

ML Session 1주차

Index

- 머신러닝
- 데이터 마이닝
- 과대적합 , 과소적합
- 교차 검증
- 과제

머신러닝 이란?

- 말그대로 ‘기계**학습**’
- ‘명시적인 프로그래밍없이 컴퓨터가 학습하는 능력을 갖추게 하는 연구분야’
- 문제풀이 > 답안채점 > 점수측정 > 오답풀이
→ 이 과정을 컴퓨터에게 !

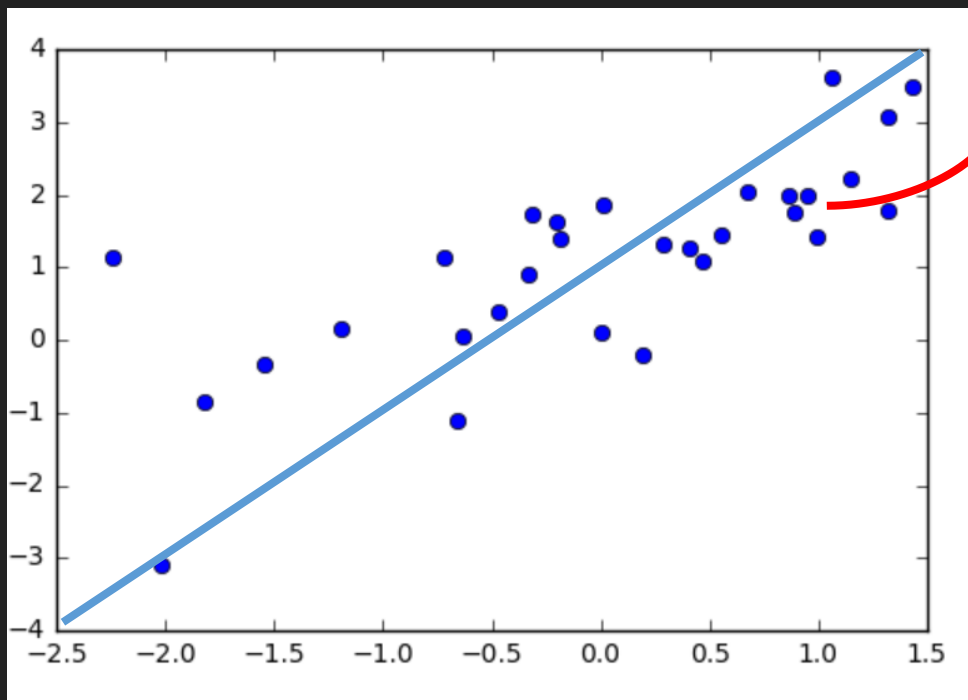
머신러닝 이란?

$$f(x) = 2x + 1$$



**F(X)를 통해 Y에 대한 정확한 예측(prediction)을
하고자 하는 것이 기계학습이다.**

머신러닝 이란?

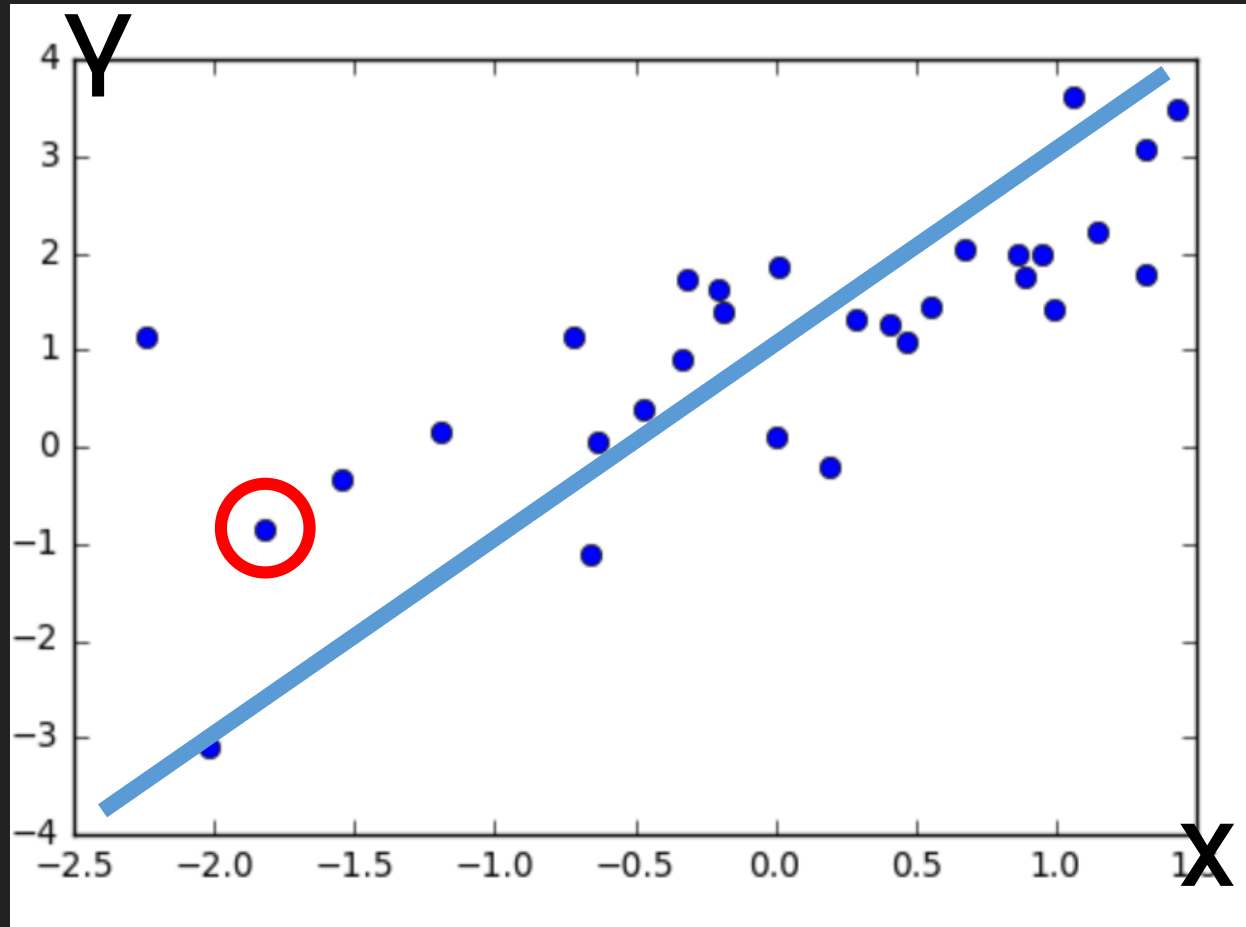


관측치(Data)

