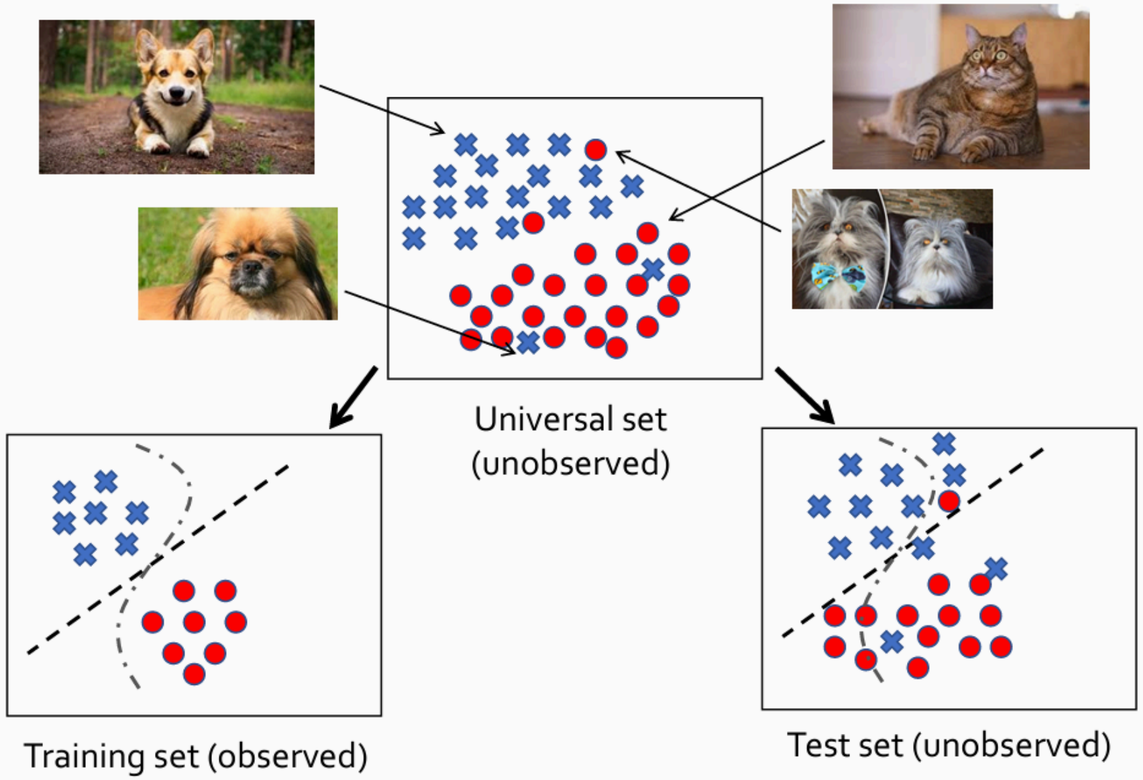
우리는 한 번도 본 적 없는 나무를 보더라도, 그것이 나무라고 분류할 수 있다. 지금까지 봐온 나무들의 일반적인 특징을 추출해 나무라는 개념을 정의했기 때문이다.

학습의 목적은 학습하는 데이터 자체에 대한 정확한 표현이 아니라 학습에 사용되지 않은 데이터도 처리할 수 있는 프로세스를 지니도록 모델을 업데이트하는 것이다.

AI명예학회
SKHU

머신러닝에서의 일반화

학습 과정에 사용된 데이터보다 알 수 없는 데이터(unseen data)에 더 좋은 성능을 보여줘야 한다.
즉, 머신러닝 모델의 **일반화 성능**이란
학습하지 않았던 새로운 데이터들에 대해
잘 예측하는 능력이다.
하지만 훈련 알고리즘은 훈련 데이터에 대한 정확도를 최대화하도록 설계되어 있다.
이 문제를 어떻게 해결할 수 있을까?
일반적으로 훈련 데이터의 일부로 테스트 데이터로 만들어 unseen data로 기능하게 한다.



AI명예학회
SKHU

Generalization Error

다르게 표현하자면, 우리의 목표는 모든 데이터에 대한 오차인 **Generalizaiton Error**를 최소화하는 것이다.

그러나 상술했듯이 전체 데이터는 우리가 알 수 없다.
훈련 과정에서 사용하지 않은 테스트 데이터에 대한 error를 통해 간접적으로 평가한다.

train error와 generalization error(실전에서는 test error)를 비교해 모델이 어떻게 학습되었는지 진단하는 방법을 알아보자.

