ORIENTATION

ML Session 1차시

빅데이터 분석 학회 D&A

CONTENTS.

01. ORIENTATION

- 수업방식
- 스터디
- 커리큘럼

02. 머신러닝이란?

- 머신러닝이란?
- 머신러닝의 종류
- 머신러닝 적용 사례

03. 머신러닝 수행과정

- 머신러닝 수행 과정

04. 머신러닝의 한계

- 머신러닝의 한계

ORIENTATION

수업 방식

세미나

매주 화요일 (18:00 ~ 20:00) ※ 10/15, 10/22 중간고사 기간이므로 수업 X
장소 : 경영관 103호
강의는 매 주차 녹화를 통해 YouTube 업로드 예정

스터디

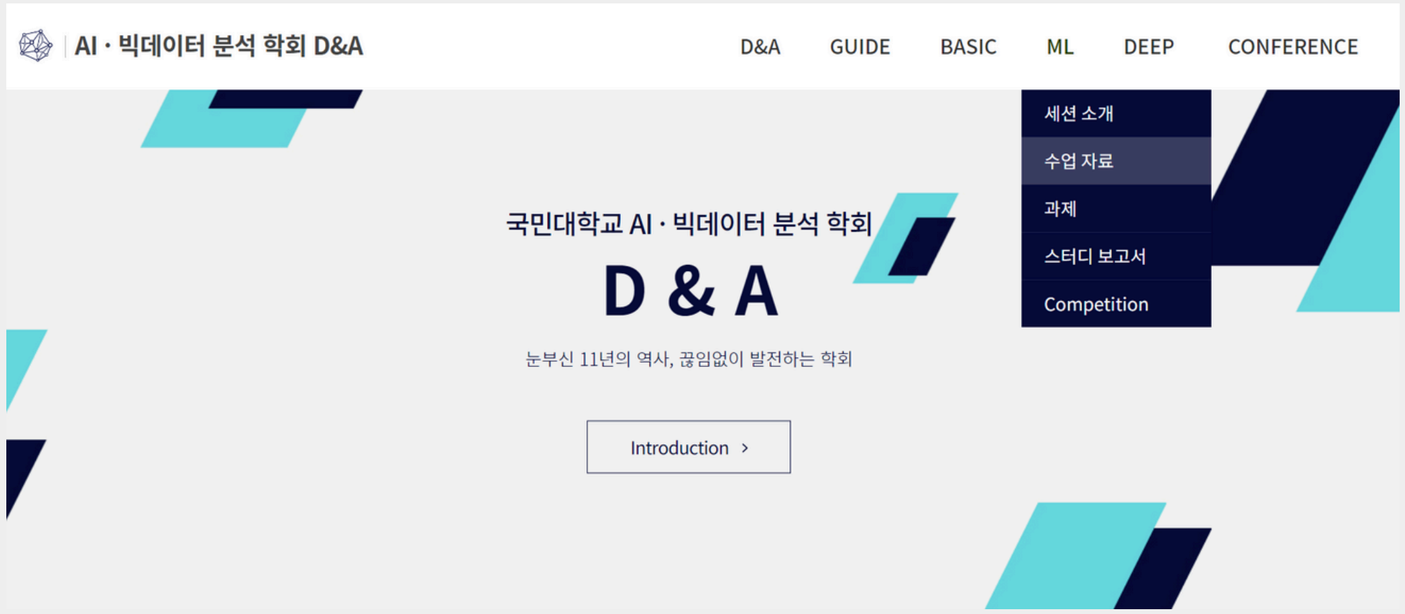
매 주 세미나 내용 Review
과제 수행 후 발표 및 어려운 점 논의
활동사진, 활동내용, 참여인원을 포함한 스터디 보고서 작성

상품

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

ORIENTATION

수업 방식



과제 제출

세미나 시작 하루 전 (**매주 월요일 23:59**)까지 제출
ML - 과제 게시판에 업로드
제출 시, 제목은 '**N주차 [조 이름] [성명]**'으로 제출
조장은 **이름 뒤에 (조장)**을 넣어서 제출