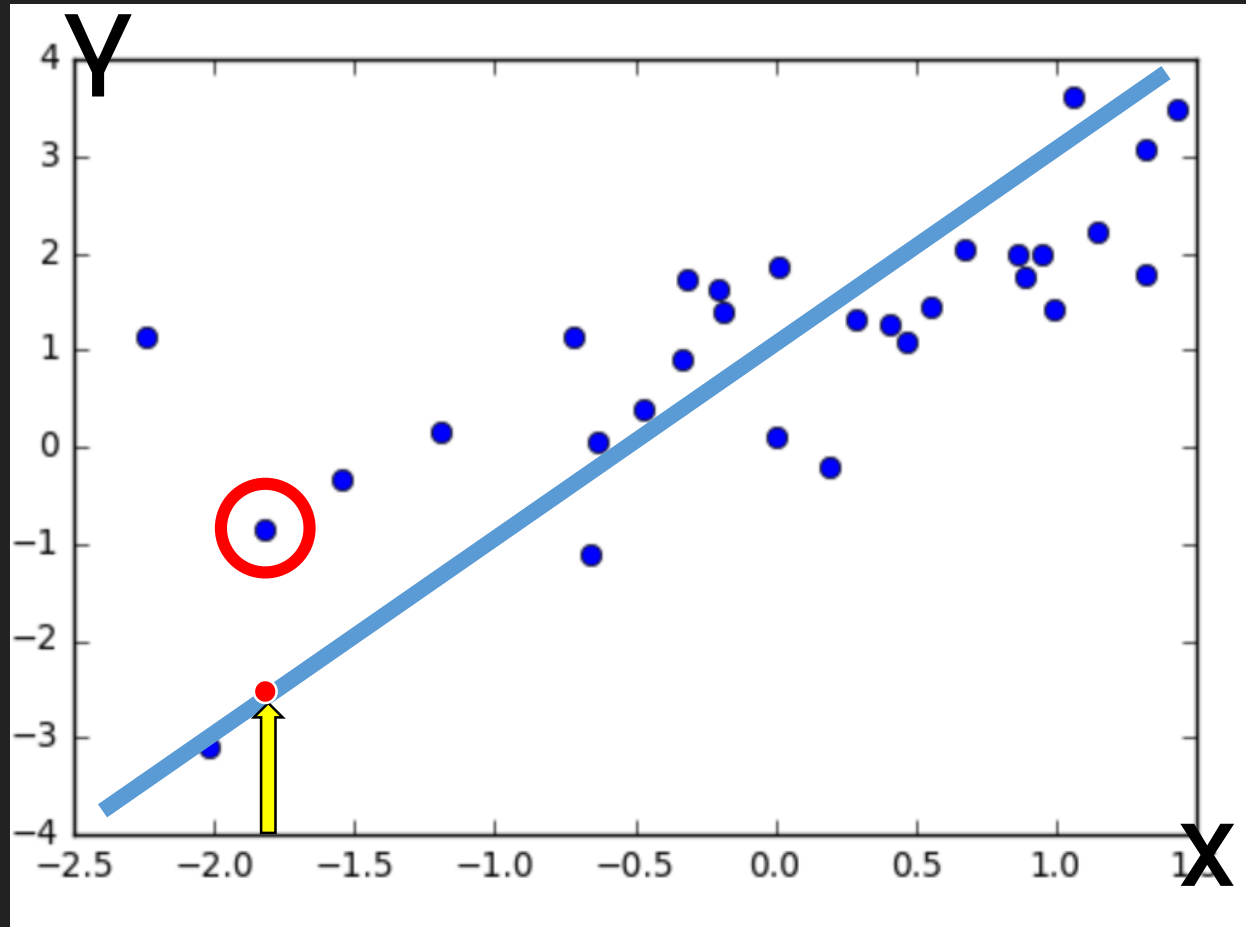
$Y = wX + b$ 로 표현할 수 있음
 w 와 b 를 '파라미터' 라고 정의

$F(x)$ 즉, 데이터의 패턴을 정의 한 것을 **모형(Model)**

머신러닝 이란?



머신러닝 이란?



대용량 데이터에서 보이지 않는 **패턴**을 발견하여
가치있는 지식을 발견

데이터마이닝 순서

데이터 생성 및 추출



전처리

(Preprocessing)



탐색적 데이터분석

(Explortory Data Analysis)



기계학습

(Machine Learning)



해석 및 활용

데이터 마이닝_ 데이터 생성 및 추출

데이터 생성

- Web Crawling
- CSV파일 등..

DBMS 추출

- 데이터모델링과 SQL
- SQL문

데이터 마이닝_ 전처리

꿈

custid	총구매액	총구매건수	평균구매액	최대구매액	총할인액	평균할인액	평균할부기	최대할부기	구입브랜드	브랜드편중	수입상품구
0	1742000	11	158363.6	455000	174200	15836.36	2.818182	3	7	0.363636	0.636364
1	2772100	26	106619.2	393000	56630	2178.077	2.461538	3	19	0.269231	0.423077
2	3750850	11	340986.4	1416000	255090	23190	3.454545	12	7	0.363636	0.090909
3	2300500	30	76683.33	621000	91660	3055.333	2.666667	5	21	0.3	0
4	1045000	4	261250	560000	21800	5450	4.5	10	4	0	0.25
5	5053759	32	157930	682000	361260	11289.38	1.875	3	21	0.34375	0.1875
6	3785029	31	122097.7	936000	315100	10164.52	1.83871	3	23	0.258065	0.096774
7	1223182	35	34948.06	202000	21930	626.5714	1.4	3	20	0.428571	0.085714
8	1267500	18	70416.67	400000	25020	1390	2.111111	3	13	0.277778	0.111111
9	4956620	59	84010.51	395000	213850	3624.576	1	1	35	0.40678	0.101695
10	1347970	24	56165.42	170000	49630	2067.917	1.916667	3	18	0.25	0.166667
11	7173999	66	108697	1780800	124130	1880.758	1.666667	3	19	0.712121	0.121212
12	2595477	28	92695.61	590000	93800	3350	2.214286	3	14	0.5	0.25
13	8789931	129	68139	497000	79950	619.7674	2.178295	3	48	0.627907	0.023256
14	325180	6	54196.67	105000	31500	5250	2	3	5	0.166667	0
15	11780260	87	135405.3	725000	449370	5165.172	2.770115	6	47	0.45977	0.08046
16	5431891	20	271594.6	1960000	114850	5742.5	2.4	3	19	0.05	0.2
17	1148397	23	49930.3	155000	26730	1162.174	1.695652	3	17	0.26087	0.043478
18	9302600	74	125710.8	573000	252600	3413.514	1.972973	3	43	0.418919	0.351351
19	1078340	17	63431.76	278000	29300	1723.529	1.705882	3	15	0.117647	0.117647
20	11422000	169	67585.8	1636000	346120	2048.047	1.639053	3	54	0.680473	0.177515
21	1387995	13	106768.8	231200	11700	900	3.230769	10	5	0.615385	0.076923
22	4649311	32	145291	785000	138550	4329.688	2.625	3	19	0.40625	0.34375
23	4733694	96	49309.31	546000	216960	2249.635	1.739167	12	49	0.489592	0.114592