과대적합과 과소적합

과소적합(Under fitting): train error > generalization error

- train error가 너무 큰 상태를 의미한다.
- 즉, 모델이 훈련 데이터에 대해서도 제대로 학습하지 못한 것이다.

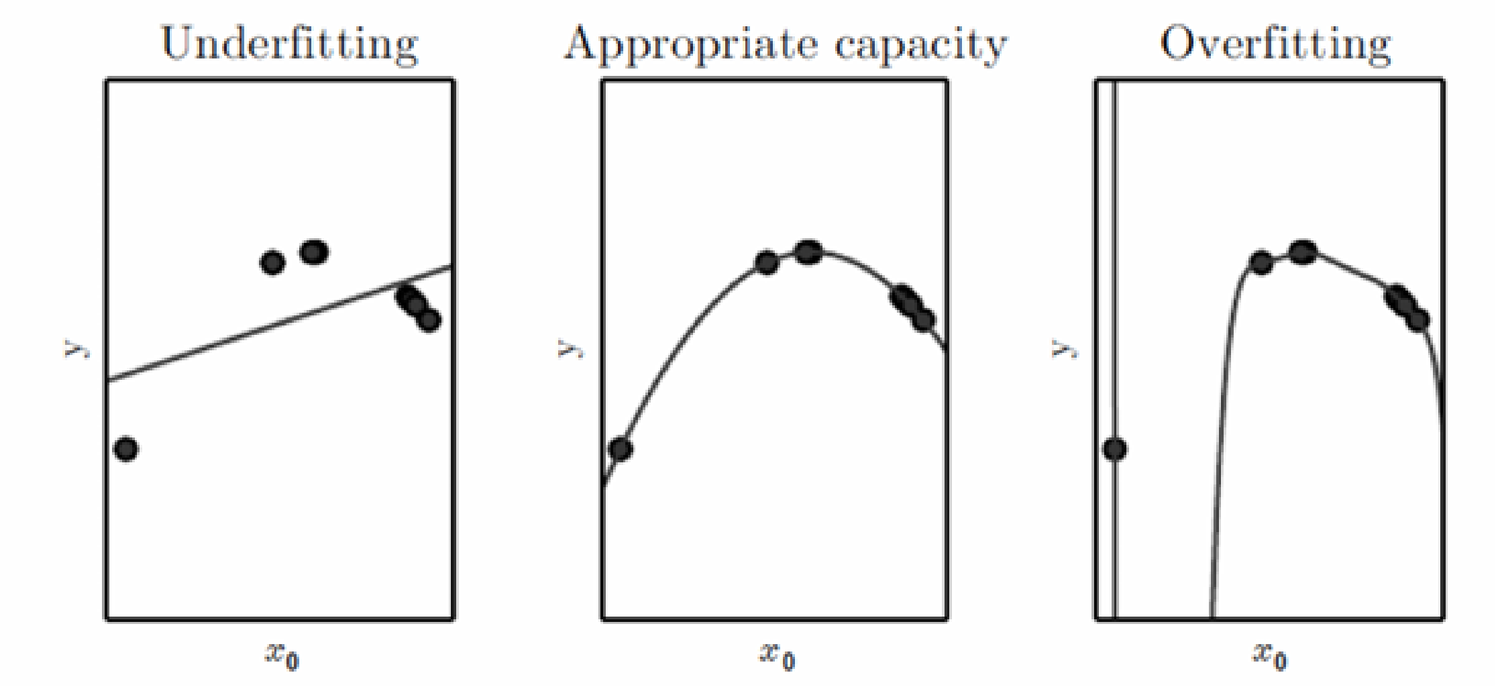
과대적합(Over fitting): train error < generalization error

- generalization error가 train error보다 훨씬 큰 상태를 의미한다.
- 모델이 일반적인 특성을 학습한 것이 아니라 개별 데이터 포인트의 오차를 줄이는 것에 집중해 학습했다.
- 즉, 훈련 데이터를 과도하게 학습했다.

AI명예학회
SKHU

모델의 수용력(capacity)와 과대적합의 관계

모델의 수용력은 학습 가능한 파라미터의 수라고도 할 수 있다. 일반적으로 수용력이 클수록 학습 데이터에 대한 오차를 낮추는 것이 유리하다. 하지만, 그만큼 과대적합의 위험도 크다.

