ML SESSION

#1
MACHINE LEARNING

DATA&ANALYTICS

INDEX

0 ORIENTATION

월A				운영진
김진호	이승학	이효빈	주민지	윤성식
월B				
전동균	김지민	한윤지	임지연	윤성식
월C				
황태균	이수현	이승준	박선혜	장성민
월D				
이성규	정승민	최여진	김태양	장성민
화A				
홍종현	이정호	하서경	이수빈	김예원
화B				
윤한빈	이예진	김희운	이현지	김예원

화C				운영진
이상준	이상우	김나은	도지은	한보혜
화D				
이경욱	한병규	구준회	이상현	한보혜
수A				
고민성	김은욱	최영재	유수빈	조영진
수B				
최민석	김건우	김진비	천예은	조영진
수C				
장성현	조문주	김현지	이수민	마민정
수D				
나요셉	권유진	권수연	이태범	마민정

0 ORIENTATION

차시	날짜	수업내용	발표자
1	09/09	OT, 머신러닝 기초 (정의, 종류, 과정, 피처)	김예원
2	09/16	모델링1 (경사하강법, 회귀&분류 모델, 파라미터 튜닝1)	마민정
3	09/23	교차검증, 평가지표 (회귀&분류 평가지표)	김예원,한보혜
4	09/30	데이터 전처리 (결측치, 이상치, Scailing, Corr, PCA, FS)	윤성식
5	10/07	모델링2 (배깅, 랜덤포레스트, 파라미터 튜닝2)	조영진
6	10/28	컴페티션1 (Adaboost, GBM, XGB, LGBM)	장성민
7	11/04	컴페티션2 (Word Embedding)	윤성식
8	11/11	컴페티션3 (Ensemble, Voting, Stacking)	장성민
9	11/18	컴페티션4 (머신러닝 활용사례)	한보혜
10	11/25	컴페티션 최종 발표	-

1 머신러닝이란?

머신러닝이란?

- Machine learning, 단어 그대로 해석하면 기계 학습

- 기존의 컴퓨터 프로그램이 인간이 만든 규칙을 기계에 입력해 답을 찾는 것이 목적이었다면 머신러닝은 이 규칙을 컴퓨터가 직접 찾을 수 있도록 학습시키는 것

구멍이 2개이고 중간 부분이 홀쭉하며 맨 위와 아래가 둥근 모양이라면 8!

- 기계가 일일이 코드로 명시하지 않은 동작을 데이터로부터 학습하여 실행할 수 있도록 하는 알고리즘을 개발하는 연구 분야라고 정의하기도 함

- 컴퓨터가 학습데이터를 통해 스스로 규칙을 찾을 수 있도록 하는 것이 머신러닝

- 우리 학과에서는 머신러닝을 통해 특정 값을 예측하는 방법을 배우게 됨

1 머신러닝이란?

기존 프로그래밍

- 스팸 메일의 일반적인 형태를 파악
- 스팸 메일로 예상되는 패턴을 컴퓨터에 직접 입력
- 프로그램을 테스트 하면서 위의 과정을 반복

↓

- 모든 패턴을 수동으로 찾는 것은 불가
- 프로그램이 너무 길어짐

[규칙을 사용자가 입력]

머신러닝

- 일반적인 메일과 비교해 스팸 메일의 패턴을 컴퓨터가 학습을 통해 발견함
- Ex) 공짜, 대출 등의 키워드

↓

- 많은 패턴을 발견해 사용 가능
- 프로그램이 짧아 관리에 용이함

[규칙을 컴퓨터가 찾음]

1 머신러닝이란?

머신러닝의 기본 구조

2 Feature

Feature

- 특색, 관찰 대상에게서 발견된 개별적이고 측정가능한 경험적 속성
- 원시 데이터의 숫자적인 표현
- Feature는 모델에 학습시킬 데이터의 특성을 알려주는 정보로 머신러닝에서 모델의 성능을 가장 크게 좌우하는 부분
- 우리가 배운 파이썬 문법을 활용해 feature를 만들어 모델로 학습시키면 찾고자 하는 예측값을 얻을 수 있음
- 크게 범주형, 수치형 데이터로 나뉘고 일반적으로 데이터 전처리를 하는 데이터는 수치형 데이터!

