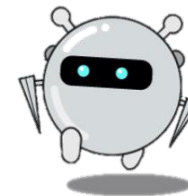
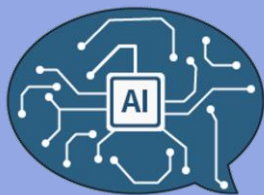


머신러닝 개요



| 목차

01 머신러닝이란

- 정의 / 구조 / 인공지능, 머신러닝, 딥러닝 / 전통적인 프로그래밍과의 비교 / 머신러닝 용어 정리

02 머신러닝의 종류

- 지도학습 / 비지도학습 / 강화학습

03 머신러닝의 단계

- 데이터 수집 -> 데이터 전처리 -> 모델링 및 훈련 -> 모델 평가 -> 모델 배포



01 머신러닝이란?



| 01 머신러닝이란?

**“ 머신러닝은 명시적인 프로그래밍 없이
컴퓨터가 학습하는 능력을 갖추게 하는 연구 분야 ”**

- 아서 새뮤얼 (Arthur Samuel, 1959)



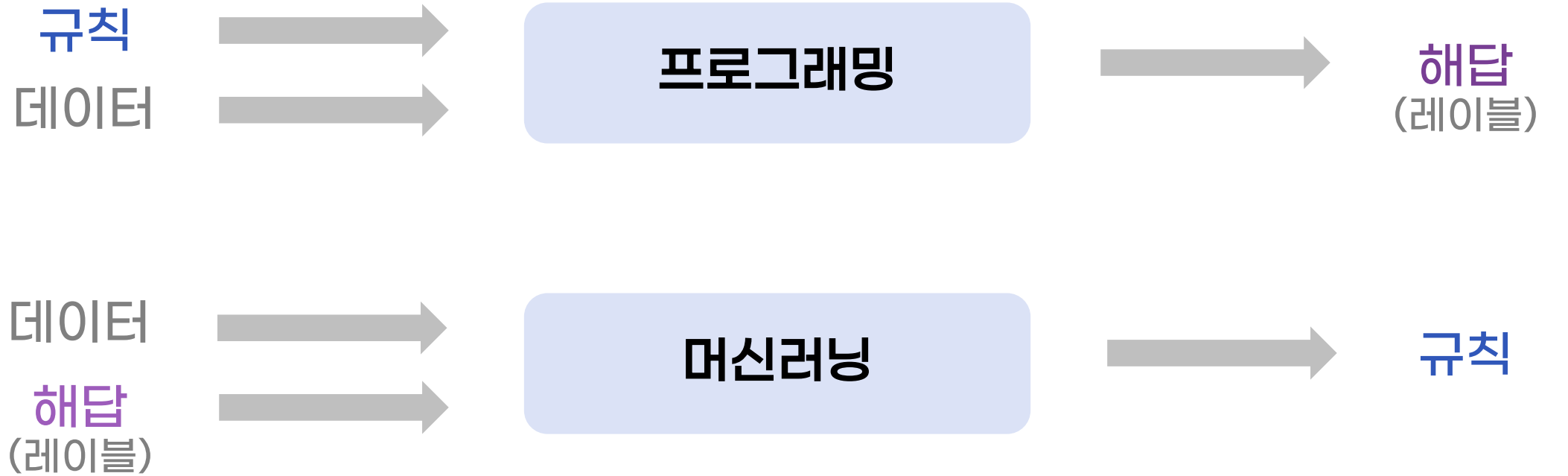
| 01 머신러닝이란?

“어떤 작업 T에 대하여 컴퓨터 프로그램의 성능을 P로 측정했을 때, 경험 E로 인해 성능이 향상되었다면, 이 컴퓨터 프로그램은 작업 T와 성능 측정 P에 대해 경험 E로부터 학습한다고 말한다.”

- 톰 미첼(Tom Mitchell, 1977)



| 01 머신러닝이란?



| 01 머신러닝이란?

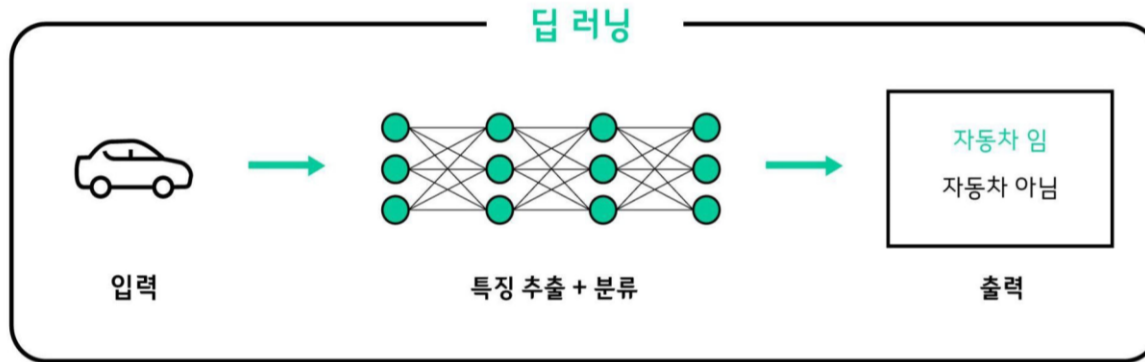


출처 : <https://hyeonjiwon.github.io> [머신러닝의 개요]

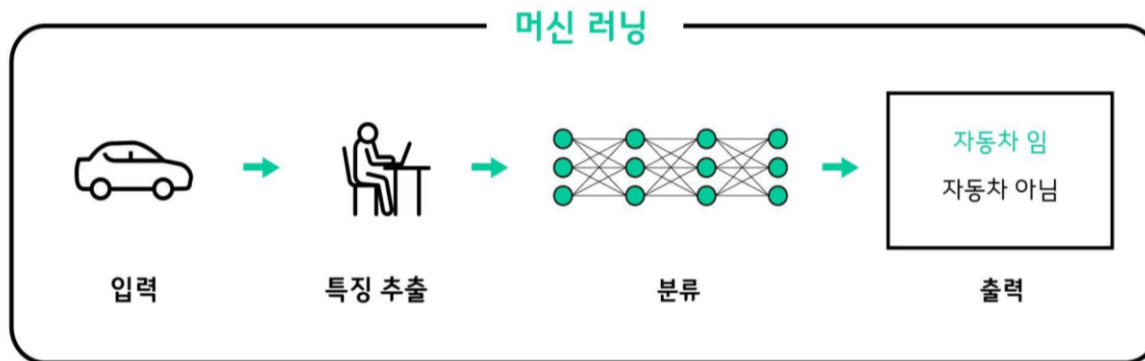


| 01 머신러닝이란?

딥 러닝과 머신 러닝 비교



-> 인공신경망을 통해
데이터 선정까지 스스로



-> 인간이 미리 데이터의
특징을 입력해야함

출처 : <https://www.freshworks.com/ko/freshdesk/kblogs/deep-learning/>



머신러닝 용어 정리



| 머신러닝 용어 정리

열 (column), 특성 (feature), 속성 (attribute), 변수 (variable), field

라벨 (label)

고객번호	성별	나이	재산 (만원)	올해 구매액 (만원)	이탈 여부
EA1243	남	21	300	520	X
EA1244	여	24	500	1120	O
EA1245	남	46	24000	5000	X
EA1246	여	61	11000	1500	X

행 (row),
개체 (instance),
관측치
기록 (record)
사례 (example)
경우 (case)



| 머신러닝 용어 정리

1. 파라미터 (Parameter)

- 머신러닝 훈련 모델에 의해 요구되는 변수
- 머신러닝 훈련 모델의 성능은 파라미터에 의해 결정됨
- 파라미터는 데이터로부터 추정 또는 학습됨
- 파라미터는 개발자에 의해 수동으로 설정하지 않는다.
- 학습된 모델의 일부로 저장된다.
- 예) 인공신경망의 가중치, SVM의 서포트 벡터, 선형 회귀에의 결정계수