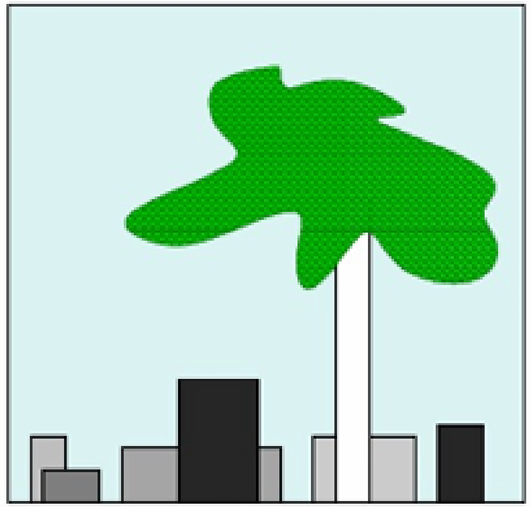


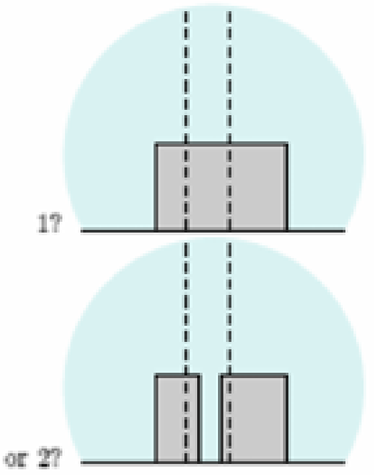
AI명예학회
SKHU

모델의 수용력(capacity)와 과대적합의 관계

만약 train과 test에서 같은 성능을 지녔다면, 어떤 모델을 선택하는 것이 좋을까?
정답은 더 단순한 모델을 고르는 것이다.

가장 단순한 설명이 모든 경우에 더 잘 적용될 확률이 크다.
설명이 더 복잡하고 요소가 많을수록 일반화되기 어려울 것이다.



