D&A

ML Session 1차시

머신러닝 기초.

2022 / 09 / 06

D&A 학회장 권유진

CONTENTS.

01. ORIENTATION

세미나

대면 강의(경영관 B107-1호)
매주 화요일 06:00PM ~ 07:40PM
강의를 녹화해서 Youtube에 업로드
45분 이론 수업, 10분 휴식, 45분 실습 수업

스터디

매주 한 명씩 번갈아 가며 세미나 내용 Review
과제 수행 후 발표 및 어려운 점 상의
활동 사진, 활동 내용, 참여 인원을 포함한 스터디 보고서 작성

과제

세미나 시작 하루 전(매주 월요일 11:59PM) 까지 과제 완성본 구글 드라이브에 제출

01. ORIENTATION

청강기간

9/6~9/27(3주간)
본인이 끝까지 학회 활동에 성실히 참여할 수 있는지 판단하는 기간(매주 세미나스터디 참가, 과제 제출)
학회의 방향이 자신과 맞지 않거나 성실한 참여가 어려울 것 같다고 느껴지면 하차 가능
단, 청강 기간이 끝난 후 책임감을 갖고 참여해야 함
청강 기간에 불성실한 참여로 다른 학회원에게 피해를 입힐 시, 세션 참여에 어려움이 있을 수 있음

상품

과제 완성도, 참여도 등을 종합하여 우수자 선정
세션 후반에 진행되는 ML Competition 우수자 선정
수료증 지급

01. ORIENTATION

구글 드라이브

세션 세미나 자료 다운로드 & 매 주 과제 수행 후 업로드
https://drive.google.com/drive/folders/1yCONwMROQ64014Y88pibx91iSFbG_OqX?usp=sharing

홈페이지

https://cms.kookmin.ac.kr/kmu-dna

카카오톡 채널(플러스 친구)

각종 질문, 건의 사항 문의
https://pf.kakao.com/_zTEGb

인스타그램, 페이스북

학회 및 세션 관련 공지 매주 세미나 요약
인스타그램: https://www.instagram.com/kmu_dna
페이스북: https://www.facebook.com/kookmin.bigdata.dna2013