2 Feature

Feature 예시

	cust_id	tran_date	store_nm	goods_id	gds_grp_nm	gds_grp_mclas_nm	amount
0	0	2007-01-19 00:00:00	강남점	127105	기초 화장품	화장품	850000
1	0	2007-03-30 00:00:00	강남점	342220	니 트	시티웨어	480000
2	0	2007-03-30 00:00:00	강남점	127105	기초 화장품	화장품	3000000
3	0	2007-03-30 00:00:00	강남점	342205	니 트	시티웨어	840000
4	0	2007-03-30 00:00:00	강남점	342220	상품군미지정	기타	20000
...	...	...	...	...	...	...	...
163553	5981	2007-01-12 00:00:00	영등포점	50105	일반가공식품	가공식품	209000
163554	5981	2007-01-12 00:00:00	영등포점	50109	상품군미지정	기타	7150
163555	5981	2007-01-12 00:00:00	영등포점	50105	햄	축산가공	9500
163556	5981	2007-01-12 00:00:00	영등포점	50105	상품군미지정	기타	9500
163557	5981	2007-03-16 00:00:00	영등포점	77198	수입식품	차/커피	174800

395562 rows × 7 columns

raw data

	cust_id	총구매액	구매건수	평균구매액	최대구매액
0	0	68282840	74	922741	11264000
1	1	2136000	3	712000	2136000
2	2	3197000	4	799250	1639000
3	3	16077620	44	365400	4935000
4	4	29050000	3	9683333	24000000
...	...	...	...	...	...
5977	5977	82581500	14	5898679	23976000
5978	5978	480000	1	480000	480000
5979	5979	260003790	71	3662025	25750000
5980	5980	88991520	18	4943973	18120000
5981	5981	623700	10	62370	209000

5982 rows × 5 columns

feature

3 머신러닝의 유형

3 머신러닝의 유형 - 지도학습

지도 학습

- 지도 학습은 학습 데이터와 결과가 있을 때 학습하는 방법
- 정답이라 가정한 내용에 맞게 컴퓨터가 예측할 수 있도록 하는 과정
- 지도 학습은 분류와 회귀 문제로 나누어집니다.