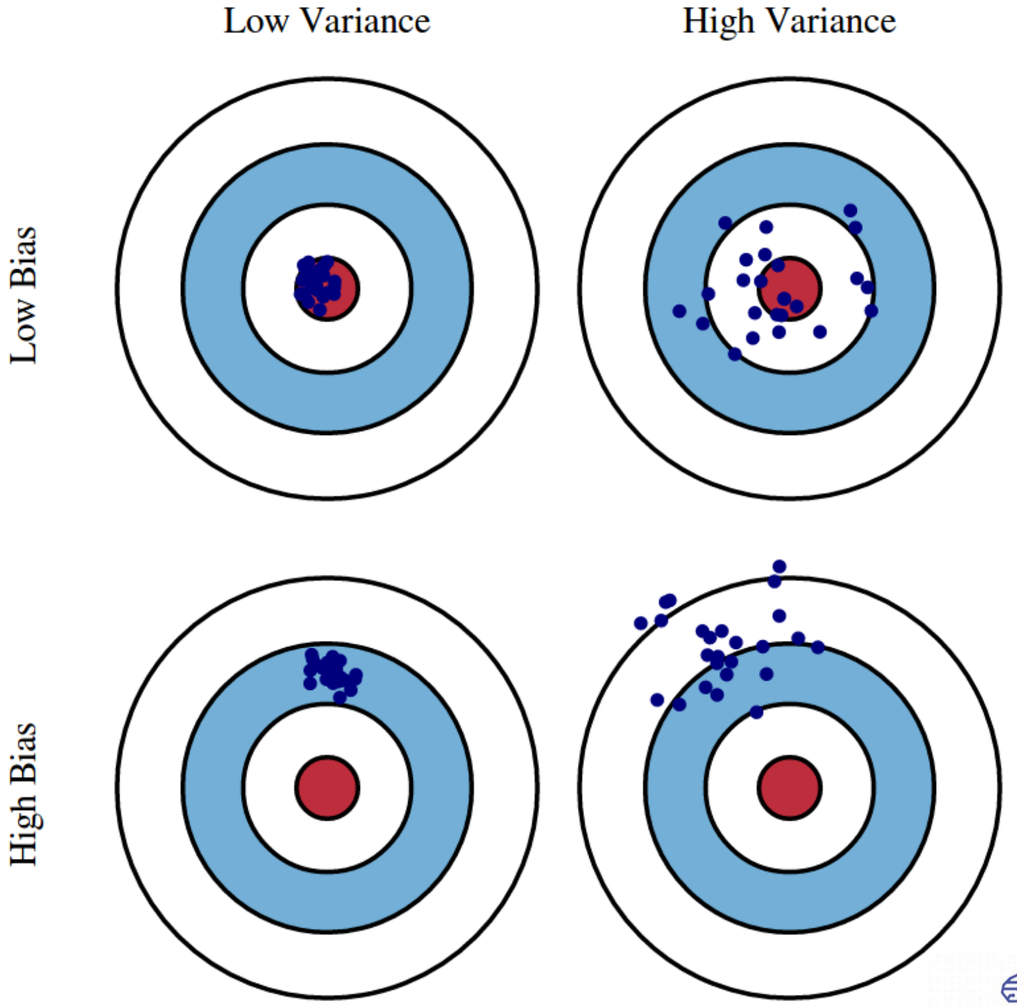


AI명예학회
SKHU

Bias-Variance Trade-off

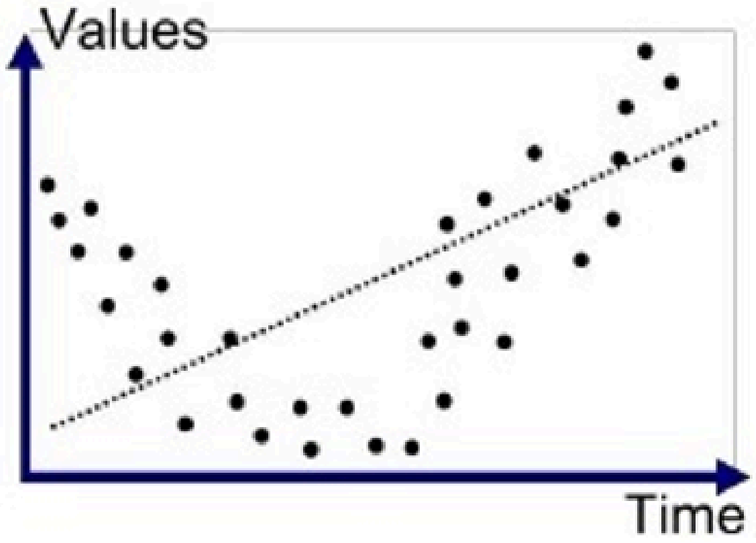
Bias (편향): 예측 값과 실제 값의 편차
Variance (분산): 서로 다른 샘플 데이터 사이의 예측 값의 편차, 즉 분산