데이터 마이닝_ 전처리

현실

STD_YM	BLOCK_CD	X_COORD	Y_COORD	TMST_00	TMST_01	TMST_02	TMST_03	TMST_04	TMST_05	TMST_06	TMST_07	TMST_08
201704	111307103000100000001	947233.991787	1953129.147389	10.54	6.78	5.6	4.85	7.9	18.17	44.82	110.29	119.79
201704	111307103000100000001	947233.991787	1953179.147389	3.84	2.49	1.96	1.78	3.64	5.2	8.1	16.99	17.75
201704	111307103000100000001	947233.991787	1953229.147389	12.49	7.32	5.29	4.53	9.29	18.1	32.7	64.44	75.16
201704	111307103000100000001	947283.991787	1953079.147389	3.31	2.14	1.69	1.52	3.09	7.11	19.7	49.28	55.76
201704	111307103000100000001	947283.991787	1953129.147389	1.75	1.12	0.9	0.8	1.56	2.45	4.02	9.08	9.31
201704	111307103000100000001	947283.991787	1953179.147389	9.26	5.87	4.67	4.35	8.31	12.91	22.66	48.58	49.87
201704	111307101010100000001	947283.991787	1953229.147389	5.62	3.42	2.79	2.61	5.21	7.84	13.01	26.04	27.92
201704	111307101010100000001	947283.991787	1953279.147389	20.13	11.53	8.57	7.41	13.59	26.57	46.9	92.24	108.29
201704	111307103000100000001	947333.991787	1953029.147389	5.02	3.66	2.81	2.65	4.3	11.06	33.08	78.65	91.51
201704	111307103000100000001	947333.991787	1953079.147389	7.55	5.17	4.09	3.64	6.97	14.41	36.44	79.68	91.17
201704	111307101010100000001	947333.991787	1953129.147389	4.87	3.12	2.46	2.22	4.58	6.78	10.67	22.69	22.77
201704	111307101010100000001	947333.991787	1953179.147389	3.19	2.05	1.68	1.37	2.75	4.27	6.68	13.36	14.06
201704	111307101010100000001	947333.991787	1953229.147389	0.74	0.42	0.36	0.32	0.66	0.86	1.41	2.95	3.05
201704	111307101010100000001	947333.991787	1953279.147389	0.26	0.16	0.12	0.12	0.26	0.43	0.71	1.52	1.59
201704	111307101000900000001	947333.991787	1953329.147389	11.01	6.54	4.89	4	7.5	13.82	23.76	45.31	52.09
201704	111307103000100000001	947383.991787	1952979.147389	1.88	1.3	1.05	0.92	1.58	4.61	14.67	32.27	37.7
201704	111307103010100000001	947383.991787	1953029.147389	5.89	4.07	3.19	2.69	6.34	14.69	40.98	102.52	115.68
201704	111307103000100000001	947383.991787	1953079.147389	15.03	9.71	7.44	6.62	12.82	19.56	30.65	62.36	63.17
201704	111307101010100000001	947383.991787	1953129.147389	4.32	2.68	1.99	1.8	3.91	5.71	8.24	17.02	19.36
201704	111307101010100000001	947383.991787	1953179.147389	2.78	1.84	1.42	1.29	2.87	3.89	5.28	10.45	10.25

Pre - Processing

- Noise(잡음)
- Outlier(이상치)
- Missing Value(결측치)
- Categorical Variable(범주형 변수)
- ⋮

파생변수 생성 (Feature Engineering)

- ☒ 모델에 입력하기 전 단계에 데이터의 특성을 잘 반영하고 성능을 높일 수 있도록 특징을 생성하고 가공하는 것
- ☒ 특징(Feature)를 만들어내는 과정

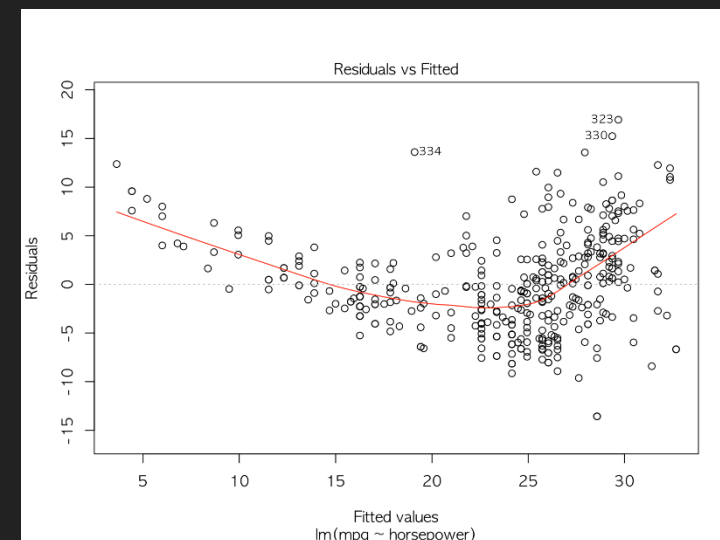
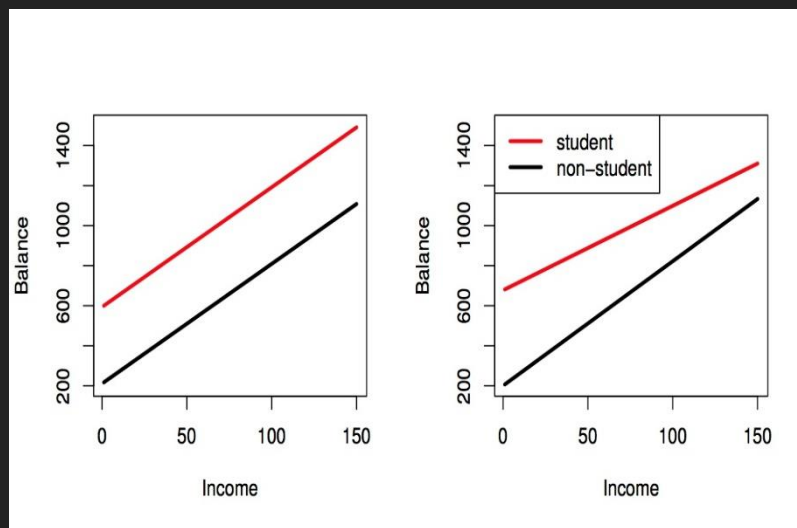
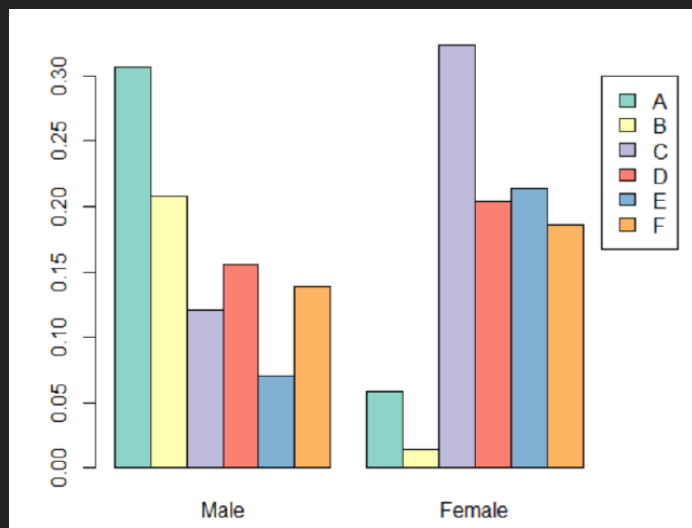
Y : 사용자의 연령대와 성별 ex.20대 여성

x : 웹 페이지 체류시간
웹 페이지 뷰
웹 페이지 카테고리
웹 페이지 이용시간



탐색적 데이터 분석

특정한 가설없이 데이터의 특성을 살펴보는 분석과정



컴퓨터에게 **패턴**을 학습시켜 예측하게 하는 과정

패턴을 정의하는 수학적 방식에 따라 여러 모델(모형)이 생김

데이터 !