수업 자료

'ML – 수업 자료' 게시판에 업로드 예정
자료 업로드 시, 카카오톡 전체 공지방을 통해 공지 예정

ORIENTATION

수업 방식

청강기간

9월 10일 ~ 9월 24일 (2주차까지)
본인이 끝까지 학회 활동에 성실히 참여할 수 있는지 판단하는 기간(매주 세미나, 스터디 참가, 과제 제출)

- 성실한 참여가 어려울 것 같다고 느껴지면 하차 가능
- 단, 청강 기간이 끝난 후에는 책임감을 가지고 참여해야 함

홈페이지

세션 세미나 자료 다운로드
https://cms.kookmin.ac.kr/dna/index.do

카카오톡

각종 질문, 건의 사항 문의
http://pf.kakao.com/_lfpJG

유튜브

실시간 강의 녹화 (매주 링크 제공)
https://www.youtube.com/@kmudna

인스타그램

학회 및 세션 관련 공지, 매주 세미나 요약
https://www.instagram.com/kmu_dna/

ORIENTATION

스터디

멘토 1	멘토 2	화요일 1조						
조현식	신지후	김태현	김선아	오준호	오창석	이정제		
		화요일 2조						
김서령	손아현	송영준	이여경	이용찬	이정훈	조은나라		
		수요일 조						
이준혁	김차미	강성원	송민승	이선민	조영우	허지원		
		목요일 조						
김예향	이지민	김유라	김지원	손현수	오서영	이상원	원예지	

ORIENTATION

커리큘럼

차시	날짜	내용	발표자
1	9/10	OT + 머신러닝 기초 1	이준혁
2	9/17	머신러닝 기초 2	김서령
3	9/24	회귀 모델 + 분류 모델	김차미
4	10/8	Data Preprocessing 1 (Data Cleansing, Feature Extraction)	손아현
5	10/29	Data Preprocessing 2 (Feature Selection, Optimization, PCA)	신지후
6	11/05	Bagging + Boosting	이지민
7	11/12	Ensemble (Voting, Stacking)	김예향
8	11/19	AutoML	조현식
9	11/26	ML competition 발표회	이준혁

머신러닝이란?

Machine Learning(기계학습)

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 과정을 반복해야 함
  - ex) 8은 구멍이 2개이고 중간 부분이 홀쭉하며 맨 위와 아래가 둥근 모양
- 머신러닝을 사용하면 프로그램이 훨씬 짧아지며 유지보수가 쉽고 정확도가 높음
  - ex) 숫자가 적힌 사진과 정답 값을 함께 입력해주면 컴퓨터가 패턴을 찾아 학습
- 머신러닝 기술을 적용해 대용량 데이터를 분석하면 겉으로는 보이지 않던 패턴 발견

머신러닝의 종류

머신러닝의 종류

여러 가지 기준에 따라 분류 가능

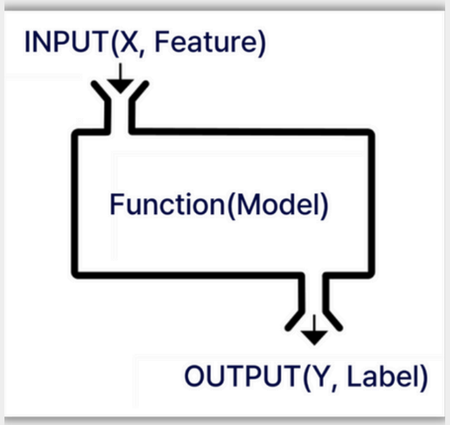
- **감독 여부**
  - 지도 학습 (Supervised Learning), 비지도 학습 (Unsupervised Learning), 강화학습 (Reinforcement Learning)
- **실시간, 점진적 학습 여부**
  - 배치 학습 (Batch Learning), 온라인 학습 (Online Learning)
- **Task 수행 방법**
  - 사례 기반 학습 (Case-Based Learning), 모델 기반 학습 (Model-Based Learning)