High Bias: 과소적합 상태
Low Bias High Variance: 과대적합 상태
Low Bias Low Variance: 최적점



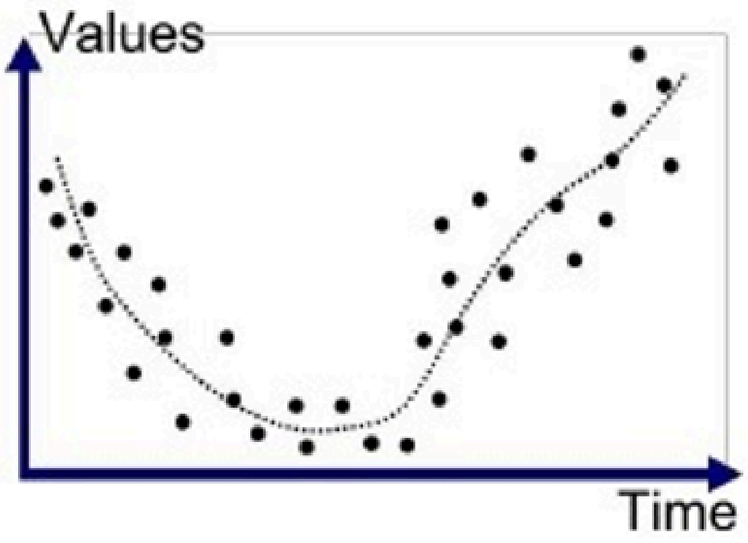
AI명예학회
SKHU

High Bias Low Variance



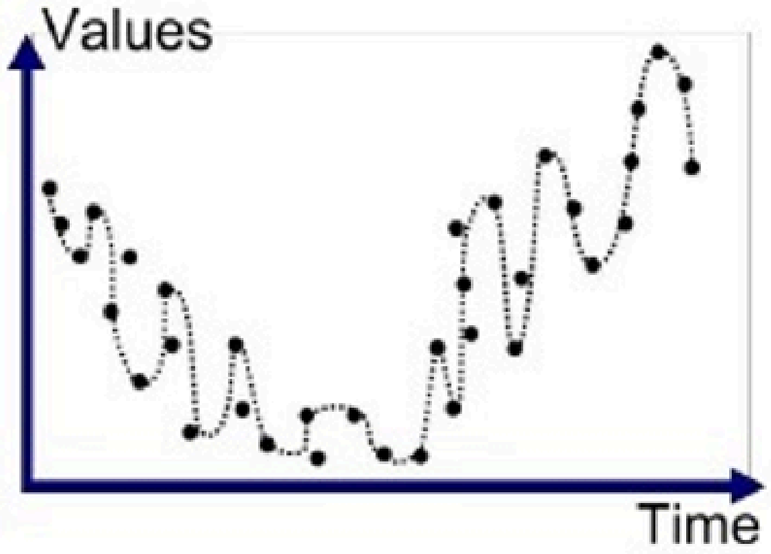
Underfitted

Low Bias Low Variance



Good Fit/Robust

Low Bias High Variance



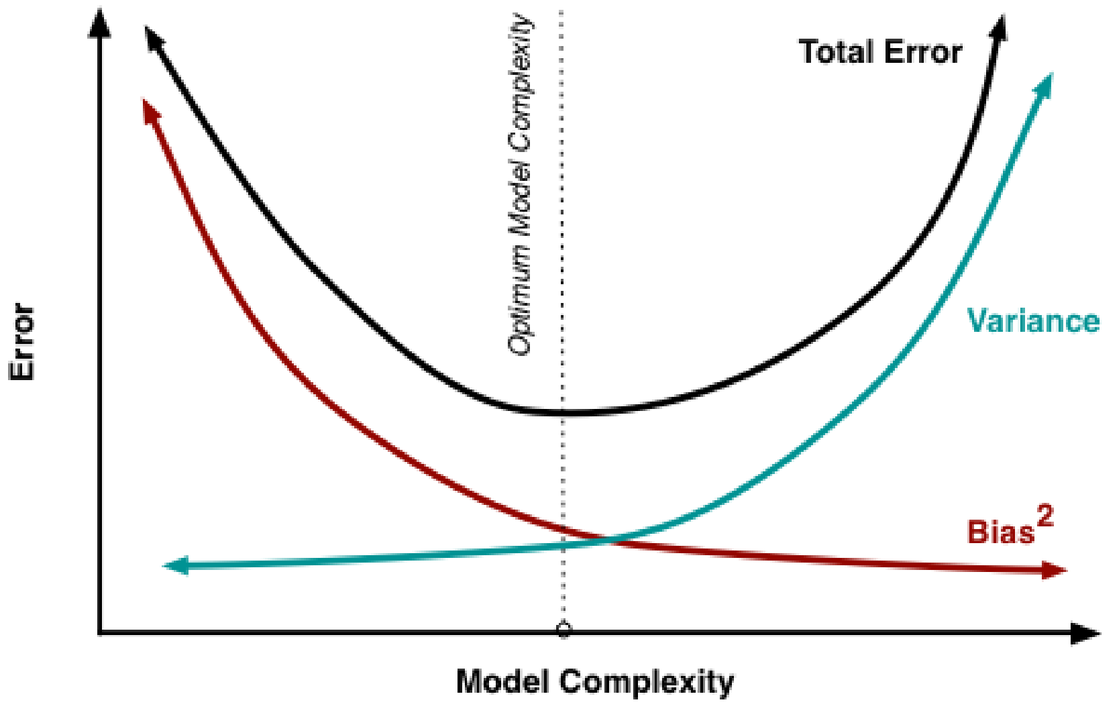
Overfitted

AI명예학회
SKHU

Bias-Variance Trade-off

test error = Bias + Variance

좋은 모델을 만들기 위해서는 이 둘 모두를 낮춰야 한다.
문제는 이 둘 사이의 trade-off 관계에 있다.
모델의 수용력이 올라갈수록 Bias는 낮아지지만, Variance는 높아진다.
반대로 모델의 수용력이 낮아질수록 Variance는 낮아지지만, Bias는 높아진다.



AI명예학회
SKHU

모든 문제에 최적인 모델은 없다

모든 문제에 대해 최적의 해를 구하는 머신러닝 알고리즘은 없다.

해결하려는 문제의 종류, 데이터의 특성과 양, 가용할 수 있는 자원,
해결함으로서 얻고자 하는 것을 종합적으로 판단해
어떤 모델이 해당 문제에 대해 제일 적합할지 선택해야 한다.

정리

머신러닝의 정의와 종류

- 모델 기반 / 사례 기반
- 지도, 비지도, 준지도, 강화

좋은 머신러닝 모델이란?

- 설명력
- 일반화
- 과대적합 / 과소적합
- Bias-Variance Trade-off

AI명예학회
SKHU