입력된 자료를 바탕으로 계산된 **결과의 성격**에 따라

1. 지도 학습 (Supervised Learning)
2. 비지도 학습 (Unsupervised Learning)
3. 강화 학습 (Reinforcement Learning)

학습시 출력되는 결과의 **올바른 답(Y)**을 알고 있는 문제

- ☒ 회귀(Regression) : Y가 연속적인 값 (주택가격 예측, 구매액 예측)
- ☒ 분류(Classification) : Y가 이산적/범주형 값 (성별분류, 스팸여부)

학습시 출력되는 결과의 **답(Y)이 없는** 문제 즉, 데이터의 유사성을 도출

☒ 차원 축소(Dimensionality Reduction)

▶ x를 효과적으로 줄이는 것

☒ 군집(Clustering)

▶ 데이터를 분류 ex.고객 집단 분류

기계학습_ 강화 학습

- ☑ 지도학습과 유사하지만 ‘보상과 처벌’을 통해 학습
- ☑ 정확한 정답(Y)가 아닌 정답이 얼마나 ‘좋은 지’만 아는 경우
- ☑ 예) 구글의 알파고(AlphaGo) : 한 수 한 수 깨우침

기계학습_ 과정

데이터 분할



모형 선택 및 훈련



테스트



평가 및 모형 선택

기계학습_ 데이터 분할

Y : 빨간색
X : 파란색

Data (1)

custid	구내액	내건수	평균구내액	최대구내액	중알인액	평균알인액	평균알무/	최대알무/	구입브랜드	브랜드변동	수입상률
0	1742000	11	158363.6	455000	174200	15836.36	2.818182	3	7	0.363636	0.636364
1	2772100	26	106619.2	393000	56630	2178.077	2.461538	3	19	0.269231	0.423077
2	3750850	11	340986.4	1416000	255090	23190	3.454545	12	7	0.363636	0.090909
3	2300500	30	76683.33	621000	91660	3055.333	2.666667	5	21	0.3	0
4	1045000	4	261250	560000	21800	5450	4.5	10	4	0	0.25
5	5053759	32	157930	682000	361260	11289.38	1.875	3	21	0.34375	0.1875
6	3785029	31	122097.7	936000	315100	10164.52	1.83871	3	23	0.258065	0.096774
7	1223182	35	34948.06	202000	21930	626.5714	1.4	3	20	0.428571	0.085714
8	1267500	18	70416.67	400000	25020	1390	2.111111	3	13	0.277778	0.111111
9	4956620	59	84010.51	395000	213850	3624.576	1	1	35	0.40678	0.101695
10	1347970	24	56165.42	170000	49630	2067.917	1.916667	3	18	0.25	0.166667
11	7173999	66	108697	1780800	124130	1880.758	1.666667	3	19	0.712121	0.121212
12	2595477	28	92695.61	590000	93800	3350	2.214286	3	14	0.5	0.25
13	8789931	129	68139	497000	79950	619.7674	2.178295	3	48	0.627907	0.023256
14	325180	6	54196.67	105000	31500	5250	2	3	5	0.166667	0
15	1780260	87	135405.3	725000	449370	5165.172	2.770115	6	47	0.45977	0.08046
16	5431891	20	271594.6	1960000	114850	5742.5	2.4	3	19	0.05	0.2
17	1148397	23	49930.3	155000	26730	1162.174	1.695652	3	17	0.26087	0.043478
18	9302600	74	125710.8	573000	252600	3413.514	1.972973	3	43	0.418919	0.351351
19	1078340	17	63431.76	278000	29300	1723.529	1.705882	3	15	0.117647	0.117647
20	1422000	169	67585.8	1636000	346120	2048.047	1.639053	3	54	0.680473	0.177515
21	1387995	13	106768.8	231200	11700	900	3.230769	10	5	0.615385	0.076923
22	4649311	32	145291	785000	138550	4329.688	2.625	3	19	0.40625	0.34375

Split

Training Set (0.7)

custid	구내액	구내건수	평균구내액	최대구내액	중알인액	평균알인액	평균알무	최대알무	구입브랜드	브랜드변동수입상률
0	1742000	11	158363.6	455000	174200	15836.36	2.818182	3	7	0.363636 0.636364
1	2772100	26	106619.2	393000	56630	2178.077	2.461538	3	19	0.269231 0.423077
2	3750850	11	340986.4	1416000	255090	23190	3.454545	12	7	0.363636 0.090909
3	2300500	30	76683.33	621000	91660	3055.333	2.666667	5	21	0.3 0
4	1045000	4	261250	560000	21800	5450	4.5	10	4	0 0.25
5	5053759	32	157930	682000	361260	11289.38	1.875	3	21	0.34375 0.1875
6	3785029	31	122097.7	936000	315100	10164.52	1.83871	3	23	0.258065 0.096774
7	1223182	35	34948.06	202000	21930	626.5714	1.4	3	20	0.428571 0.085714
8	1267500	18	70416.67	400000	25020	1390	2.111111	3	13	0.277778 0.111111
9	4956620	59	84010.51	395000	213850	3624.576	1	1	35	0.40678 0.101695
10	1347970	24	56165.42	170000	49630	2067.917	1.916667	3	18	0.25 0.166667
11	7173999	66	108697	1780800	124130	1880.758	1.666667	3	19	0.712121 0.121212
12	2595477	28	92695.61	590000	93800	3350	2.214286	3	14	0.5 0.25
13	8789931	129	68139	497000	79950	619.7674	2.178295	3	48	0.627907 0.023256
14	325180	6	54196.67	105000	31500	5250	2	3	5	0.166667 0

Test Set (0.3)

15	1780260	87	135405.3	725000	449370	5165.172	2.770115	6	47	0.45977	0.08046
16	5431891	20	271594.6	1960000	114850	5742.5	2.4	3	19	0.05	0.2
17	1148397	23	49930.3	155000	26730	1162.174	1.695652	3	17	0.26087	0.043478
18	9302600	74	125710.8	573000	252600	3413.514	1.972973	3	43	0.418919	0.351351
19	1078340	17	63431.76	278000	29300	1723.529	1.705882	3	15	0.117647	0.117647
20	1422000	169	67585.8	1636000	346120	2048.047	1.639053	3	54	0.680473	0.177515
21	1387995	13	106768.8	231200	11700	900	3.230769	10	5	0.615385	0.076923
22	4649311	32	145291	785000	138550	4329.688	2.625	3	19	0.40625	0.34375

컴퓨터가 단순히 데이터를 '외웠'는지 아니면
데이터의 패턴을 잘 깨우쳤는지 구별하기 위해
데이터 중 일부를 테스트용으로 나눠놓는 것이다.