유튜브

세미나 강의 녹화 (매주 링크 제공)
https://www.youtube.com/channel/UCaypwX7F1SKXM0X7J6Uzs7A

01. ORIENTATION

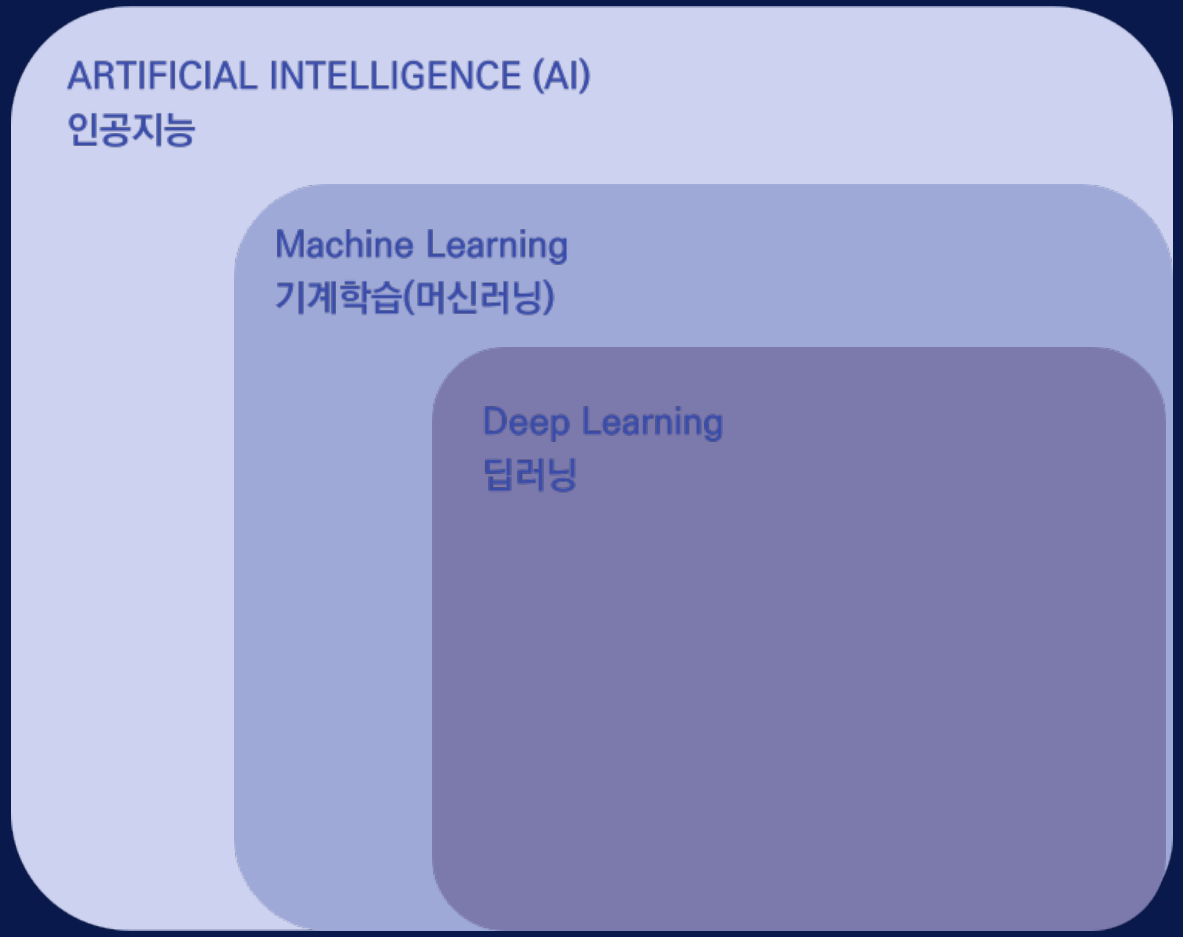
월A					운영진	월B					운영진
김서령	김예향	류병하	이은지	황건하	권유진	강민수	김승혁	김채원	김현조	우호경	나요셉
월C						수A					
김지은	노명진	심재민	이서연	이창훈	이수빈	남현서	원대인	이수인	전영호	주현민	이예진
수B						목A					
김종민	박지민	신수옥	윤태양	최다은	윤경서	김해나	박상수	배지환	윤상진	정가연	이경욱
목B						금A					
김해우	송유나	장예진	최민지	한준규	윤경서	배민성	박수현	신기성	신재웅	이재혁	김정하
금B											
김강연	김수지	임형빈	황태균		이경욱						

01. ORIENTATION

차시	날짜	내용	발표자
1	09/06	머신러닝 기초	권유진
2	09/13	분류/회귀, 데이터 분할, 교차검증, 평가지표	윤경서
3	09/20	회귀 모델	이경욱
4	09/27	분류 모델	이예진
5	10/04	데이터 전처리, 모델 튜닝	김정하
6	10/11	Bagging	윤경서
7	11/08	Boosting	권유진
8	11/15	Ensemble, Voting, Stacking	이수빈
9	11/22	Feature Extraction	나요셉
10	11/29	ML Competition 발표회	-

Competition

02. 머신러닝이란?


ARTIFICIAL INTELLIGENCE (AI)
인공지능
Machine Learning
기계학습(머신러닝)
Deep Learning
딥러닝

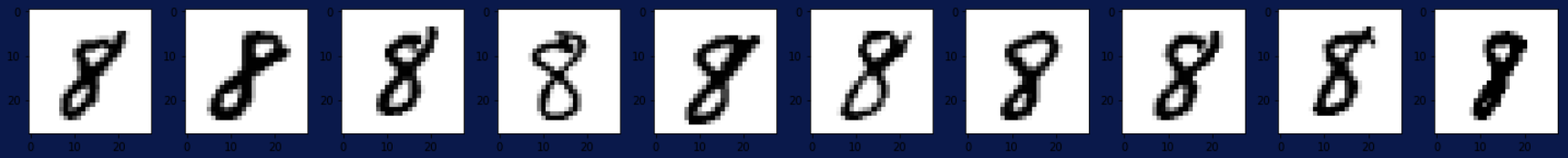
02. 머신러닝이란?

머신러닝이란?