분류

- 지도 학습 결과가 이산형, 범주형인 변수
- Ex) 성별 예측, 연령대 예측 등

회귀

- 지도 학습 결과가 연속형인 변수
- Ex) 키 예측, 가격 예측 등

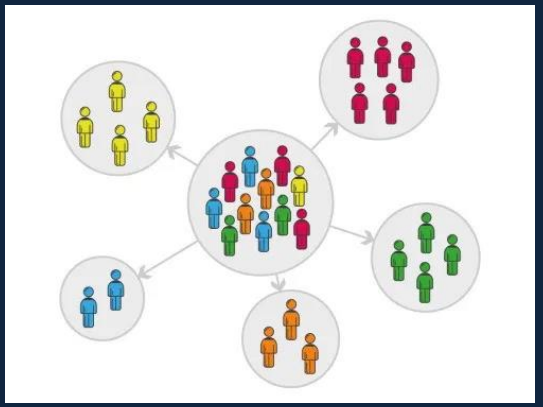
4 머신러닝의 유형 - 비지도학습

비지도 학습

- 비지도 학습은 학습 데이터만 있고 결과가 없을 때 학습하는 방법
- 컴퓨터가 스스로 학습하는 것이라 평가하는 것이 쉽지 않음

군집

- 데이터를 비슷한 것들끼리 모으는 분석 방법
- Ex) 뉴스 기사 분류, 블로그 글의 주제 구분 등



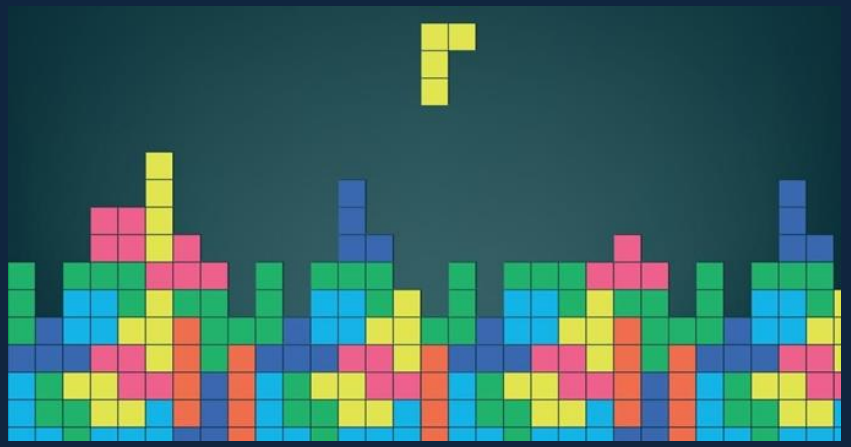
차원 축소

- 변수의 차원을 줄이는 방법
- Ex) 시각화를 위해 데이터를 2차원으로 바꾸는 것, 많은 텍스트에서 주제 찾기 등

5 머신러닝의 유형 - 강화학습

강화 학습

- 강화 학습은 행동에 대한 보상을 받으며 학습하는 것
- 보상을 최대화하는 행동, 행동 순서를 선택하는 방법
- 대표적으로 게임이 있으며 게임에서 이기면 보상이 주어지는 것과 같은 원리



테트리스, 공으로 천장뚫기와 같은 게임들을
학습시키면 성능이 잘 나올 가능성이 높음!

강화학습 예시 영상 – 자동차 게임

6 머신러닝의 과정

stage 1	stage 2	stage 3	stage 4	stage 5	stage 6
Domain Understanding and Data Collection	Data Preprocessing	Modeling and Ensemble	Prediction	Evaluation	Deployment

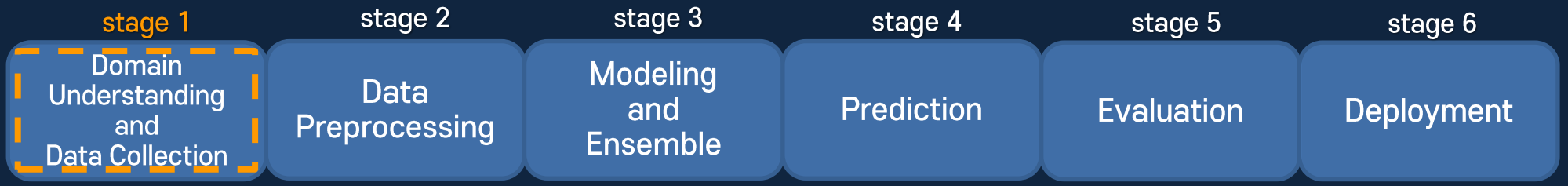
Domain Understanding and Data Collection

- 진행하고자 하는 프로젝트에 대해 전반적으로 이해하고 앞으로의 계획을 세움
- 어떤 데이터를 통해 어떤 예측값을 결과로 만들어낼 지에 대해 이해

	cust_id	tran_date	store_nm	goods_id	gds_grp_nm	gds_grp_mclas_nm	amount
0	0	2007-01-19 00:00:00	강남점	127105	기초 화장품	화장품	850000
1	0	2007-03-30 00:00:00	강남점	342220	니 트	시티웨어	480000
2	0	2007-03-30 00:00:00	강남점	127105	기초 화장품	화장품	3000000
3	0	2007-03-30 00:00:00	강남점	342205	니 트	시티웨어	840000
4	0	2007-03-30 00:00:00	강남점	342220	상품군미지정	기타	20000
...	...	...	...	...	...	...	...
163553	5981	2007-01-12 00:00:00	영등포점	50105	일반가공식품	가공식품	209000
163554	5981	2007-01-12 00:00:00	영등포점	50109	상품군미지정	기타	7150