즉, 모델의 일반화를 위해서! (유연성)

기계학습_ 실습코드

y_train

X_train

custid	총구매액	총구매건수	평균구매액	최대구매액	평균할인액	평균할부기	최대할부기	구입브랜드	브랜드편중	수입상품구
0	1742000	11	158363.6	455000	174200	158363.6	2.818182	3	7	0.363636
1	2772100	26	106619.2	393000	56630	2178.077	2.401519	3	19	0.269231
2	3750850	11	340986.4	1416000	255090	23190	3.454545	7	7	0.363636
3	2300500	30	76683.33	621000	91660	3055.333	2.666667	5	21	0.3
4	1045000	4	261250	560000	21800	5450	4.5	10	4	0
5	5053759	32	157930	682000	361260	11289.38	1.875	3	21	0.34375
6	3785029	31	122097.7	500000	21880	3624.576	1	1	35	0.40678
7	1223182	35	34948.06	170000	49630	2067.917	1.916667	3	18	0.25
8	1267500	18	70416.67	400000	25020	1390	2.111111	3	13	0.277778
9	4956620	59	84010.51	395000	213850	3624.576	1	1	35	0.40678
10	1347970	24	56165.42	170000	49630	2067.917	1.916667	3	18	0.25
11	7173999	66	108697	1780800	124130	1880.758	1.666667	3	19	0.712121
12	2595477	28	92695.61	590000	93800	3350	2.214286	3	14	0.5
13	8789931	129	68139	497000	79950	619.7674	2.178295	3	48	0.627907
14	325180	6	54196.67	105000	31500	5250	2	3	5	0.166667

Training Set

① 학습

학습되지 않은 모델

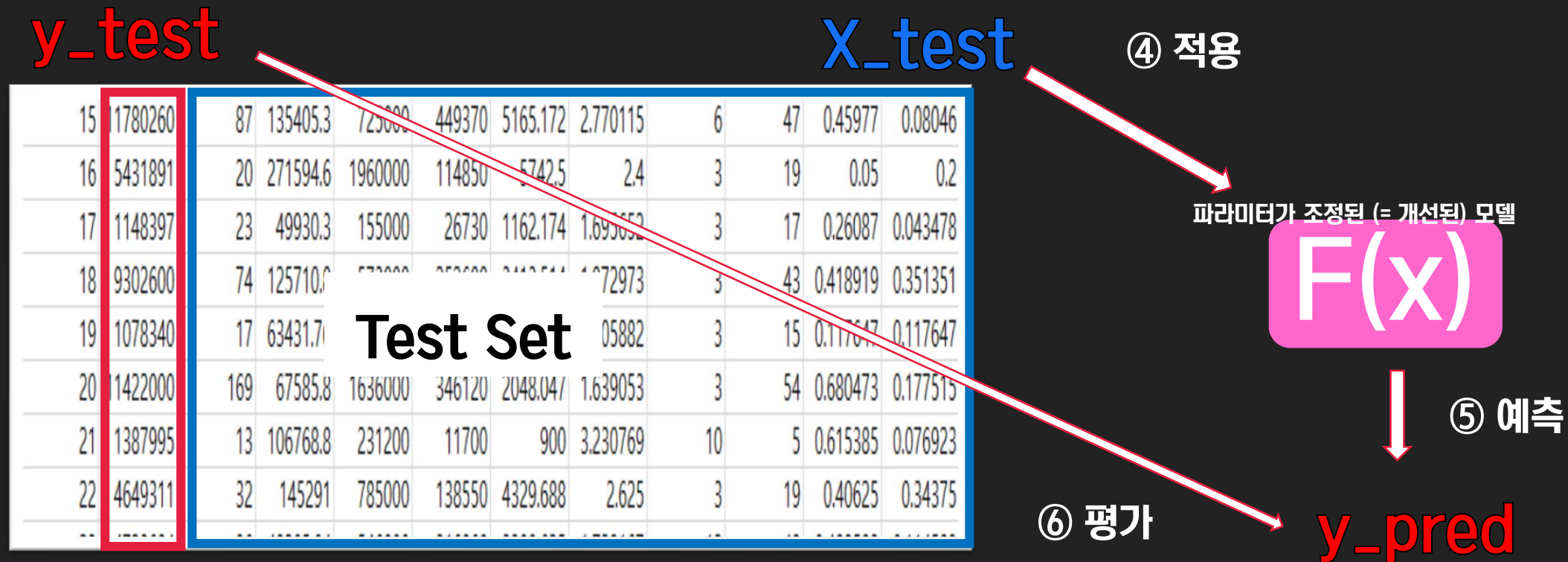
F(x)