머신러닝의 종류

감독 여부

훈련 데이터의 라벨(Label) 여부에 따른 분류

- **특성** (feature)
  - 입력 변수 (회귀 모델에서 변수 X에 해당)
- **라벨** (Label)
  - 예측되는 변수 (회귀 모델에서 변수 y에 해당)
  - Target, Class 라고도 부름



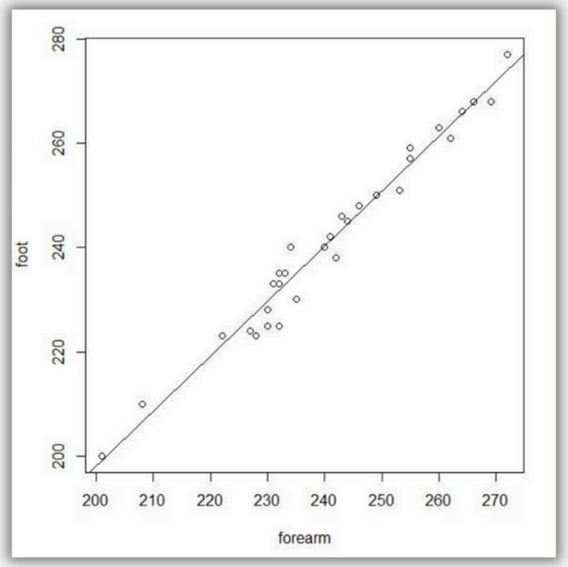
머신러닝의 종류

지도학습

- 라벨이 존재하는 학습 데이터를 학습하는 것
- 정답이라고 가정한 내용에 맞게 컴퓨터가 예측
- 분류와 회귀로 나뉨

Sepal length	Sepal width	petal length	petal width	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 특성**을 통해 **범주형 데이터**인 label(class)를 예측
    - 성별 예측, 붓꽃 종류 예측 등
- **회귀** (Regression)
  - **데이터의 특성**을 통해 **연속형 데이터**인 target을 예측
    - 나이 예측, 주가 예측 등



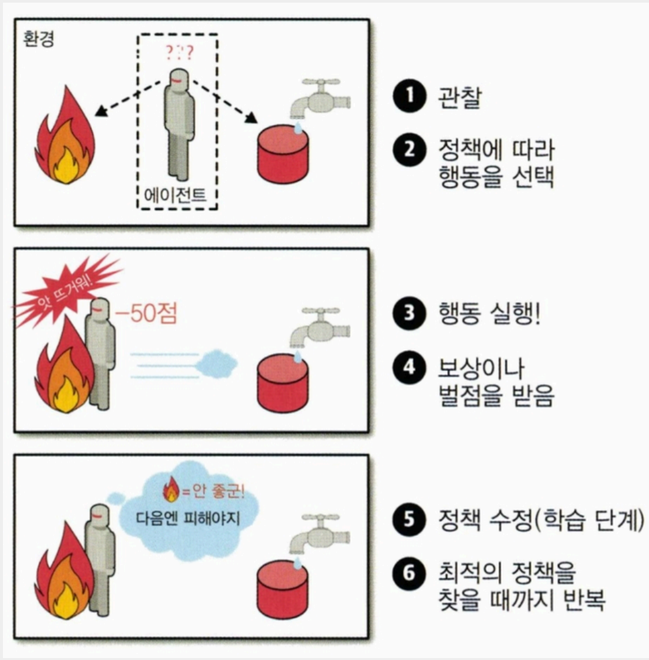
머신러닝의 종류

비지도 학습

- **라벨이 존재하지 않는 훈련 데이터**를 학습하는 방법
  - 군집화, 차원 축소, 연관 분석 등

강화학습

- **환경 안에서 정의된 에이전트**가 현재의 상태를 인식하여, 선택 가능한 행동들 중 **보상을 최대화하는 행동 혹은 행동 순서를 선택**하는 방법



머신러닝이란?