Ex) 고객의 구매정보를 통해
고객의 성별을 예측하는 프로젝트

6 머신러닝의 과정



Domain Understanding and Data Collection

- 프로젝트를 진행할 데이터를 수집하고 이해하는 단계

- 데이터가 가지고 있는 특성을 파악하고 EDA를 통해 데이터를 분석함

- EDA(탐색적 데이터 분석)는 데이터를 분석하고 결과를 내는 과정에 있어서 지속적으로 해당 데이터에 대한 '탐색과 이해' 를 기본으로 가져야 한다는 것을 의미

- 데이터 분포, 결측값, 이상치 등을 시각화를 통해 확인하면서 데이터를 분석함

- 데이터 자체에 대한 해석이 잘못되면 이후에 진행되는 모든 과정들이 적절한 방향으로 진행될 수 없기 때문에 데이터 전처리 과정, 특히 EDA가 머신러닝에서 매우 중요!

6 머신러닝의 과정

stage 1	stage 2	stage 3	stage 4	stage 5	stage 6
Domain Understanding and Data Collection	Data Preprocessing	Modeling and Ensemble	Prediction	Evaluation	Deployment

Data Preprocessing (데이터 전처리)

- 데이터 전처리 과정으로 머신러닝에서 가장 많은 시간과 노력을 투자해야하는 단계

- 결측값, 이상치를 처리하고 feature를 만듦. 일반적으로 많은 feature를 만들어 두고 유의미하다고 판단되는 feature를 feature selection을 통해 걸러내서 사용

- 모델이 값을 잘 예측할 수 있는 유의미한 feature를 제공해야 성능이 좋은 모델을 만들 수 있기 때문에 유의미한 feature를 만드는 것이 가장 중요!

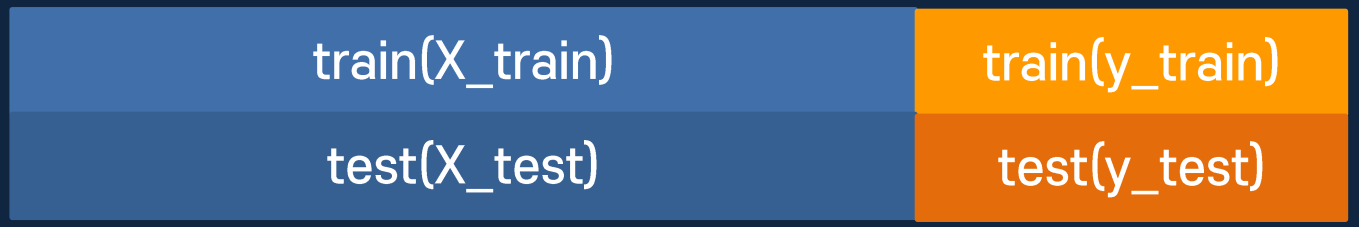
- 예를 들어, 백화점에서 수집한 고객 자료를 토대로 고객의 성별을 예측하는 모델을 만들고 싶을 때 주방용품 구매 비율 (주방용품 구매 횟수/전체 구매 횟수)과 같은 피처를 만들게 되면 모델이 여성을 잘 구분할 확률이 높아질 수 있음

6 머신러닝의 과정

stage 1	stage 2	stage 3	stage 4	stage 5	stage 6
Domain Understanding and Data Collection	Data Preprocessing	Modeling and Ensemble	Prediction	Evaluation	Deployment

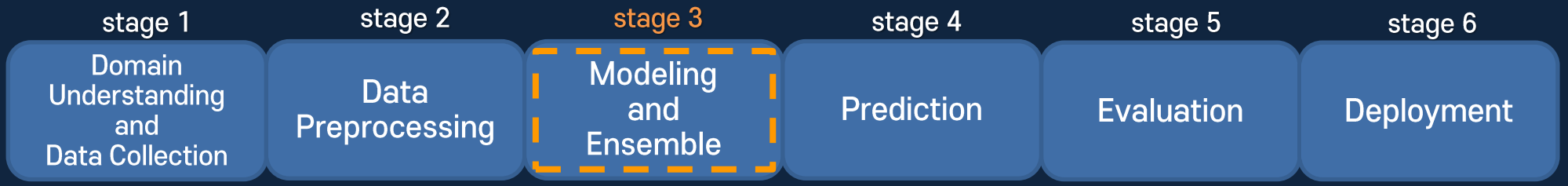
Data Preprocessing (데이터 전처리)