② 예측

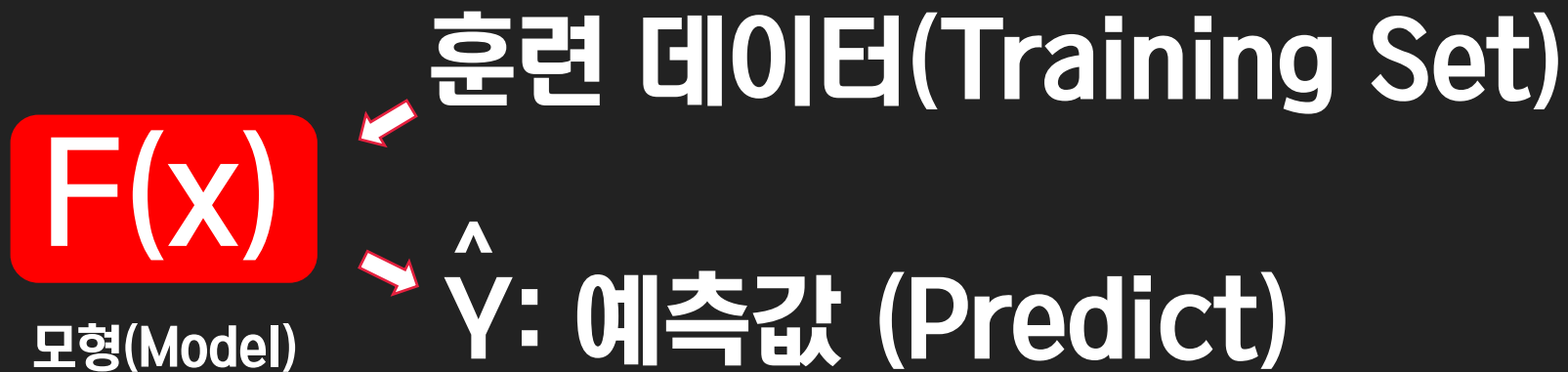
\hat{y}_{train}

③ 평가

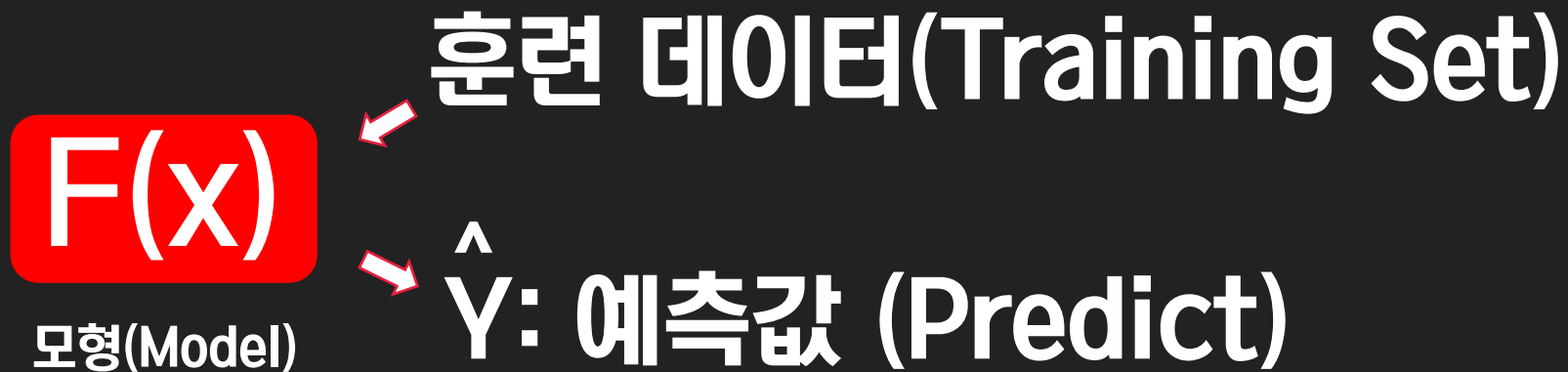
기계학습_ 실습코드



기계학습_ 모형 훈련 및 테스트



기계학습_ 모형 훈련 및 테스트



VS

테스트 데이터의 Y

\hat{Y} : 예측값 (Predict) Y : 테스트 데이터(Test Set)

예측값과 정답사이의 평가

평가지표

$$MSE = \frac{1}{n} \sum (y - \hat{y})^2$$

- 회귀(Y:연속형) : 평균제곱근오차(RMSE), 평균제곱오차(MSE) 등
- 분류(Y:범주형) : 정확도(Accuracy)

정답 맞춘 개수/전체 개수

행렬을 열심히 공부했다.
행렬 문제집도 풀었고 많이 맞았다.

모의고사에는 미분이 나왔다...