2. 하이퍼파라미터 (Hyperparameter)

- 최적의 훈련 모델을 구현하기 위해 모델에 설정하는 변수
- 개발자에 의해 수동으로 설정할 수 있다.
- 학습 알고리즘의 샘플에 대한 일반화를 위해 조절된다.
- 예) 학습률, 손실함수, k-NN의 k값, 은닉층의 개수, epoch 수



02 머신러닝의 종류



| 02 머신러닝의 종류

머신러닝

지도학습 Supervised Learning

- 레이블된 데이터로 학습
- 출력 및 미래 예측

분류
Classification

회귀
Regression

비지도학습 Unsupervised Learning

- 레이블되지 않은 데이터로 학습
- 데이터에서 숨겨진 구조 찾기

군집화
Clustering

차원 축소
Dimensionality
Reduction

강화학습 Reinforcement Learning

- 특정 목표를 위해 최선의 전략을 선택하도록 학습
- 피드백이나 보상을 통해 학습



| 02 머신러닝의 종류 - 지도학습

분류 (Classification)

- 데이터를 기반으로 새로운 샘플의 범주형 클래스 레이블 예측이 목표
- 이진 분류(binary classification): 클래스 레이블 종류가 두 개인 경우
- 다중 클래스 분류 (multi-class classification): 클래스 레이블 종류가 세 개 이상

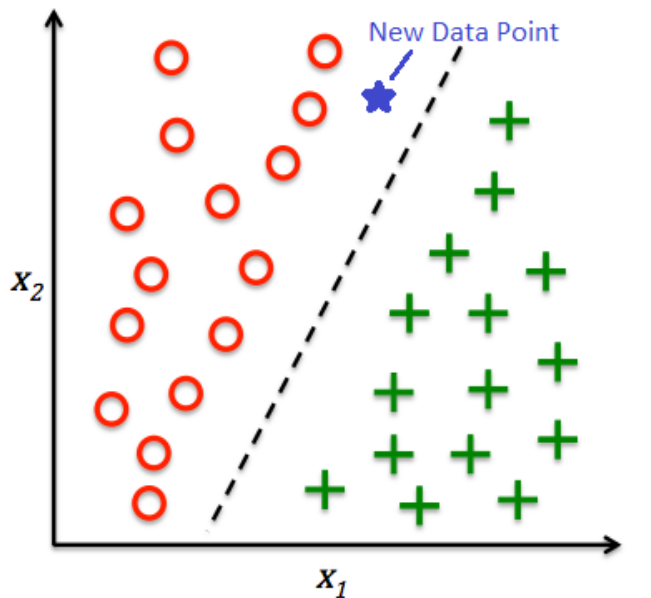
회귀 (Regression)

- 연속적인 출력 값에 대한 예측이 목표
- 예) 시험 점수 예측,
주식 가격 예측,
부동산 가격 예측



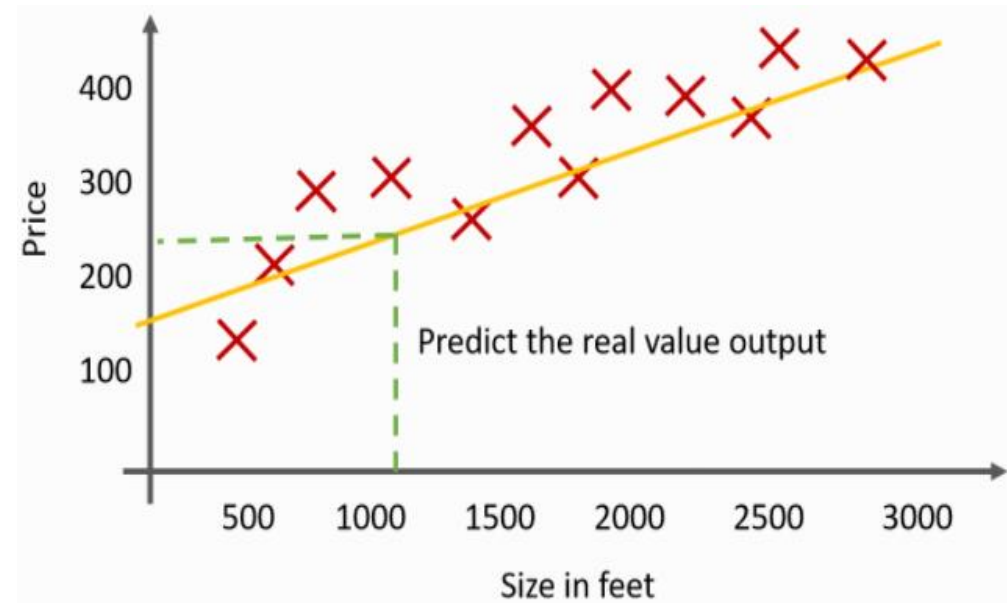
| 02 머신러닝의 종류 - 지도학습

분류 (Classification)



문서 분류, 이미지 분류

회귀 (Regression)



주식 가격, 부동산 가격 예측



출처 : <https://velog.io/@dohyunkyoung>

지도학습의 이해



| 지도학습의 이해 - Train Set, Test Set

환자번호	기침	발열	미각상실	복통	오한	두통	기타	양성 여부
EA1243	1	1	1	1	1	1	0	1
EA1244	1	0	0	0	1	0	1	0
EA1245	1	0	0	0	0	1	1	1
EA1246	1	0	0	1	1	1	0	0
EA1247	1	1	1	1	1	1	0	1