- Feature를 다 만들면 프로젝트에서 설계할 모델을 내부적으로 평가하기 위해 정답으로 가정할 데이터를 분리함
- 일반적으로 가지고 있는 데이터의 7~80%를 학습데이터로, 2~30%를 평가데이터(정답으로 가정할 데이터)로 사용



- 위에서 분리한 test는 우리가 만든 모델의 성능을 자체적으로 평가하기 위한 데이터이고 실제 데이터의 정답은 따로 있음!
- train_test_split(X_train, y_train, test_size=0.3)

6 머신러닝의 과정



Modeling and Ensemble

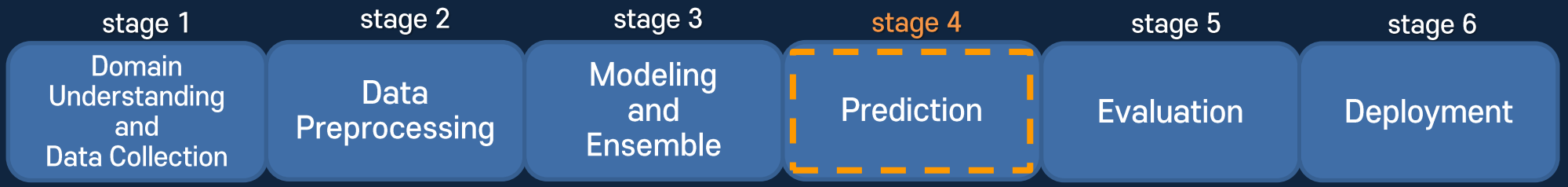
- 데이터에 적합한 모델을 설계하는 과정으로 feature들이 X(입력값)가 되고 예측할 내용은 y(결과값)가 되어 학습함

- 정답이라고 가정한 데이터의 값과 모델을 통해 예측한 값의 차이가 적어질 수 있도록 학습 데이터를 통해 학습을 진행함

- 모델에서 사용되는 하이퍼파라미터를 조정하기도 함 (* 하이퍼파라미터는 모델링할 때 사용자가 직접 세팅해주는 값)

- model.fit(X_train, y_train)

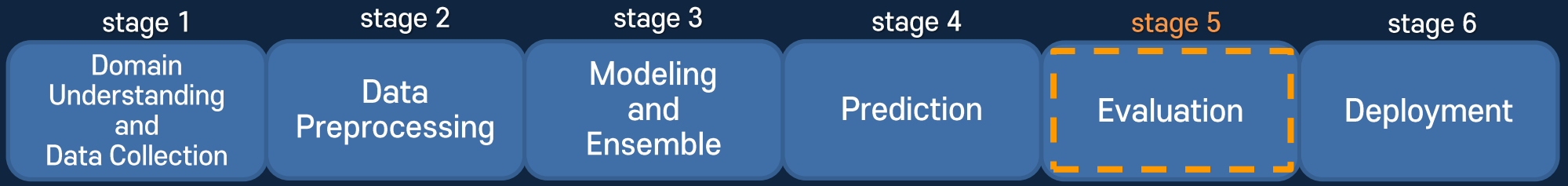
6 머신러닝의 과정



Prediction

- 학습을 진행한 후에 결과값을 예측하는 단계
- X_train과 y_train으로 모델을 학습시켜 값을 예측할 수 있는 모델을 만들 수 있음
- 학습된 모델에 X_test값을 입력값으로 넣어주면 모델이 예측값을 반환
- model.predict(X_test)

6 머신러닝의 과정



Evaluation

- 실제 정답과 모델을 통해 예측한 값의 차이 정도(오차)를 통해 잘 학습된 모델인지 아닌지 평가할 수 있음

- 이 때 과적합에 유의! (과적합 – 과대적합, 과소적합)

- 과대적합은 모델이 예측을 잘 하지만 너무 복잡해 일반성이 떨어진다는 의미이고 훈련한 데이터 외에 다른 데이터를 입력하면 성능이 현저히 떨어지는 모습을 보임