Machine Learning (기계 학습)

컴퓨터가 데이터에서부터 규칙을 찾아 학습하도록 프로그래밍

인간이 감지할 수 없는 어렵고 복잡한 문제의 패턴을 감지하여 판단에 좋은 기준을 자동으로 학습



머신러닝을 사용하지 않으면 직접 패턴을 발견하고 알고리즘으로 작성하는 과정 반복해야 함

ex) 8은 구멍이 2개이고 중간 부분이 홀쭉하며 맨 위와 아래가 둥근 모양

머신러닝을 사용하면 프로그램이 훨씬 짧아지고 유지보수가 쉽고 정확도가 높다.

ex) 숫자가 적힌 사진과 라벨(정답 값)을 함께 입력해주면 컴퓨터가 패턴을 찾아 학습

머신러닝 기술을 적용해 대용량 데이터를 분석하면 겉으로는 보이지 않던 패턴 발견(데이터 마이닝)

02. 머신러닝이란?

머신러닝 종류

여러 가지 기준에 따라 분류 가능

감독 여부

지도 학습(Supervised Learning) / 비지도 학습(Unsupervised Learning) / 강화학습(Reinforcement Learning)

실시간, 점진적 학습 여부

배치학습(Batch Learning) / 온라인 학습(Online Learning)

Task 수행 방법

사례 기반 학습(Cased-Based Learning) / 모델 기반 학습(Model-Based Learning)

02. 머신러닝이란?

감독 여부

지도 학습(Supervised Learning) / 비지도 학습(Unsupervised Learning) / 강화학습(Reinforcement Learning)

훈련 데이터의 라벨(Label)의 여부에 따라 분류

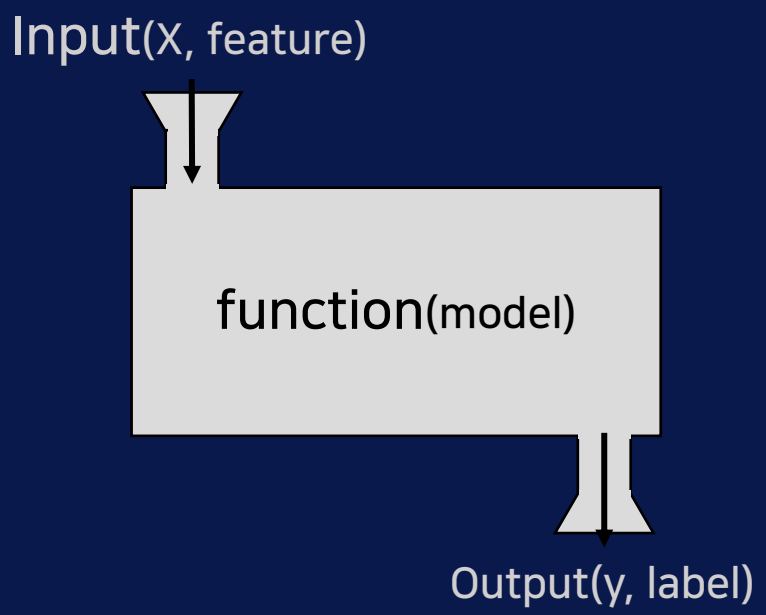
특성(feature)

입력 변수(회귀 모델에서 변수 X에 해당)

라벨(Label)

예측하는 항목(회귀 모델에서 변수 y에 해당)

Target, Class 라고도 부름



02. 머신러닝이란?

지도학습

라벨이 존재하는 학습 데이터를 학습하는 것

정답이라고 가정한 내용에 맞게 컴퓨터가 예측

분류, 회귀로 나뉨

Sepal length(cm)	Sepal width(cm)	Petal length(cm)	Petal width(cm)	label
5.1	3.5	1.4	0.2	setosa
5.6	3.	4.5	1.5	veriscolor
5.9	3.	5.1	1.8	virginica

분류(Classification)