머신러닝의 종류

실시간, 점진적 학습 여부

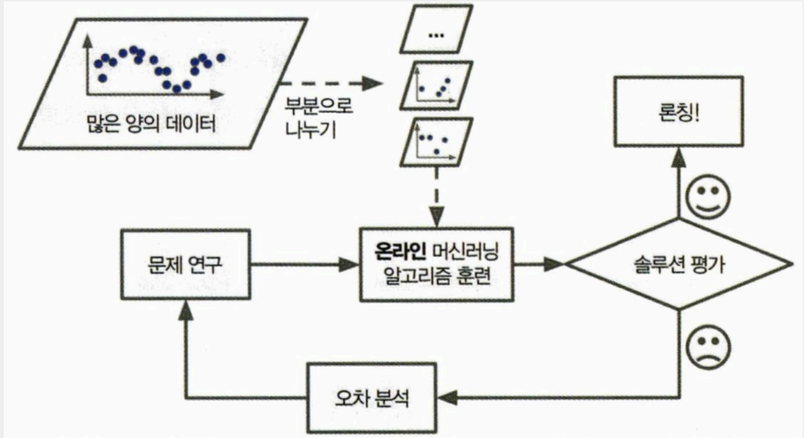
- 입력 데이터의 stream으로부터 **점진적으로 학습할 수 있는가?**
- 배치 학습 (batch Learning), 온라인 학습 (Online Learning)

배치학습

- 모델을 학습시킨 뒤, 더 이상의 학습 없이 시스템에 적용
- 가용한 데이터를 모두 사용해서 훈련
  - 많은 자원과 시간이 소요
- 새로운 데이터 학습을 위해서는 새로운 버전을 처음부터 다시 훈련

온라인 학습

- 학습된 모델이 시스템에 적용된 상태에서도, 미니배치 단위로 점진적인 학습을 추가적으로 진행
  - 속도가 빠르고 비용이 적게 듦
  - 학습률이 중요함



온라인 학습을 이용한 대용량 데이터 처리

머신러닝의 종류

Task 수행 방법

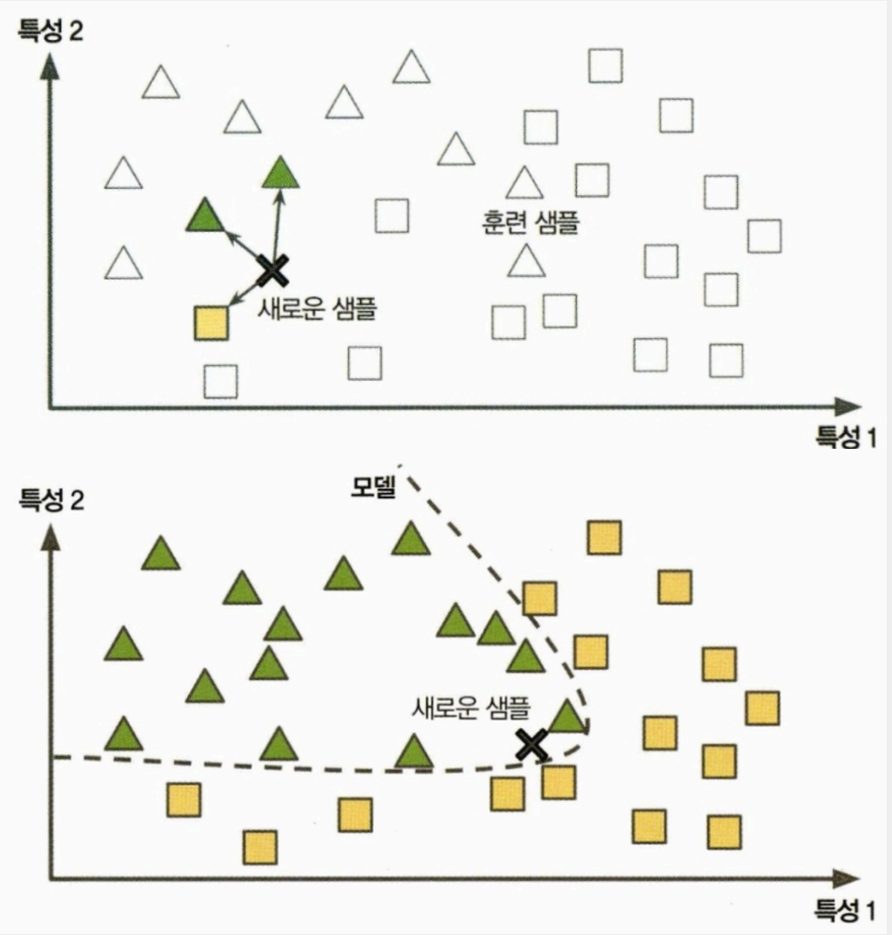
- 사례 기반 학습 (Case-Based Learning)
- 모델 기반 학습 (Model-Based Learning)