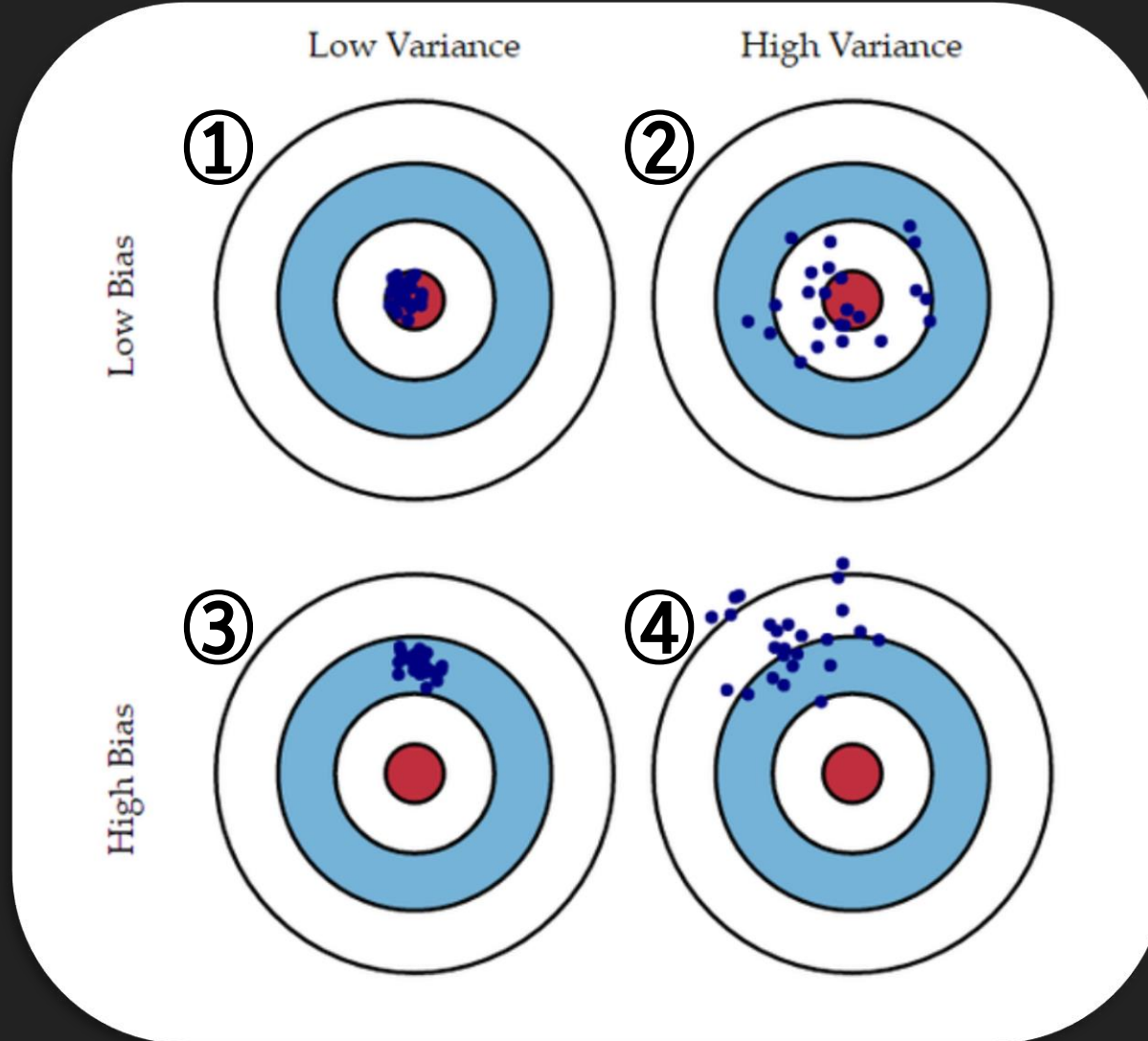
과대적합과 과소적합

☑ Bias(편향) : 예측값과 정답(관측치)의 떨어진 정도
 \hat{Y} Y

☑ Variance(분산) : 예측값이 흩어진 정도
 \hat{Y}

과대적합과 과소적합

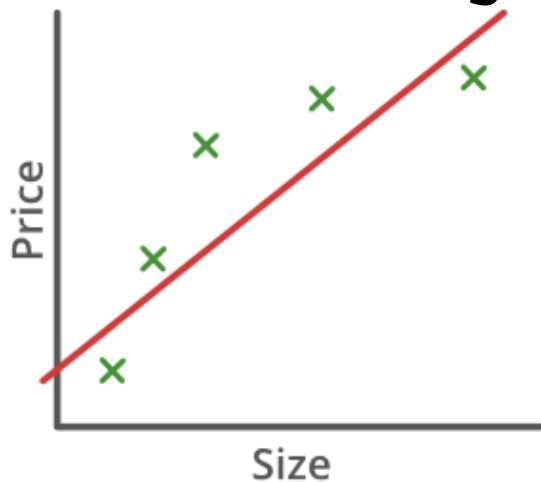
과녁의 **빨간점** 잘 맞추는게 **정답**
파란점은 **예측치**



과대적합과 과소적합

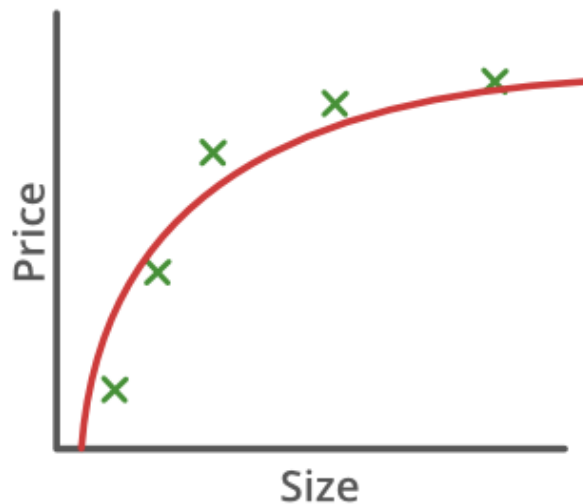
Fitting == 적합시킨다
== 훈련시킨다
== 학습시킨다

Underfitting



$$\theta_0 + \theta_1 x$$

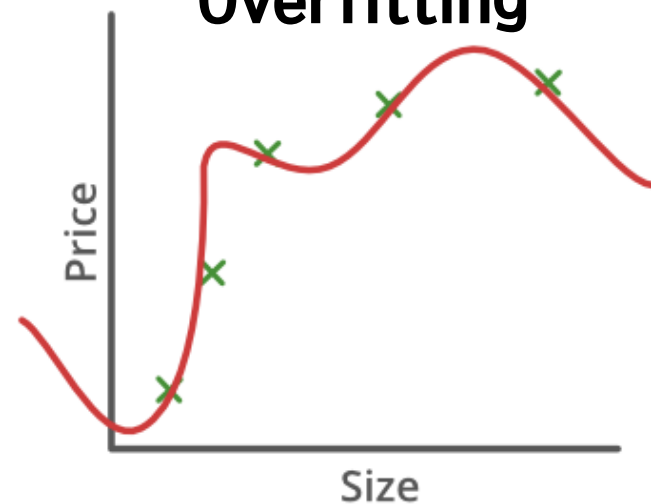
High bias (underfit)



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

High bias (underfit)

Overfitting



$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_2 x^2 + \theta_2 x^2$$

High variance
(overfit)



행렬을 열심히 공부했다.
행렬 문제집도 풀었고 많이 맞았다.

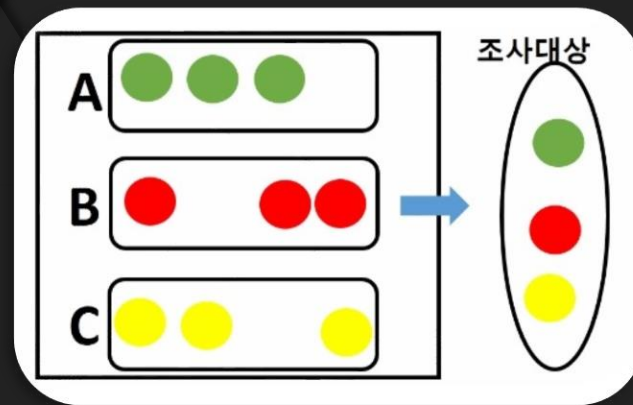
모의고사에는 미분이 나왔다...

머신러닝에서도 앞선 상황이 발생하는 이유는 ?

과대적합과 과소적합

1. 데이터 분할이 랜덤하게 (무작위 추출) 이루어졌다.
2. 데이터 분할이 단 한 번만 이루어졌다.
(즉, 테스트 셋이 하나다.)

과대적합과 과소적합



∴ 증화추출법

1. 데이터 분할이 랜덤하게 (**무작위 추출**) 이루어졌다.
2. 데이터 분할이 단 한 번만 이루어졌다.
(즉, 테스트 셋이 하나다.)

과대적합과 과소적합

1. 데이터 분할이 랜덤하게 (무작위 추출) 이루어졌다.
2. 데이터 분할이 **단 한 번**만 이루어졌다.
(즉, 테스트 셋이 하나다.

∴ 검증세트 추가

교차검증

Y : 빨간색
X : 파란색

Training Set

custid	총구매액	구매건수	평균구매액	최대구매액	중알인액	평균알인액	평균알부	최대알부	구입브랜드	브랜드변동수	입상률
0	1742000	11	158363.6	455000	174200	15836.36	2.818182	3	7	0.363636	0.636364
1	2772100	26	106619.2	393000	56630	2178.077	2.461538	3	19	0.269231	0.423077
2	3750850	11	340986.4	1416000	255090	23190	3.454545	12	7	0.363636	0.090909
3	2300500	30	76683.33	621000	91660	3055.333	2.666667	5	21	0.3	0
4	1045000	4	261250	560000	21800	5450	4.5	10	4	0	0.25
5	5053759	32	157930	682000	361260	11289.38	1.875	3	21	0.34375	0.1875
6	3785029	31	122097.7	936000	315100	10164.52	1.83871	3	23	0.258065	0.096774
7	1223182	35	34948.06	202000	21930	626.5714	1.4	3	20	0.428571	0.085714
8	1267500	18	70416.67	400000	25020	1390	2.111111	3	13	0.277778	0.111111
9	4956620	59	84010.51	395000	213850	3624.576	1	1	35	0.40678	0.101695
10	1347970	24	56165.42	170000	49630	2067.917	1.916667	3	18	0.25	0.166667
11	7173999	66	108697	1780800	124130	1880.758	1.666667	3	19	0.712121	0.121212
12	2595477	28	92695.61	590000	93800	3350	2.214286	3	14	0.5	0.25
13	8789931	129	68139	497000	79950	619.7674	2.178295	3	48	0.627907	0.023256
14	5419667	6	54196.67	105000	31500	5250	2	3	5	0.166667	0