| 지도학습의 이해 - Train Set, Test Set

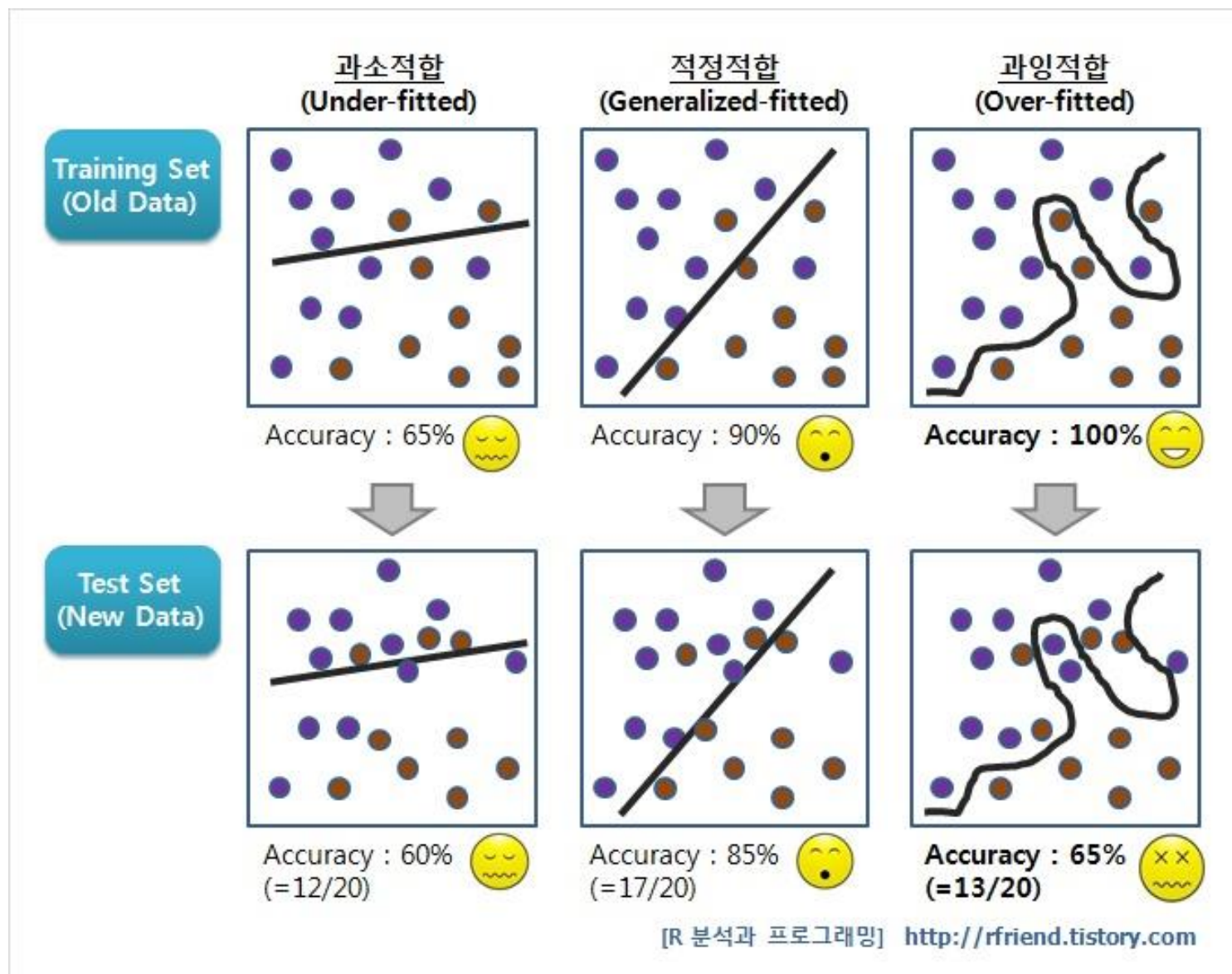
Train Set

환자번호	기침	발열	미각상실	복통	오한	두통	기타	양성 여부
EA1243	1	1	1	1	1	1	0	1
EA1244	1	0	0	0	1	0	1	0
EA1245	1	0	0	0	0	1	1	1
EA1246	1	0	0	1	1	1	0	0

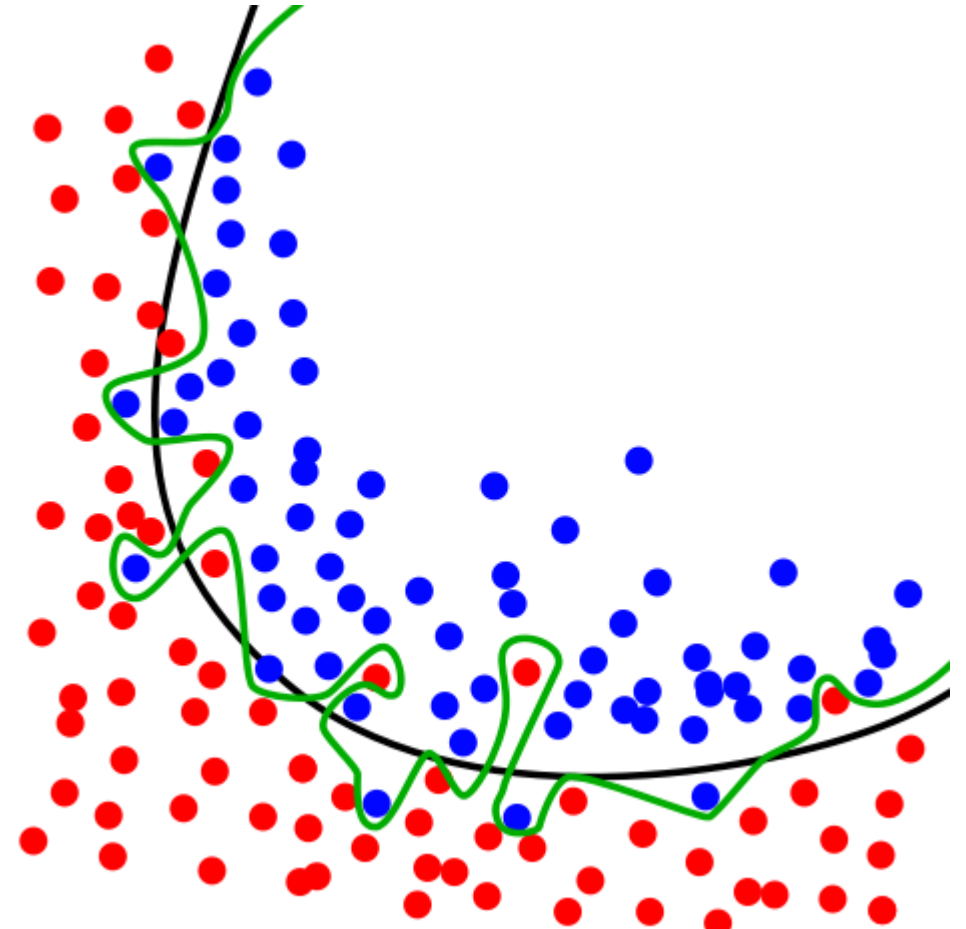
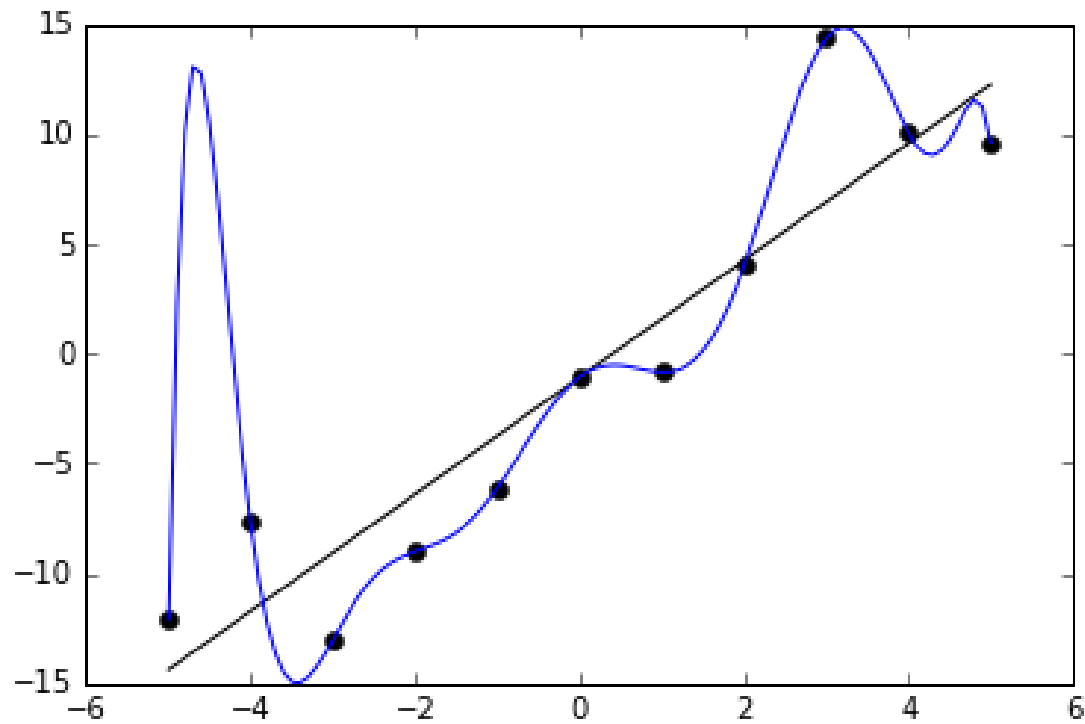
Test Set