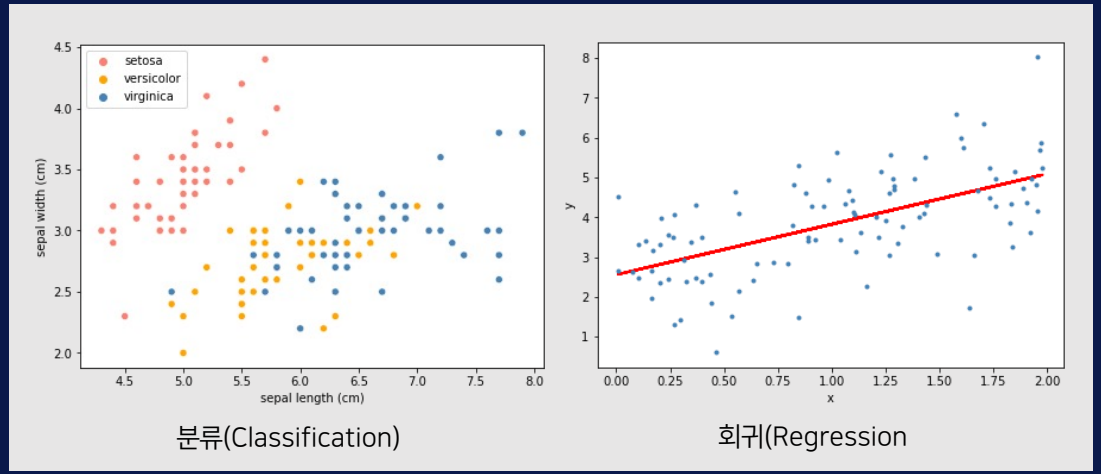
데이터의 특성을 통해 범주형 데이터인 label(class)를 예측

ex) 성별 예측, 붓꽃 종류 예측 …

회귀(Regression)

데이터의 특성을 통해 연속형 데이터인 target 수치 예측

ex) 나이 예측, 주가 예측 …



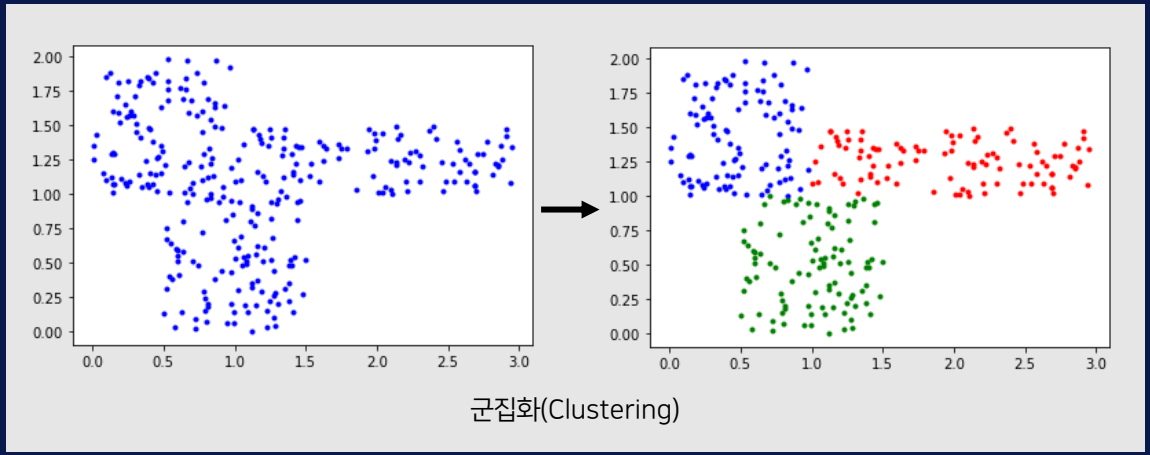
분류(Classification) 회귀(Regression

02. 머신러닝이란?

비지도학습

라벨이 존재하지 않는 학습 데이터를 학습하는 것

군집화, 차원 축소, 연관 분석 등

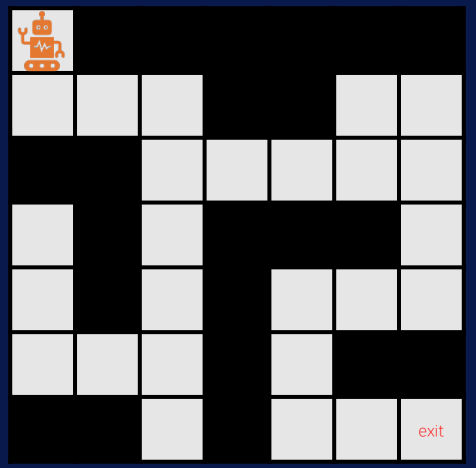


군집화(Clustering)