사례 기반 학습

- 단순히 모든 데이터를 기억 후 새로운 데이터와 비교
  - 분류의 경우, 모든 데이터와의 유사도를 측정 후, 훈련 데이터 중 가장 유사한 데이터의 클래스로 해당 데이터를 예측

모델 기반 학습

- 데이터에 알맞는 모델을 만들어 예측에 사용
- 알맞은 모델 선택 -> 파라미터 조정 -> 비용 함수 정의 후 성능 측정 -> 추론



머신러닝 적용 사례

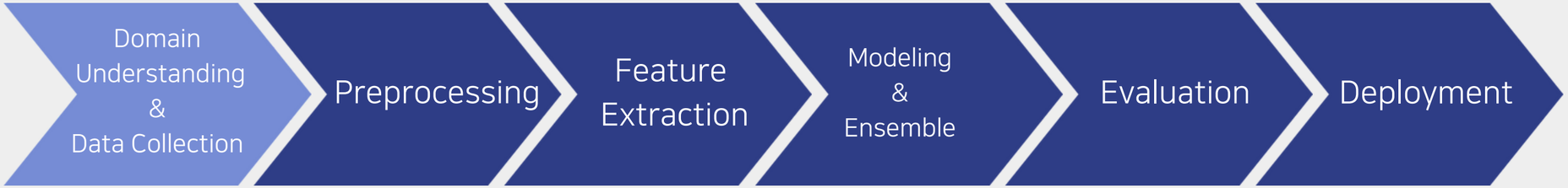
머신러닝 적용 사례

- 생산 라인에서 제품 이미지를 분석해 자동으로 분류
- 뇌를 스캔하여 종양 진단
- 자동으로 뉴스 기사 분류
- 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

머신러닝 적용 사례

Domain Understanding & Data Collection

프로젝트를 진행할 데이터를 수집하고 이해하는 단계

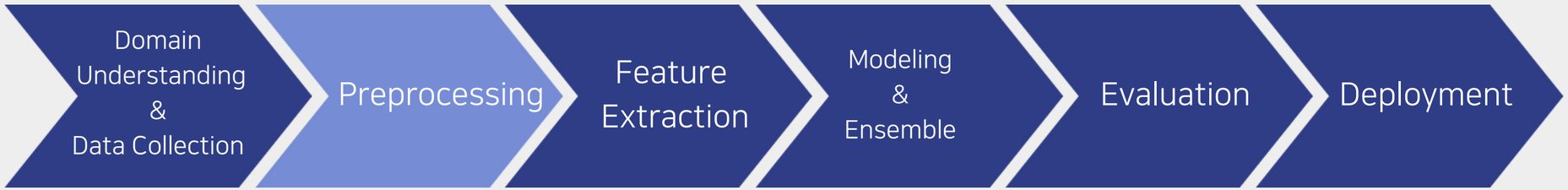


- **데이터가 갖고 있는 특성을 파악**하고 EDA를 진행
  - 지속적으로 해당 데이터에 대한 탐색과 이해를 기본적으로 가져야 함
- 데이터 분포, 결측값, 이상치 등을 시각화를 통해 확인하면서 데이터 분석
- 데이터 자체에 대한 해석이 잘못되면 이후에 진행되는 모든 과정들이 적절한 방향으로 진행될 수 없음