환자번호	기침	발열	미각상실	복통	오한	두통	기타	양성 여부
EA1247	1	1	1	1	1	1	0	 1

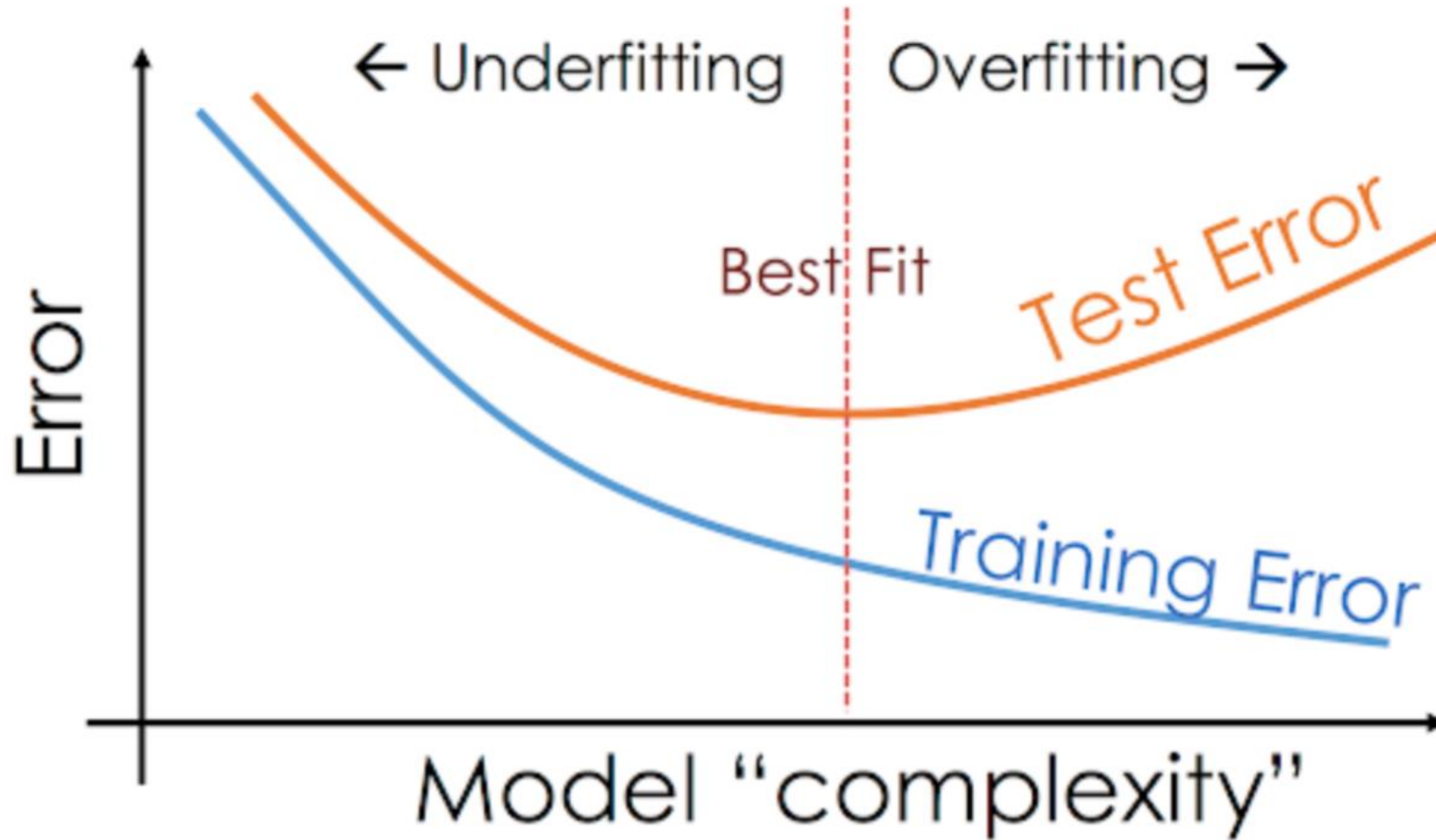
| 지도학습의 이해 - Overfitting & Underfitting



| 지도학습의 이해 - Overfitting & Underfitting



| 지도학습의 이해 - Overfitting & Underfitting



<https://vitalflux.com/overfitting-underfitting-concepts-interview-questions/>



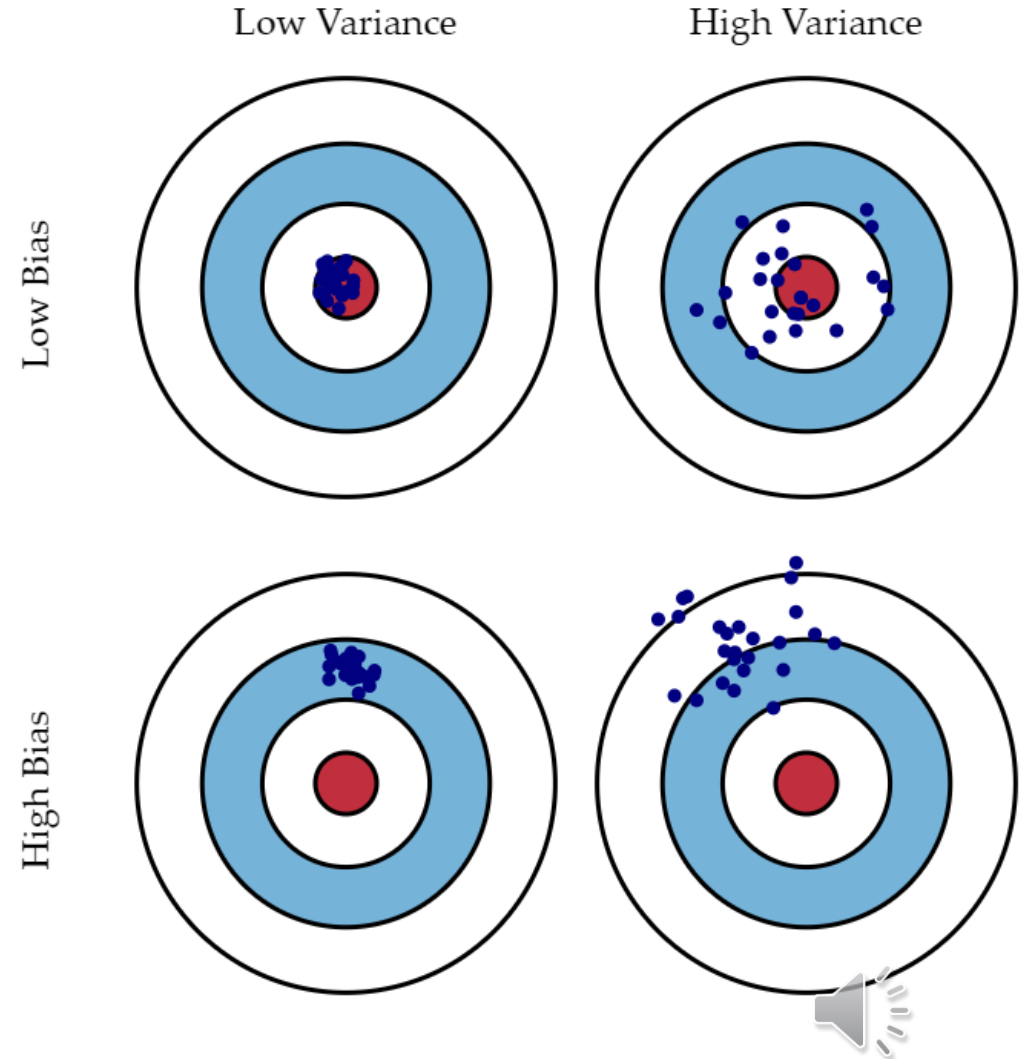
| 지도학습의 이해 - 편향과 분산 Trade-off

편향 (Bias)