강화학습

환경을 관찰해 행동을 실행하고 그 결과로 보상을 최대화하는 최상의 정책 학습

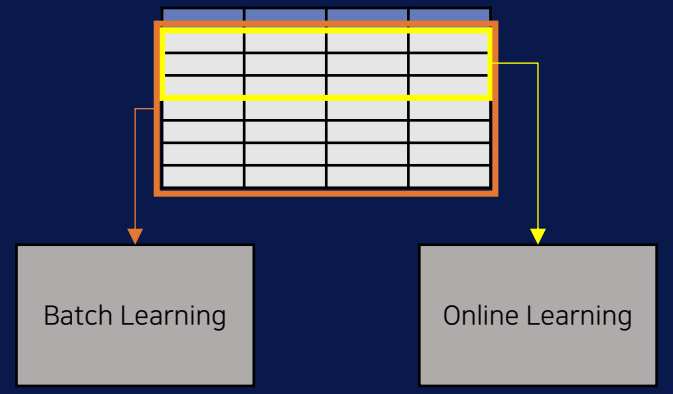
ex) 알파고



02. 머신러닝이란?

실시간, 점진적 학습 여부

배치학습(Batch Learning) / 온라인 학습(Online Learning)



배치 학습(오프라인 학습)

한번 학습하면 끝인 학습 방법

가용한 데이터를 모두 사용해 훈련

→ 자원과 시간이 많이 소요

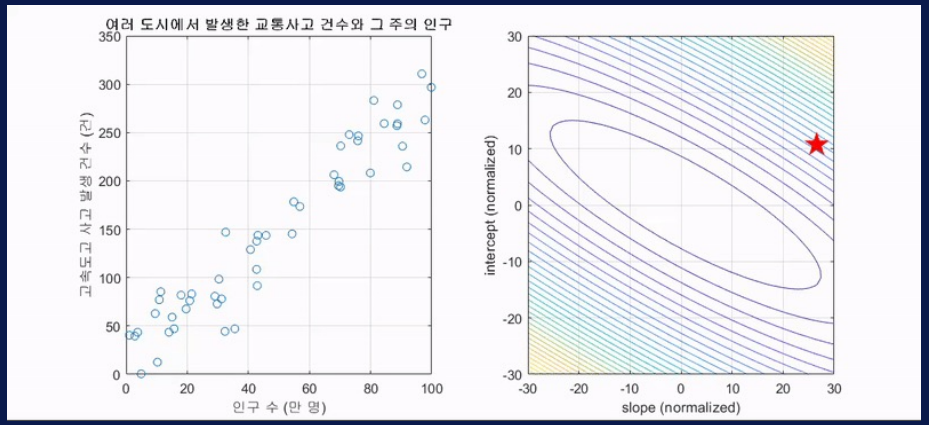
새로운 데이터 학습을 위해서는 새로운 버전을 처음부터 다시 훈련

온라인 학습

데이터를 순차적으로 한 개씩 또는 미니배치 단위로 학습

속도가 빠르고 비용이 적게 듦

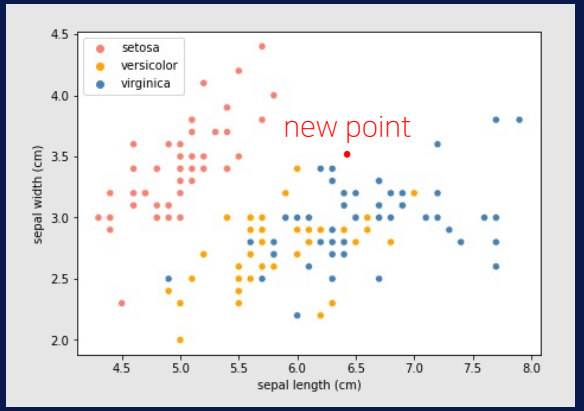
중요한 파라미터로 학습률 존재



02. 머신러닝이란?