머신러닝 적용 사례

Data Preprocessing

데이터 전처리 과정으로 머신러닝에서 가장 많은 시간과 노력을 투자해야 하는 단계

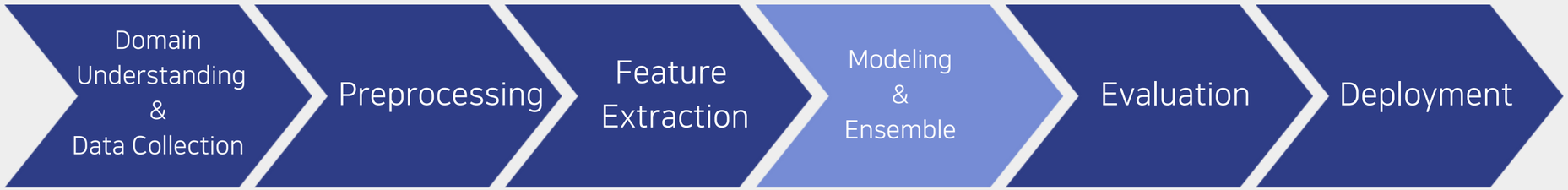


- **결측값과 이상치를 처리하고 feature를 만듦**
  - 많은 feature를 만들고 유의미하다고 판단되는 feature를 feature selection을 통해 골라서 사용
- 모델이 값을 잘 예측할 수 있는 **유의미한 feature**를 제공해야 성능이 좋은 모델을 만들 수 있음
  - Garbage in, Garbage Out

머신러닝 적용 사례

Modeling & Ensemble

데이터에 적합한 모델을 설계하는 과정



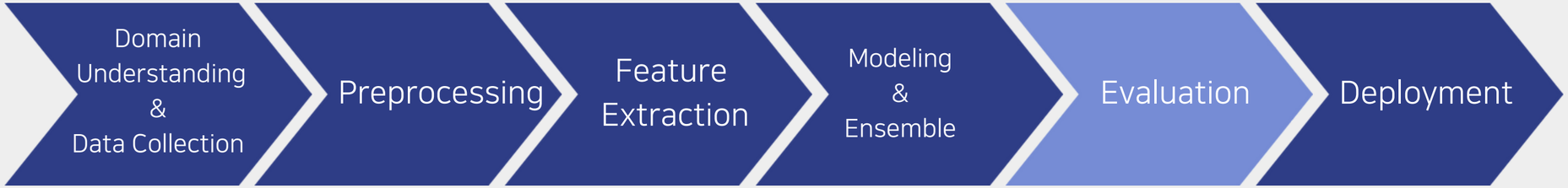
- 정답으로 가정한 데이터의 값과 모델을 통해 예측한 값의 **차이가 적어질 수 있도록** 학습
  - 모델에서 사용되는 *하이퍼 파라미터를 조정

* 하이퍼 파라미터 : 모델링할 때, 사용자가 직접 세팅해주는 값

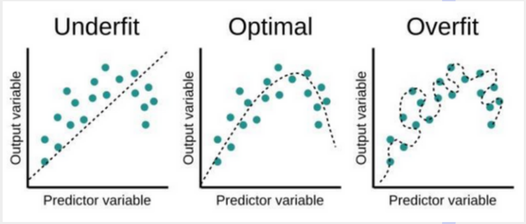
머신러닝 적용 사례

Evaluation

실제 정답과 모델의 예측값의 차이 정도를 통해 해당 모델이 잘 학습된 모델인지 평가. 이때, 과적합에 유의해야 함.