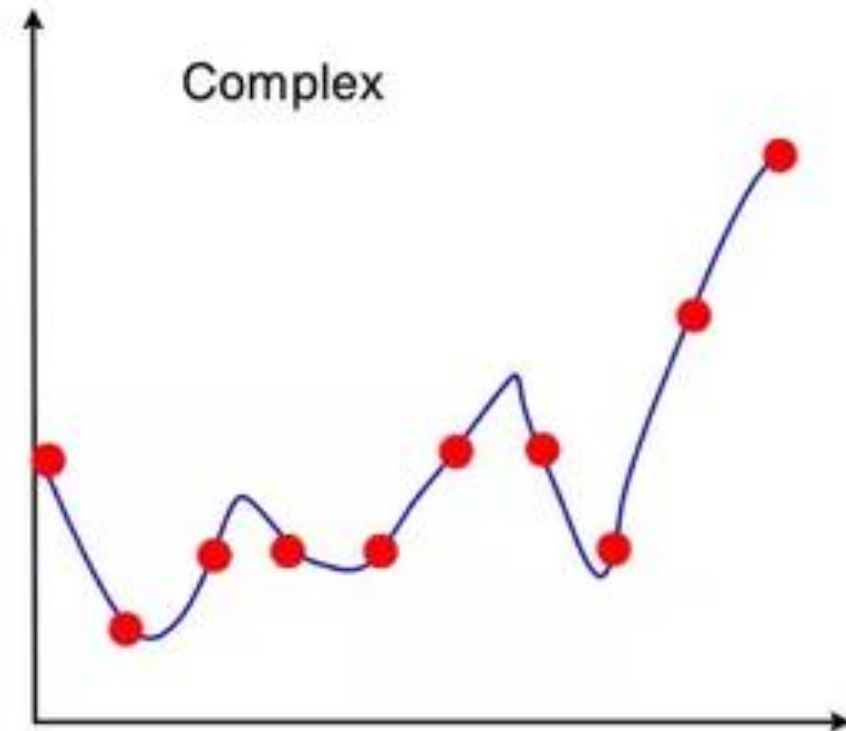
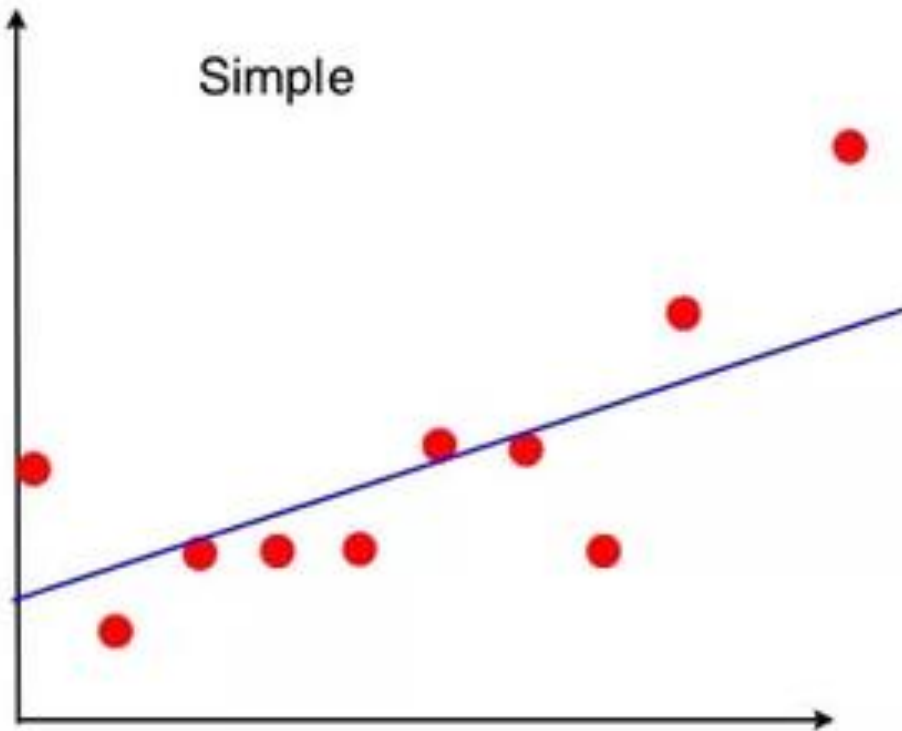
- 지나치게 단순한 모델로 인한 오류
- 편향이 크면 과소적합을 야기함.

분산 (Variance)

- 지나치게 복잡한 모델로 인한 오류
- 분산이 크면 과대적합을 야기함.

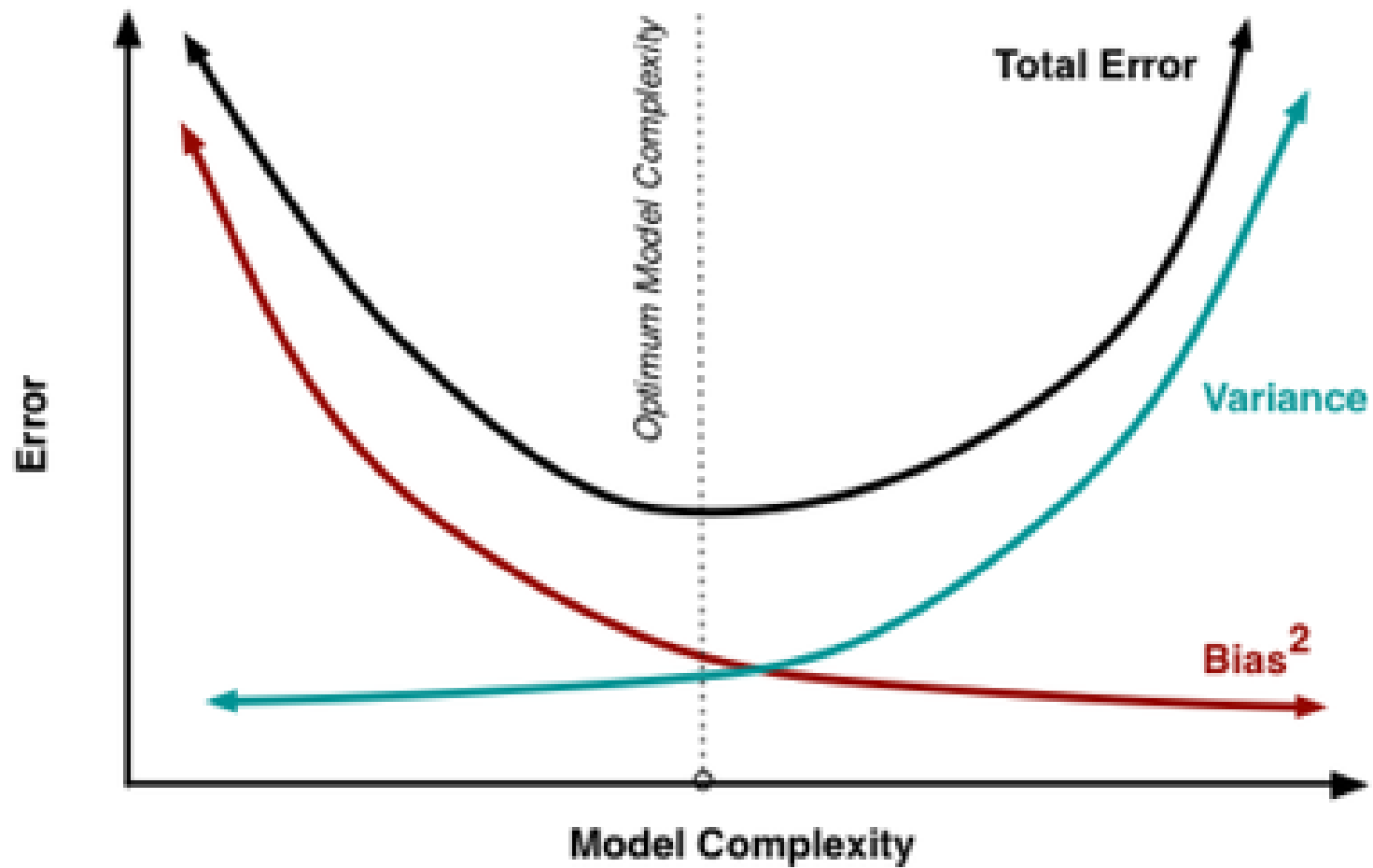


| 지도학습의 이해 - 편향과 분산 Trade-off



<https://bkshin.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%88%9D-12-%ED%8E%B8%ED%96%A5Bias%EC%99%80-%EB%B6%84%EC%82%B0Variance-Trade-off>


| 지도학습의 이해 - 편향과 분산 Trade-off



| 지도학습의 이해 - 편향과 분산 Trade-off

$$Y = f(X) + \varepsilon \quad \text{이라고 하면}$$

$$\underbrace{E[(Y - \hat{f}(X))^2]}_{\text{Expected mean-squared error (MSE) on the validation sample (for prediction modeling)}} = \underbrace{E[(\hat{f}(X) - E[\hat{f}(X)])^2]}_{\text{Variance of the fit}} + \underbrace{(E[\hat{f}(X)] - f(X))^2}_{\text{(Bias of the fit)}^2} + \underbrace{\sigma^2}_{\text{Variance of the error (noise)}}$$