- 과소적합은 모델이 너무 단순해서 데이터의 내재된 구조를 학습하지 못하는 것을 의미

- model.score(X_test, y_test)

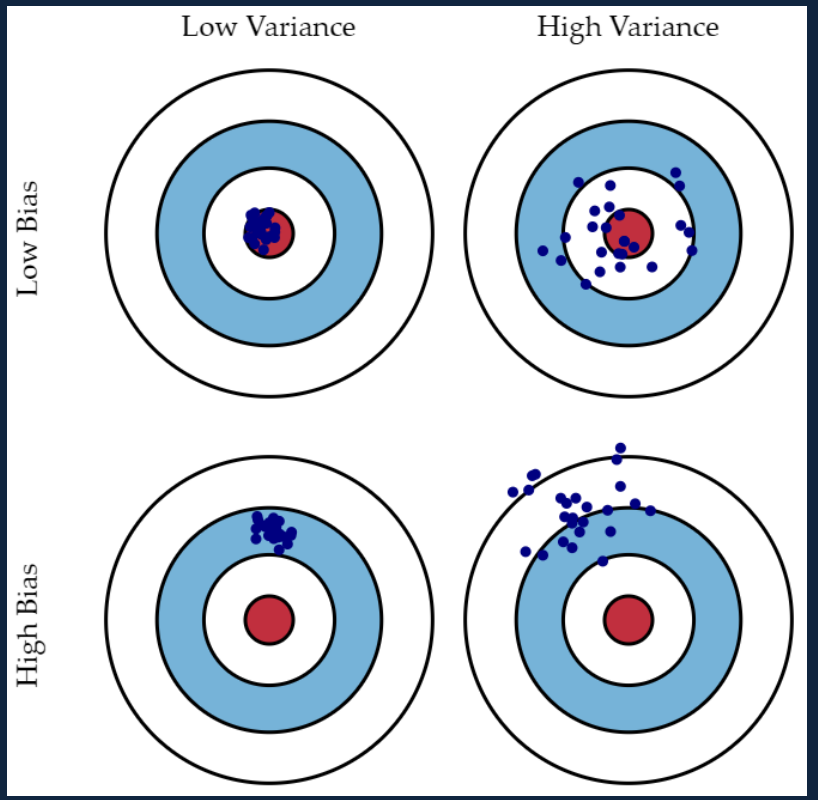
6 머신러닝의 과정

stage 1	stage 2	stage 3	stage 4	stage 5	stage 6
Domain Understanding and Data Collection	Data Preprocessing	Modeling and Ensemble	Prediction	Evaluation	Deployment