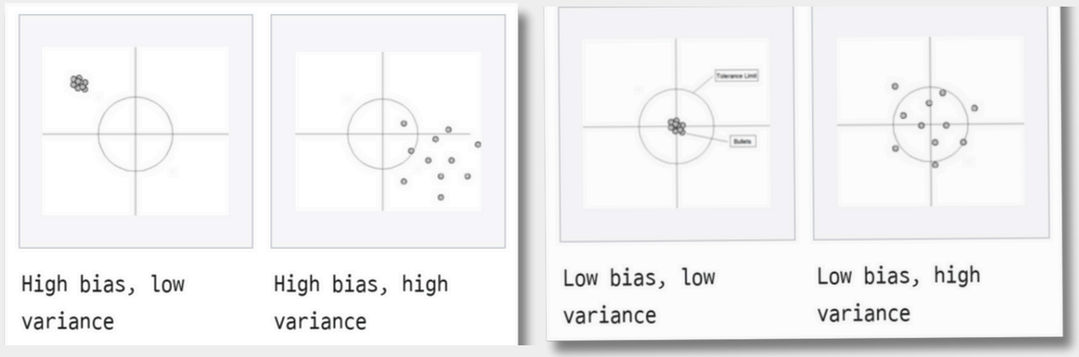
- 과대적합 (Overfitting) : 모델이 훈련 데이터에 너무 최적화된 상황
  - 훈련 데이터의 경우, 예측을 잘 하지만 훈련 데이터가 아닌 데이터는 잘 예측하지 못함 (훈련 성능은 높지만 검증 성능은 낮음) -> 모델이 너무 복잡해서 **일반성이 떨어진다**는 의미
- 과소적합 (Underfitting) : 모델이 너무 단순해서 데이터에 내재된 구조를 학습하지 못하는 것



머신러닝 적용 사례

Evaluation

실제 정답과 모델의 예측값의 차이 정도를 통해 해당 모델이 잘 학습된 모델인지 평가. 이때, 과적합에 유의해야 함.



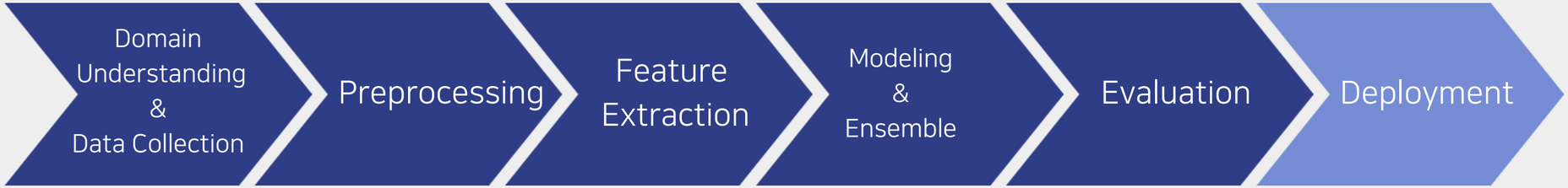
- 편향: **학습 알고리즘이 잘못된 가정**을 했을 때 발생하는 오차
  - 편향이 크면 과소적합
- 분산: **학습 데이터에 내재된 작은 변동** 때문에 발생하는 오차
  - 분산이 크다는 것은 노이즈까지 모델링에 포함시켰다는 의미.(과대적합)
- 편향과 분산은 trade-off 관계