 *Trade-off (Dilemma)*

* 참고 : Y는 새로운 관측값(from validation set)임. $\hat{f}(X)$ 는 이전 Y값(from training set)으로부터 계산한 값임.
따라서 Y와 $\hat{f}(X)$ 는 서로 상관성이 없음

* 주의 : (Bias of the fit) 이 아니라 (Bias of the fit)² 임



| 지도학습의 이해 - Validation Set

Train Set

환자번호	기침	발열	미각상실	복통	오한	두통	기타	양성 여부
EA1243	1	1	1	1	1	1	0	1
EA1244	1	0	0	0	1	0	1	0
EA1245	1	0	0	0	0	1	1	1

Validation Set

환자번호	기침	발열	미각상실	복통	오한	두통	기타	양성 여부
EA1246	1	0	0	1	1	1	0	0

Test Set

환자번호	기침	발열	미각상실	복통	오한	두통	기타	양성 여부
EA1247	1	1	1	1	1	1	0	1



| 02 머신러닝의 종류 - 비지도학습

군집화 (Clustering)

- 레이블이 없는 데이터를 분류
- 비지도 분류라고도 함
- 같은 그룹 내에서는 유사성 가짐
- 다른 그룹 간에는 다른 성질을 가짐

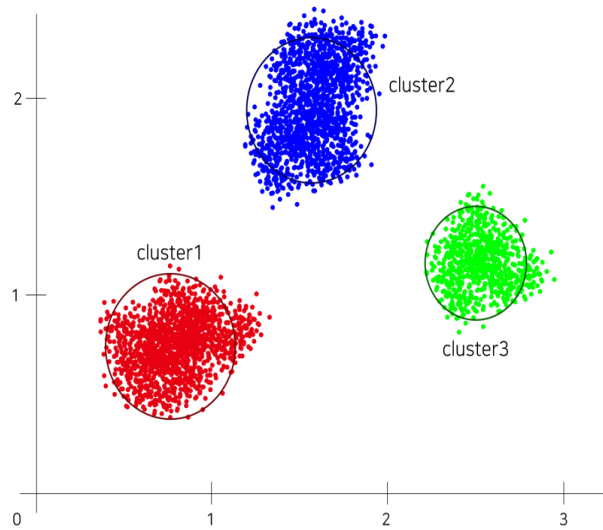
차원 축소 (Dimensionality Reduction)

- 고차원 데이터(변수가 많은 데이터)를 중요한 정보는 유지하면서 더 작은 차원으로 압축
- 저장 공간을 줄이고 잡음(noise) 데이터를 제거해 알고리즘의 예측 성능을 높이는 것이 목표



| 02 머신러닝의 종류 - 비지도학습

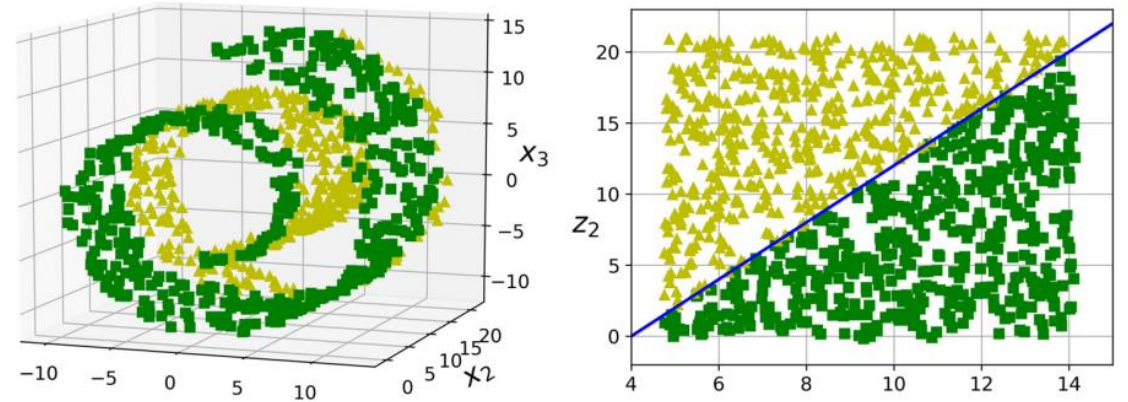
군집화 (Clustering)



출처 : <https://muzukphysics.tistory.com/entry>

고객 분류, 유사 단어 및 이미지 군집화

차원 축소 (Dimensionality Reduction)



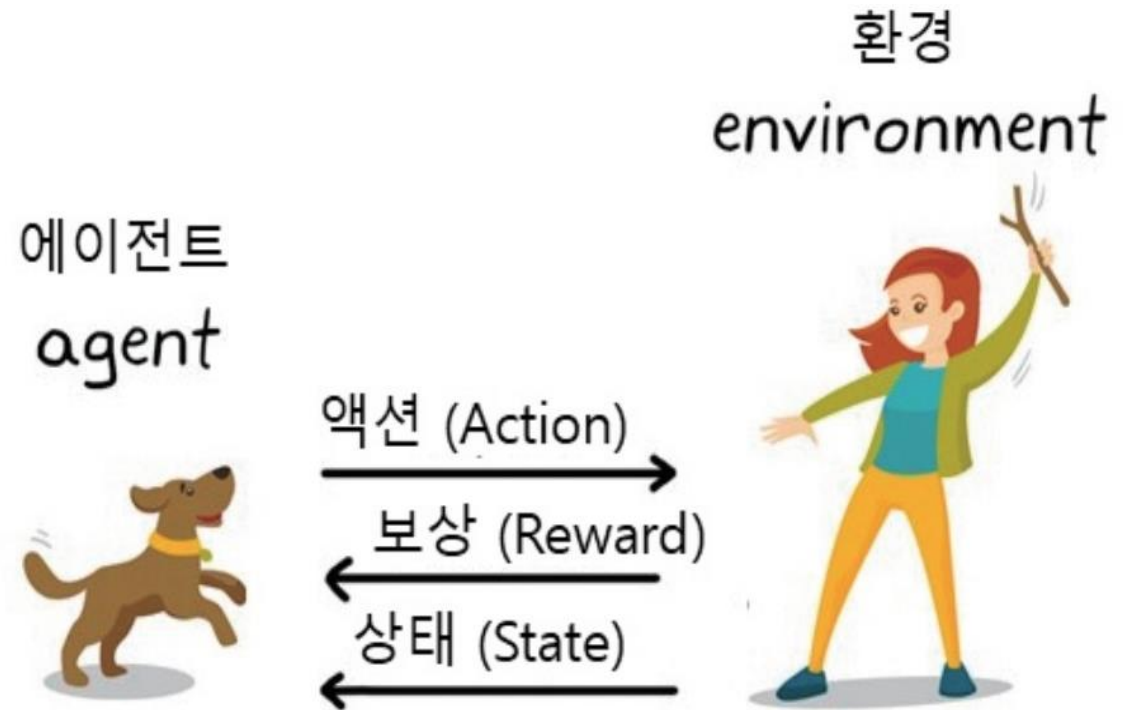
출처 : <https://velog.io/@chaelm>

과적합 해결, 모델 성능 향상



| 02 머신러닝의 종류 - 강화 학습

- 환경으로부터의 피드백을 기반으로 행위자 (agent)의 행동을 분석하고 최적화하여 목표를 달성할 수 있도록 학습을 진행
- 주어진 환경에서 행위자가 최대의 보상을 얻기 위해 시행 착오를 통해 최선의 전략을 스스로 학습
- 예) 체스 인공지능, 알파고