Evaluation

- 편향(bias) : 예측이 정답에서 얼마나 떨어져 있는지
- 분산(variance) : 예측의 변동폭이 얼마나 큰지

Ex) 편향이 크면 과소적합, 분산이 크면 과대적합

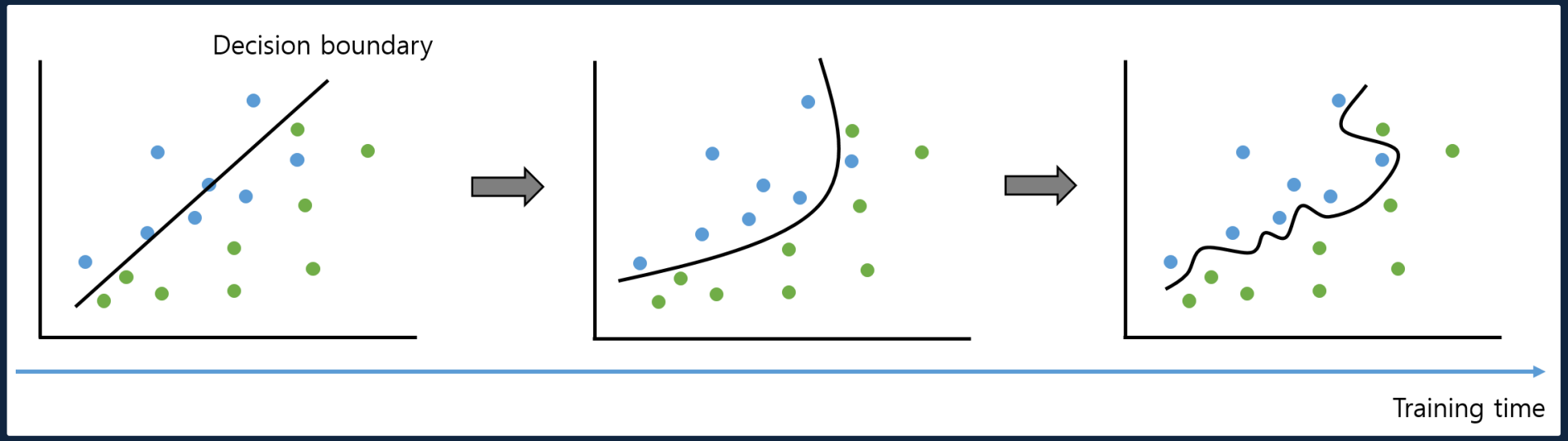


6 머신러닝의 과정

stage 1	stage 2	stage 3	stage 4	stage 5	stage 6
Domain Understanding and Data Collection	Data Preprocessing	Modeling and Ensemble	Prediction	Evaluation	Deployment

Evaluation

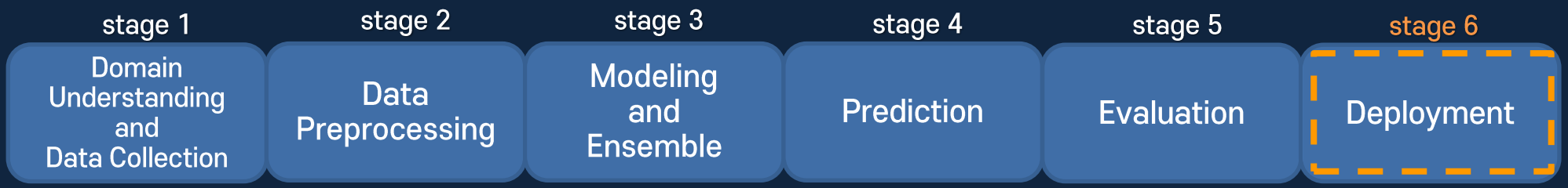
- 과대적합은 보다 단순한 모델을 사용하거나 feature의 수를 줄이는 등의 방법으로 해결



과소적합(underfitting)

과대적합(overfitting)

6 머신러닝의 과정



Deployment

- 머신러닝의 마지막 단계

- 모델을 학습시켰을 때 가장 성능이 좋았던 모델로 예측값을 도출하고 최종 제출할 submission 파일을 생성

7 머신러닝의 한계

- 과적합, 과도한 일반화로 인해 성능 향상에 한계가 있음
- 정답이 있는 대량의 데이터 필요
- 도출 결과의 설명력이 부족
- 기존 학습 모델의 재사용이 어려움 (금융분야에서 학습된 모델을 법률분야에 적용X)

8 과제

파이썬 문법들을 활용해 여러가지
숫자형(numeric) feature 만들어 보기
[10개 이상]

조별로 조 이름과 조장, 스터디 시간, 발표 순서 정해서 스터디 보고서와 함께 구글 드라이브에 업로드하기

+

깃 강의 영상 시청하기

8 과제

PRODUCT

CLNT_ID	고객 id
SESS_ID	세션 id
HITS_SEQ	하트일련번호
PD_C	상품코드
PD_ADD_NM	상품추가정보
PD_BRA_NM	상품브랜드
PD_BUY_AM	단일상품금액
PD_BUY_CT	구매건수

SEARCH

KWD_NM	검색키워드명
SEARCH_CNT	검색건수

MASTER

PD_C	상품코드
PD_NM	상품명
CLAC1_NM	상품대분류명
CLAC2_NM	상품중분류명
CLAC3_NM	상품소분류명

SESSION

SESS_SEQ	세션일련번호
SESS_DT	세션일자
TOP_PAG_VIW_CT	총페이지조회건수
TOT_SESS_HR_V	총세션시간값
DVC_CTG_NM	기기유형
ZON_NM	지역대분류
CITY_NM	지역중분류

THANK YOU