*중앙 : 실제값

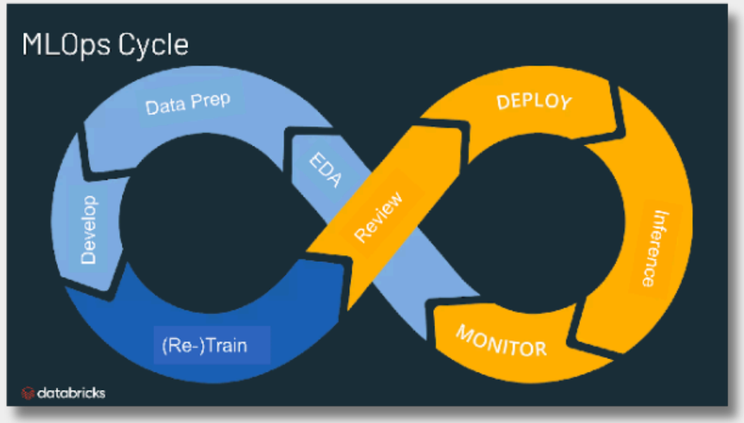
머신러닝 적용 사례

Deployment

최종 모델을 선정해 실제로 사용할 수 있도록 상용화



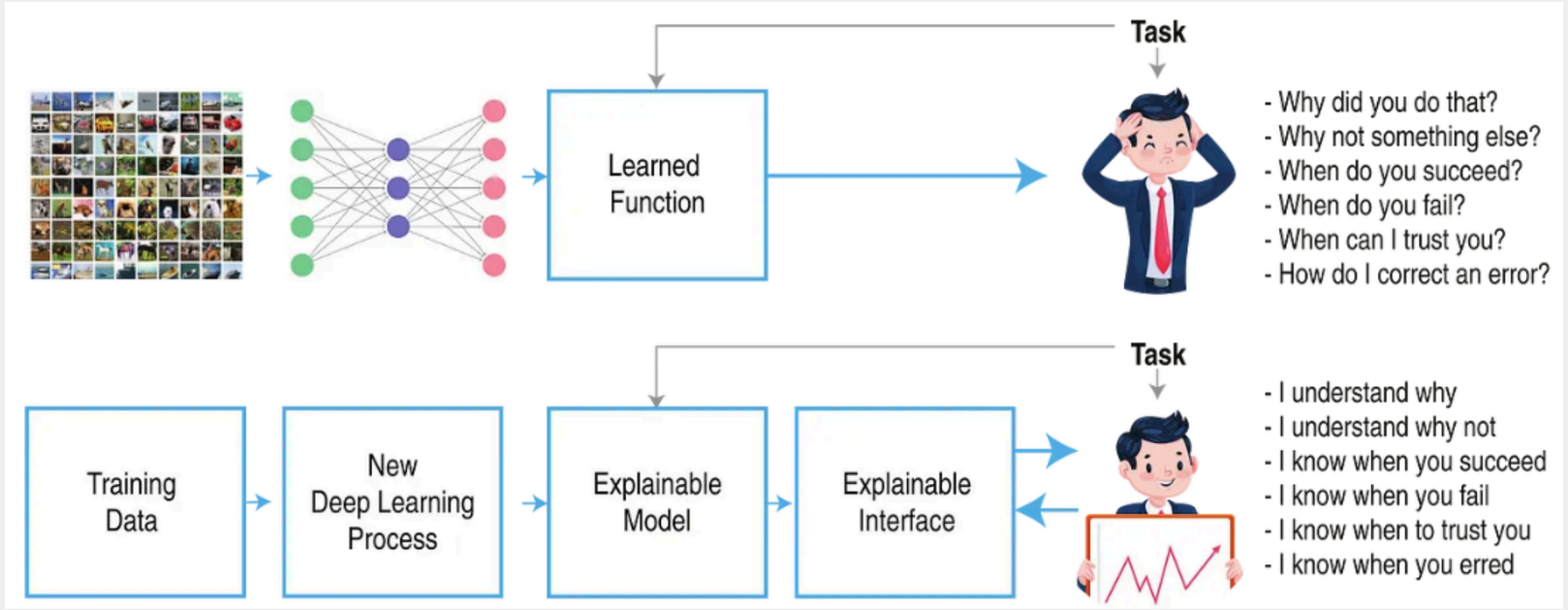
- 상용화 후 일정 간격으로 실시간 성능 체크 및 모니터링



머신러닝의 한계

머신러닝의 한계

1. 과적합
2. 정답이 있는 대량의 데이터 필요
3. 기존 학습 모델의 재사용 어려움
4. 도출 결과의 설명력 부족



과제

1. 파이썬 문법을 활용해 여러가지 numeric feature 5개 이상 만들기
2. 만든 feature를 활용하여 머신러닝 수행 코드 돌려보기
3. 조 별로 조이름과 조장, 발표 순서 정하기
4. 스터디 보고서 작성해서 홈페이지에 업로드하기

THANK YOU

ML Session 1차시