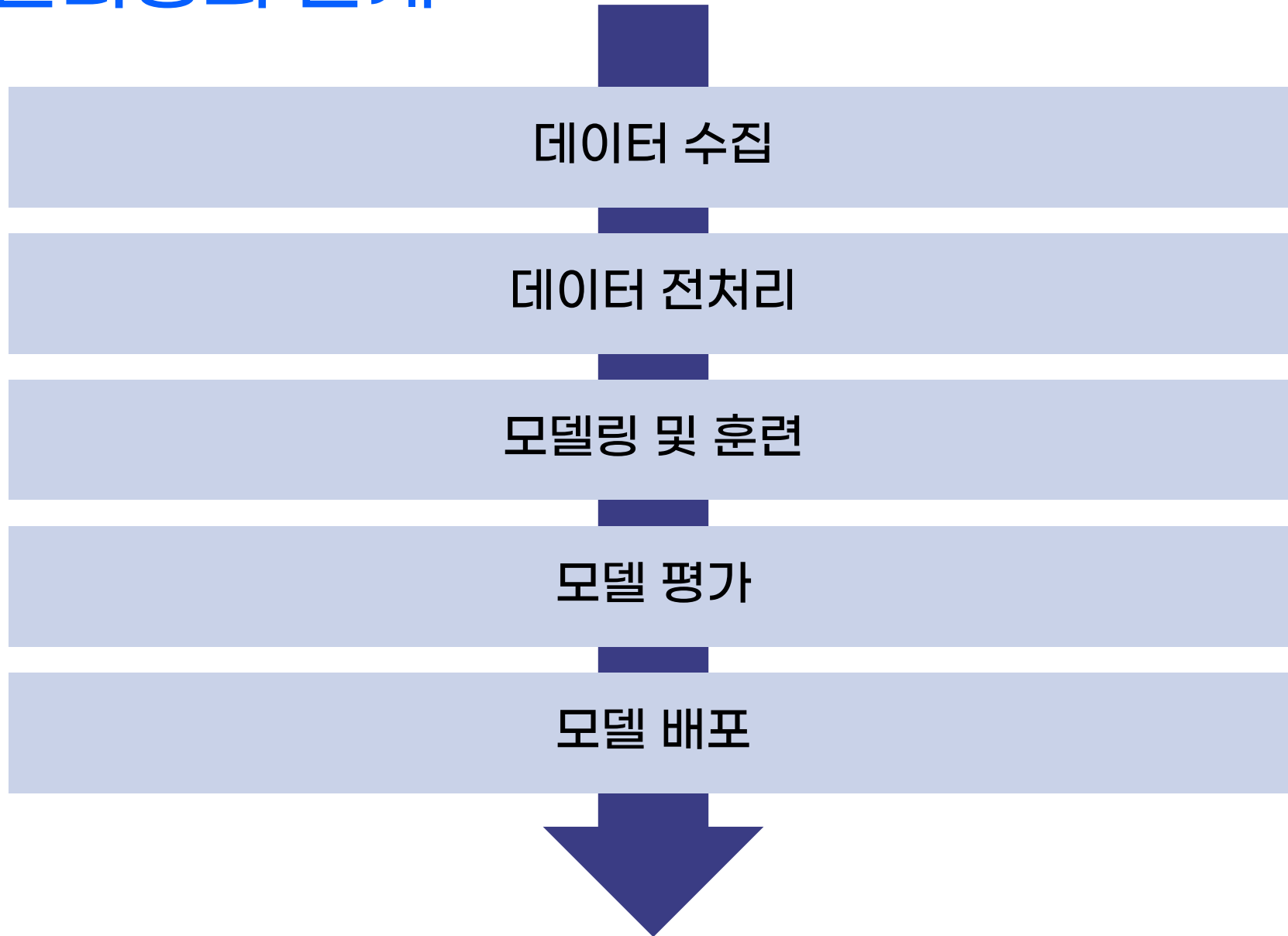
출처 : <https://www.aitimes.com/news/articleView.html?idxno=136181>



03 머신러닝의 단계



| 03 머신러닝의 단계



| 03 머신러닝의 단계 - 1) 데이터 수집

- 머신러닝 분야에서 필수적인 단계

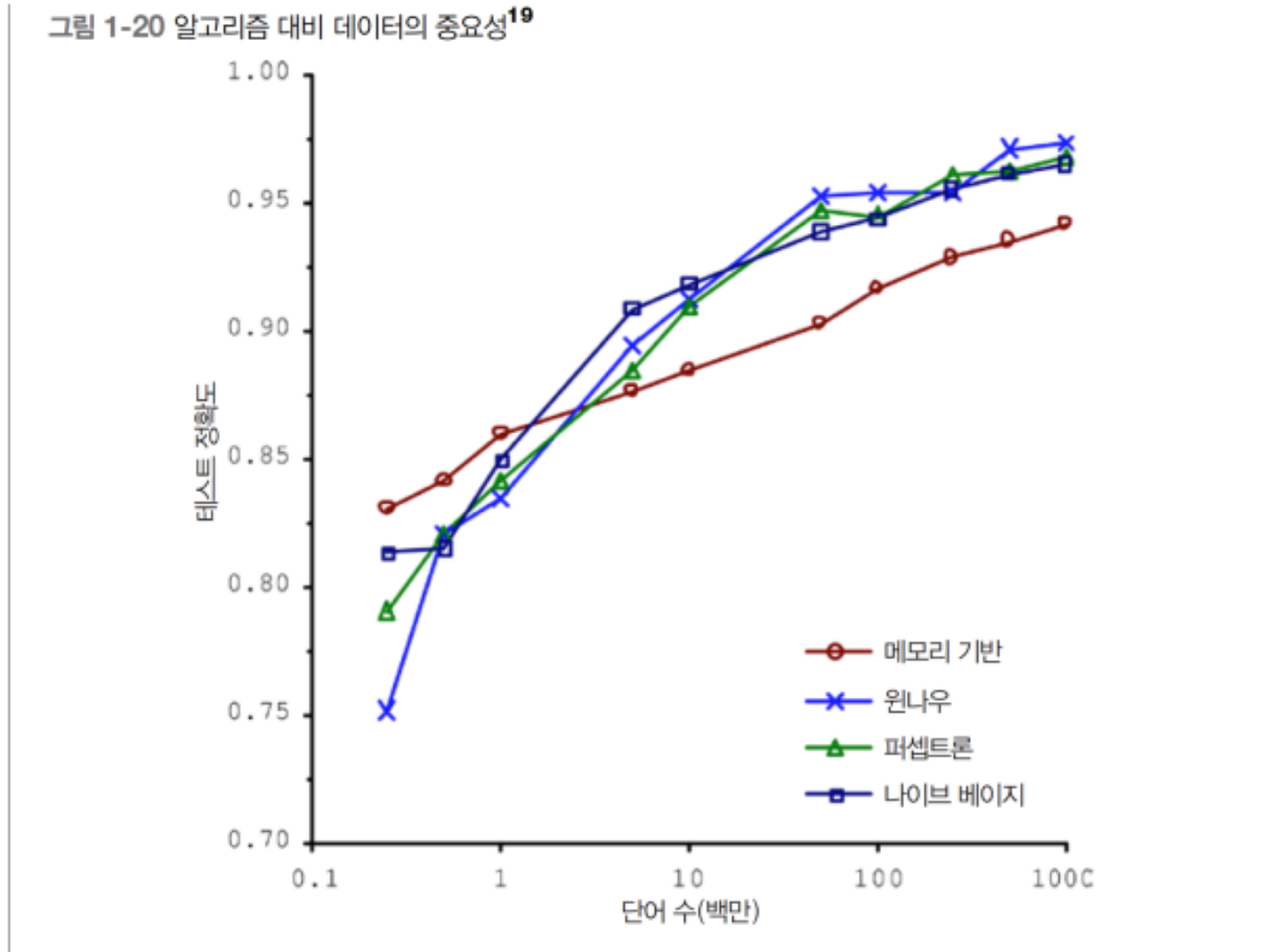
- 데이터의 품질이 학습된 모델의 품질을 결정

 - > 많고 다양한 데이터를 수집해야 함

- 데이터의 형태에 따라 다양한 방법으로 수집 가능



| 03 머신러닝의 단계 - 1) 데이터 수집



| 03 머신러닝의 단계 - 1) 데이터 수집

대표성 있는 훈련 데이터

타겟 집단을 대표해야 함

샘플링 편향 (샘플 수 크더라도 발생 가능)