Training Set

custid	총구매액	구매건수	평균구매액	최대구매액	중알인액	평균알인액	평균알부	최대알부	구입브랜드	브랜드변동수	입상률
0	1742000	11	158363.6	455000	174200	15836.36	2.818182	3	7	0.363636	0.636364
1	2772100	26	106619.2	393000	56630	2178.077	2.461538	3	19	0.269231	0.423077
2	3750850	11	340986.4	1416000	255090	23190	3.454545	12	7	0.363636	0.090909
3	2300500	30	76683.33	621000	91660	3055.333	2.666667	5	21	0.3	0
4	1045000	4	261250	560000	21800	5450	4.5	10	4	0	0.25
5	5053759	32	157930	682000	361260	11289.38	1.875	3	21	0.34375	0.1875
6	3785029	31	122097.7	936000	315100	10164.52	1.83871	3	23	0.258065	0.096774
7	1223182	35	34948.06	202000	21930	626.5714	1.4	3	20	0.428571	0.085714
8	1267500	18	70416.67	400000	25020	1390	2.111111	3	13	0.277778	0.111111
9	4956620	59	84010.51	395000	213850	3624.576	1	1	35	0.40678	0.101695
10	1347970	24	56165.42	170000	49630	2067.917	1.916667	3	18	0.25	0.166667

Split

Validation Set (검증세트)

custid	총구매액	구매건수	평균구매액	최대구매액	중알인액	평균알인액	평균알부	최대알부	구입브랜드	브랜드변동수	입상률
2	2595477	28	92695.61	590000	93800	3350	2.214286	3	14	0.5	0.25
3	8789931	129	68139	497000	79950	619.7674	2.178295	3	48	0.627907	0.023256
4	3251000	6	54196.67	105000	31500	5250	2	3	5	0.166667	0

Test Set

custid	총구매액	구매건수	평균구매액	최대구매액	중알인액	평균알인액	평균알부	최대알부	구입브랜드	브랜드변동수	입상률
15	1780260	81	135405.3	725000	449370	5165.172	2.770115	6	41	0.45977	0.08046
16	5431891	20	271594.6	1960000	114850	5742.5	2.4	3	19	0.05	0.2
17	1148397	23	49930.3	155000	26730	1162.174	1.695652	3	17	0.26087	0.043478
18	9302600	74	125710.8	573000	252600	3413.514	1.972973	3	43	0.418919	0.351351
19	1078340	17	63431.76	278000	29300	1723.529	1.705882	3	15	0.117647	0.117647
20	1422000	169	67585.8	1636000	346120	2048.047	1.639053	3	54	0.680473	0.177515
21	1387995	13	106768.8	231200	11700	900	3.230769	10	5	0.615385	0.076923
22	4649311	32	145291	785000	138550	4329.688	2.625	3	19	0.40625	0.34375

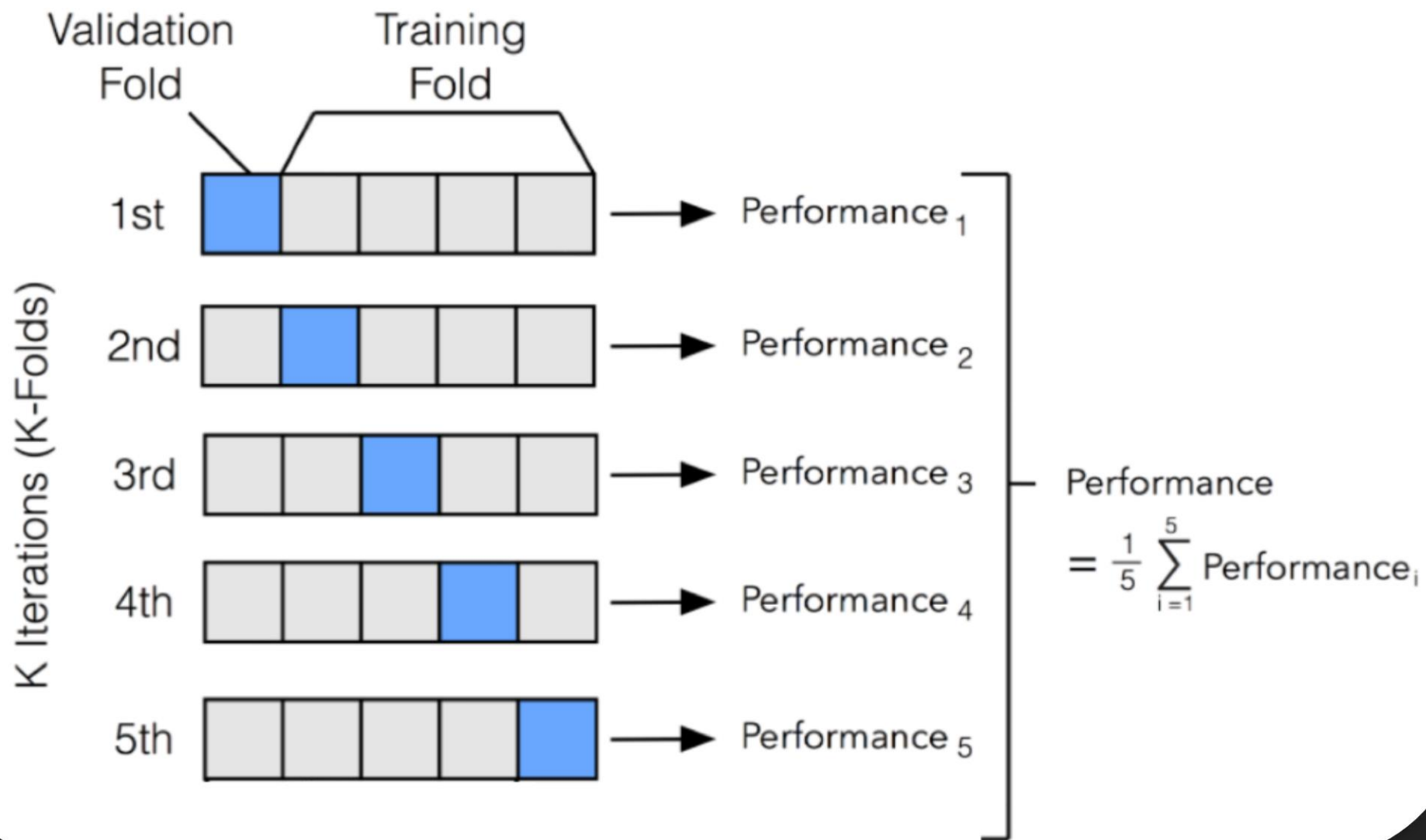
K- fold 교차검증 (K – fold Cross Validation)

K- fold 교차검증 (K – fold Cross Validation)

: 훈련용 데이터를 K등분하고 그 중 한 세트를 검증용으로 사용, 나머지 세트를 훈련용 세트로 활용, 이 과정을 K번 반복한다.

∴ 한 번 학습할 걸 K번 학습할 수 있게 됨!

K- fold 교차검증 (K – fold Cross Validation)



1. 편향(Bias)과 분산(Variance)이 동시에 줄어들 수 없는 이유(Bias-Variance TradeOff)를 조사해서 레포트 작성

* 꼭 **수식적 설명**이 포함되어야 함 !

2. 지도학습(회귀/분류)의 모델과 모델에 대한 간단한 설명이 포함된 레포트 작성

파일명: ML_1주차_홍길동

제출형식 : PDF

THANK YOU FOR YOUR ATTENTION