Task 수행 방법

사례 기반 학습(Cased-Based Learning) / 모델 기반 학습(Model-Based Learning)



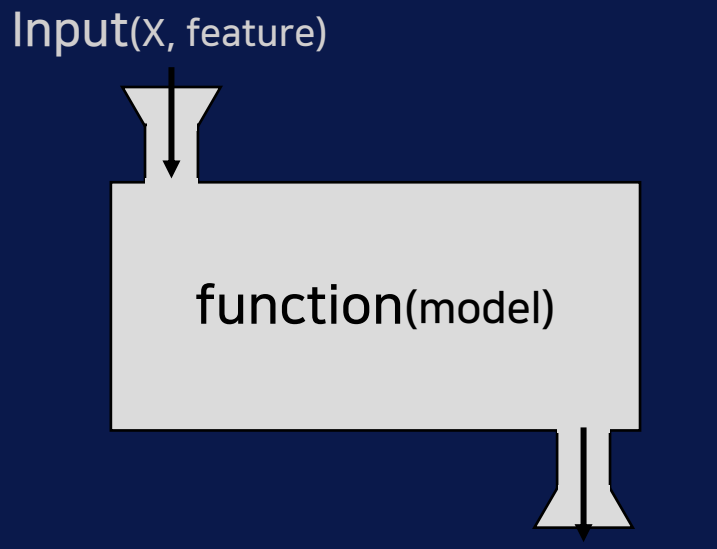
사례 기반 학습

단순히 모든 데이터를 기억 후 새로운 데이터와 비교
→ 분류의 경우, 모든 데이터와의 유사도를 측정 후 가장 유사한 데이터의 클래스로 예측

모델 기반 학습

데이터에 알맞는 모델을 만들어 예측에 사용

과정: 알맞는 모델 선택 → 파라미터 조정 → 효용/비용 함수 정의 후 성능 측정 → 추론



02. 머신러닝이란?

머신러닝 적용 사례

생산 라인에서 제품 이미지를 분석해 자동으로 분류

뇌를 스캔해 종양 진단

자동으로 뉴스 기사를 분류

토론 포럼에서 부정적인 코멘트를 자동으로 구분

다양한 성능 지표를 기반으로 회사의 내년도 수익 예측

신용카드 부정 거래 감지

구매 이력을 기반으로 고객을 나누고 각 집합마다 다른 마케팅 전략 계획

과거 구매 이력을 기반으로 고객이 관심을 가질 수 있는 상품 추천

지능형 게임 봇 만들기

등...

03. 머신러닝 수행 과정

Domain Understading & Data Collection	Data Preprocessing	Modeling & Ensemble	Prediction	Evaluation	Deployment

Domain Understanding & Data Collection

프로젝트를 진행할 데이터를 수집하고 이해하는 단계

데이터가 갖고 있는 특성을 파악하고 EDA를 통해 데이터 분석
→ 지속적으로 해당 데이터에 대한 탐색과 이해를 기본적으로 가져야 함!

데이터 분포, 결측값, 이상치 등을 시각화를 통해 확인하면서 데이터 분석

데이터 자체에 대한 해석이 잘못되면 이후에 진행되는 모든 과정들이 적절한 방향으로 진행될 수 없다!
→ 매우 중요!

03. 머신러닝 수행 과정

Domain Understading & Data Collection	Data Preprocessing	Modeling & Ensemble	Prediction	Evaluation	Deployment

Data Preprocessing

데이터 전처리 과정으로 머신러닝에서 가장 많은 시간과 노력을 투자해야하는 단계

결측값, 이상치를 처리하고 feature를 만듦

많은 feature를 만들고 유의미하다고 판단되는 feature를 feature selection을 통해 골라서 사용

모델이 값을 잘 예측할 수 있는 유의미한 feature를 제공해야 성능이 좋은 모델을 만들 수 있다.

→ Garbage In, Garbage Out!