The Literary Digest

NEW YORK OCTOBER 31, 1936

Topics of the day

LANDON, 1,293,669; ROOSEVELT, 972,897

Final Returns in The Digest's Poll of Ten Million Voters

Well, the great battle of the ballots in the Poll of ten million voters, scattered throughout the forty-eight States of the Union, is now finished, and in the table below we record the figures received up to the hour of going to press.

These figures are exactly as received from more than one in every five voters polled in our country—they are neither weighted, adjusted nor interpreted.

Never before in an experience covering more than a quarter of a century in taking polls have we received so many different varieties of criticism—praise from many; condemnation from many others—and yet it has been just of the same type that has come to us every time a Poll has been taken in all these years.

A telegram from a newspaper in California asks: "Is it true that Mr. Hearst has purchased THE LITERARY DIGEST?" A telephone message only the day before these lines were written: "Has the Repub-

lican National Committee purchased THE LITERARY DIGEST?" And all types and varieties, including: "Have the Jews purchased THE LITERARY DIGEST?" "Is the Pope of Rome a stockholder of THE LITERARY DIGEST?" And so it goes—all equally absurd and amusing. We could add more to this list, and yet all of these questions in recent days are but repetitions of what we have been experiencing all down the years from the very first Poll.

Problem—Now, are the figures in this Poll correct? In answer to this question we will simply refer to a telegram we sent to a young man in Massachusetts the other day in answer to his challenge to us to wager \$100,000 on the accuracy of our Poll. We wired him as follows:

"For nearly a quarter century, we have been taking Polls of the voters in the forty-eight States, and especially in Presidential years, and we have always merely mailed the ballots, counted and recorded those

returned and let the people of the Nation draw their conclusions as to our accuracy. So far, we have been right in every Poll. Will we be right in the current Poll? That, as Mrs. Roosevelt said concerning the President's reelection, is in the 'lap of the gods.'

"We never make any claims before election but we respectfully refer you to the opinion of one of the most quoted citizens to-day, the Hon. James A. Farley, Chairman of the Democratic National Committee. This is what Mr. Farley said October 14, 1932:

"Any sane person can not escape the implication of such a gigantic sampling of popular opinion as is embraced in THE LITERARY DIGEST straw vote. I consider this conclusive evidence as to the desire of the people of this country for a change in the National Government. THE LITERARY DIGEST poll is an achievement of no little magnitude. It is a Poll fairly and correctly conducted."

In studying the table of the voters from

The statistics and the material in this article are the property of Funk & Wagnalls Company and have been copyrighted by it; neither the whole nor any part thereof may be reprinted or published without the special permission of the copyright owner.

| 03 머신러닝의 단계 - 2) 데이터 전처리

- 데이터를 머신러닝의 형태로 변환하는 역할

-> 대부분의 머신러닝 모델은 숫자 데이터를 입력 받기 때문에 수치형 자료로 변환

- 결측값 및 이상치를 처리하여 데이터 정제

- 머신러닝의 단계에서 가장 까다롭고 오래 걸리는 과정

- 데이터 전처리를 통해 변수 간 관계를 파악하고 변수들의 특징을 더 잘 이해할 수 있음



| 03 머신러닝의 단계 - 3) 모델링 및 훈련

- 적절한 알고리즘을 선정하여 코드를 작성하는 모델링 진행 후, 기계 훈련

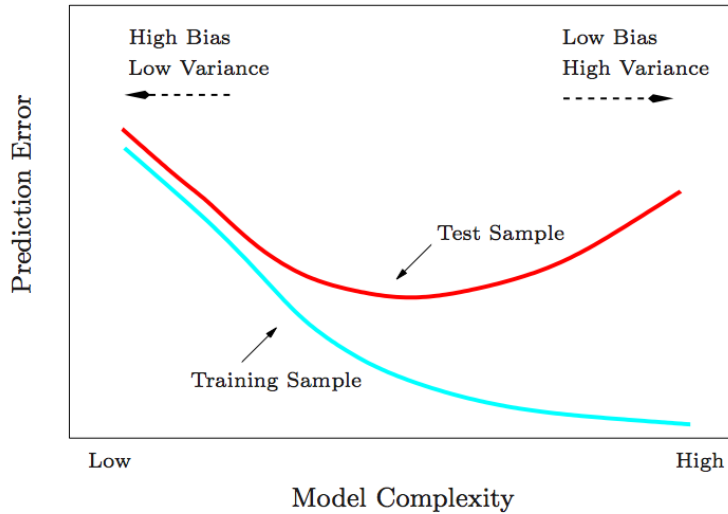
- 데이터 타입에 따라서 모델 구축/ 목적에 따라 선정하는 모델이 다름

<모델에 대한 4가지 지표>

- 모델 정확도 (Accuracy): 모델의 예측/분류 등 성과에 대한 지표
- 모델 복잡도 (Complexity): 모델의 학습 과정에서 소요되는 계산량
- 모델 해석도 (Interpretability): 학습 결과에 대한 해석력이 높을수록 모델 피드백이 쉬움
- 모델 확장도 (Scalability): 데이터의 추가에 따른 모델의 정확도 상승이 어디까지 가능한지에 대한 지표

| 03 머신러닝의 단계 - 3) 모델링 및 훈련

모델 복잡도



출처 : <https://towardsdatascience.com/model-complexity-accuracy-and-interpretability-59888e69ab3d>

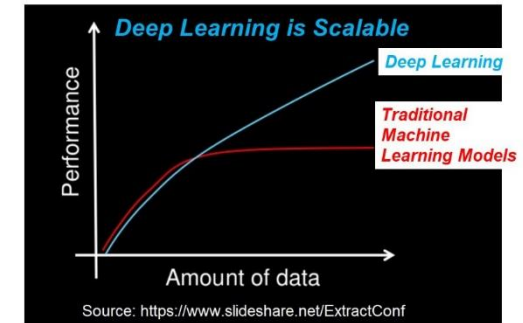
모델 해석도



출처 : <https://machine-learning.paperspace.com/wiki/interpretability>

모델 확장도

Machine Learning vs Deep Learning: Scalability



Source: <https://www.slideshare.net/ExtractConf>
Michio Sugino (Deep Origami), CFA: <https://www.linkedin.com/in/reversalpoint/>



| 03 머신러닝의 단계 - 4) 모델 평가 / 5) 모델 배포

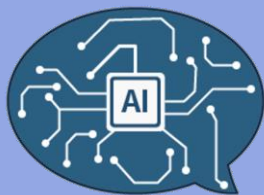
- 평가 지표를 이용해 모델을 평가

- > 사용 모델에 따라 다양한 평가 지표가 있음

- 배포 후, 성능을 지속적으로 체크

- 사용자의 피드백을 통해 성능 개선 / 새로운 모델의 개발로 이어질 수 있음





감사합니다