03. 머신러닝 수행 과정

Domain Understading & Data Collection	Data Preprocessing	Modeling & Ensemble	Prediction	Evaluation	Deployment

Modeling & Ensemble

데이터에 적합한 모델을 설계하는 과정

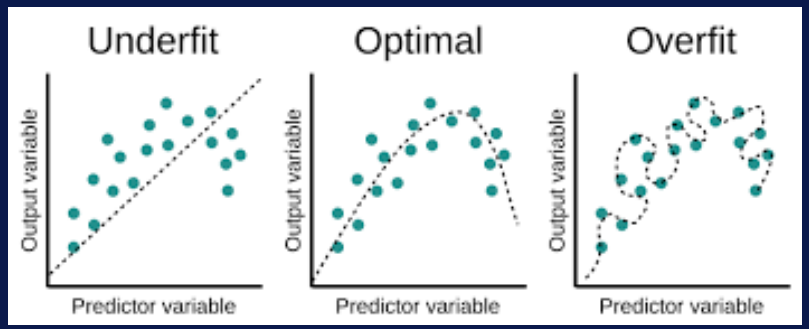
정답으로 가정한 데이터의 값과 모델을 통해 예측한 값의 차이가 적어질 수 있도록 학습

모델에서 사용되는 하이퍼 파라미터를 조정

→ 하이퍼 파라미터: 모델링할 때 사용자가 직접 세팅하여 주는 값

03. 머신러닝 수행 과정

Domain Understading & Data Collection	Data Preprocessing	Modeling & Ensemble	Prediction	Evaluation	Deployment



Evaluation

실제 정답과 모델의 예측값의 차이 정도를 통해 해당 모델이 잘 학습된 모델인지 평가

이때 과적합에 유의!

과대적합(Overfitting): 모델이 너무 학습이 잘돼 훈련 데이터에 너무 최적화된 상황

→ 훈련 데이터의 경우, 예측을 잘 하지만 훈련 데이터가 아닌 데이터는 잘 예측하지 못함(훈련 성능은 높지만 검증 성능이 낮음)

→ 모델이 너무 복잡해 일반성이 떨어진다는 의미

과소적합(Underfitting): 모델이 너무 단순해서 데이터의 내재된 구조를 학습하지 못하는 것을 의미

03. 머신러닝 수행 과정

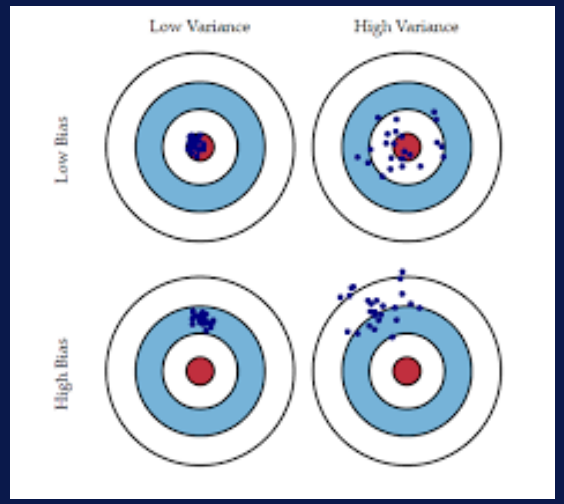
Domain Understading & Data Collection	Data Preprocessing	Modeling & Ensemble	Prediction	Evaluation	Deployment

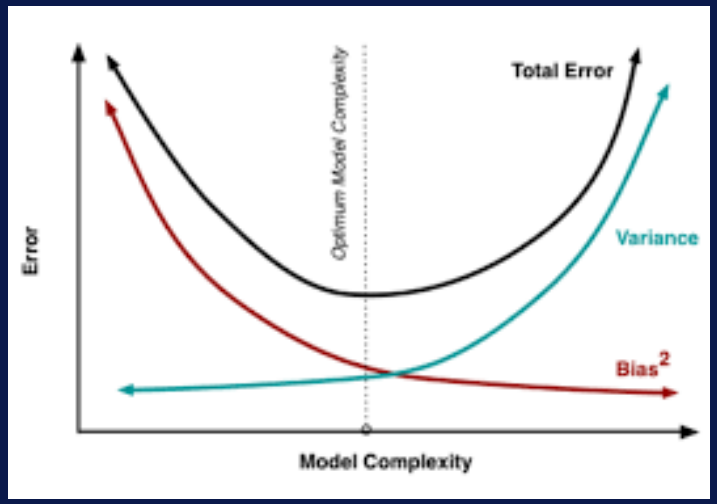
Evaluation

편향(bias)
예측이 정답에서 얼마나 떨어져 있는지 의미
→ 편향이 크면 과소적합

분산(variance)
예측의 변동폭이 얼마나 큰지 의미
→ 분산이 크면 과대적합

편향과 분산은 trade-off 관계
→ 1개가 증가하면 1개가 감소한다.





03. 머신러닝 수행 과정

Domain Understading & Data Collection	Data Preprocessing	Modeling & Ensemble	Prediction	Evaluation	Deployment

Deployment

최종 모델을 선정해 실제로 사용할 수 있도록 상용화

상용화 후, 일정 간격으로 실시간 성능 체크 및 모니터링

03. 머신러닝 수행 과정

Domain Understading & Data Collection	Data Preprocessing	Modeling & Ensemble	Prediction	Evaluation	Deployment

예시

실제로 우리가 머신러닝을 사용하게 된다면 위 과정을 어떻게 거칠까?

ex) titanic dataset

Passenger ID	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked	Survived
1	3	Harris	male	22.0	1	0	A/5 21171	7.25	NaN	S	0

04. 머신러닝의 한계

머신러닝 한계

과적합, 과도한 일반화로 인해 성능 향상에 한계
→ 과대적합 해결법: 모델의 경량화, 학습 데이터 수 증가, 모델 규제, 앙상블
(과대적합은 데이터에 비해 모델이 복잡하여 발생하기 때문)
→ 과소적합 해결법: 모델 복잡도 증가
(과소적합은 데이터에 비해 모델이 가벼워 발생하기 때문)

정답이 있는 대량의 데이터 필요

도출 결과의 설명력 부족
대부분 머신러닝 모델은 BlackBox 모델(대출 가능 여부를 머신러닝 모델을 통해 결정한다고 할 때, 고객에게 기각된 이유를 설명 불가능)
→ BlackBox: 모델의 예측 결과를 설명할 수 없는 모델
→ WhiteBox: 모델의 예측 결과를 설명할 수 있는 모델

기존 학습 모델의 재사용 어려움
→ 금융 분야에서 학습된 모델을 법률 분야에서 적용 불가능

APPENDIX. Scikit-Learn

사용 과정	모듈	설명
Data Cleansing & Feature Engineering	sklearn.preprocessing	데이터 전처리 (인코딩, 정규화 등)
	sklearn.feature_selection	Feature 선택
	sklearn.feature_extraction	Feature 추출
Model Evaluation	sklearn.model_selection	데이터 분리, 검증 및 모델 튜닝
	sklearn.metrics	성능 평가
Supervised Learning	sklearn.linear_model	선형 모델
	sklearn.svm	서포트 벡터 머신
	sklearn.tree	의사결정나무
	sklearn.ensemble	앙상블 알고리즘
Unsupervised Learning	sklearn.cluster	군집 분석
	sklearn.decomposition	차원 축소
Utility&Dataset	sklearn.pipeline	워크플로우 효율화
	sklearn.datasets	예제 데이터셋

00. 과제

파이썬 문법을 활용해 여러가지 숫자형(numeric) feature 5개 이상 만들기
만든 feature를 활용하여 머신러닝 수행 코드 돌려보기

+

조별로 조 이름과 조장, 스터디 시간, 발표 순서를 정해서
스터디 보고서를 작성해 구글 드라이브에 업로드

첨부자료 출처

02. 머신러닝이란?

경사하강법_gif출처: https://angeloyeo.github.io/2020/08/16/gradient_descent.html

03. 머신러닝 수행과정

과소적합, 과대적합_이미지출처: https://www.educative.io/answers/overfitting-and-underfitting
분산, 편향_이미지출처: https://opentutorials.org/module/3653/22071
분산, 편향 trade-off_이미지출처: https://bkshin.tistory.com/entry/%EB%A8%B8%EC%8B%A0%EB%9F%AC%EB%8B%9D-12-%ED%8E%B8%ED%96%A5Bias%EC%99%80-%EB%B6%84%EC%82%B0Variance-Trade-off

---

폰트

네이버 글꼴 모음 _ 나눔 스퀘어 사용
출처 : https://hangeul.naver.com/font

D&A

ML Session 1차시 머신러닝 기초

Thank You.

2022 / 09 / 06

D&A 학회장 권유진