예측 모델 학습 프로세스

김성범

고려대학교
KOREA UNIVERSITY

(다변량) 데이터

Y (결과): 종속변수, 반응변수, 출력변수

X (원인): 독립변수, 예측변수, 입력변수

변수 / 관측치	X_1	…	X_i	…	X_p	Y
N_1	x_{11}	…	x_{1i}	…	x_{1p}	20.5
N_2	x_{21}	…	x_{2i}	…	x_{2p}	22.2
…	…	…	…	…	…	…
N_{n-1}	$x_{n-1\,1}$	…	$x_{n-1\,i}$	…	$x_{n-1\,p}$	72.3
N_n	x_{n1}	…	x_{ni}	…	x_{np}	82.8

많은 현상을 X와 Y로 설명할 수 있어...

어떤 고객들이 이탈할까?

고장을 미리 예측 할 수 있을까?

최적의 투자전략은 무엇인가?

식품 판매량 (수요) 예측?

보험 과다 청구 여부?

출시 예정 상품이 시장에서 어떤 반응을 보일까?

핵심 아이디어...

X와 Y의 관계를 찾는 것!

우리의 주 관심은 Y (예측하려는 대상)

Y를 설명하는 X변수는 보통 여러 개

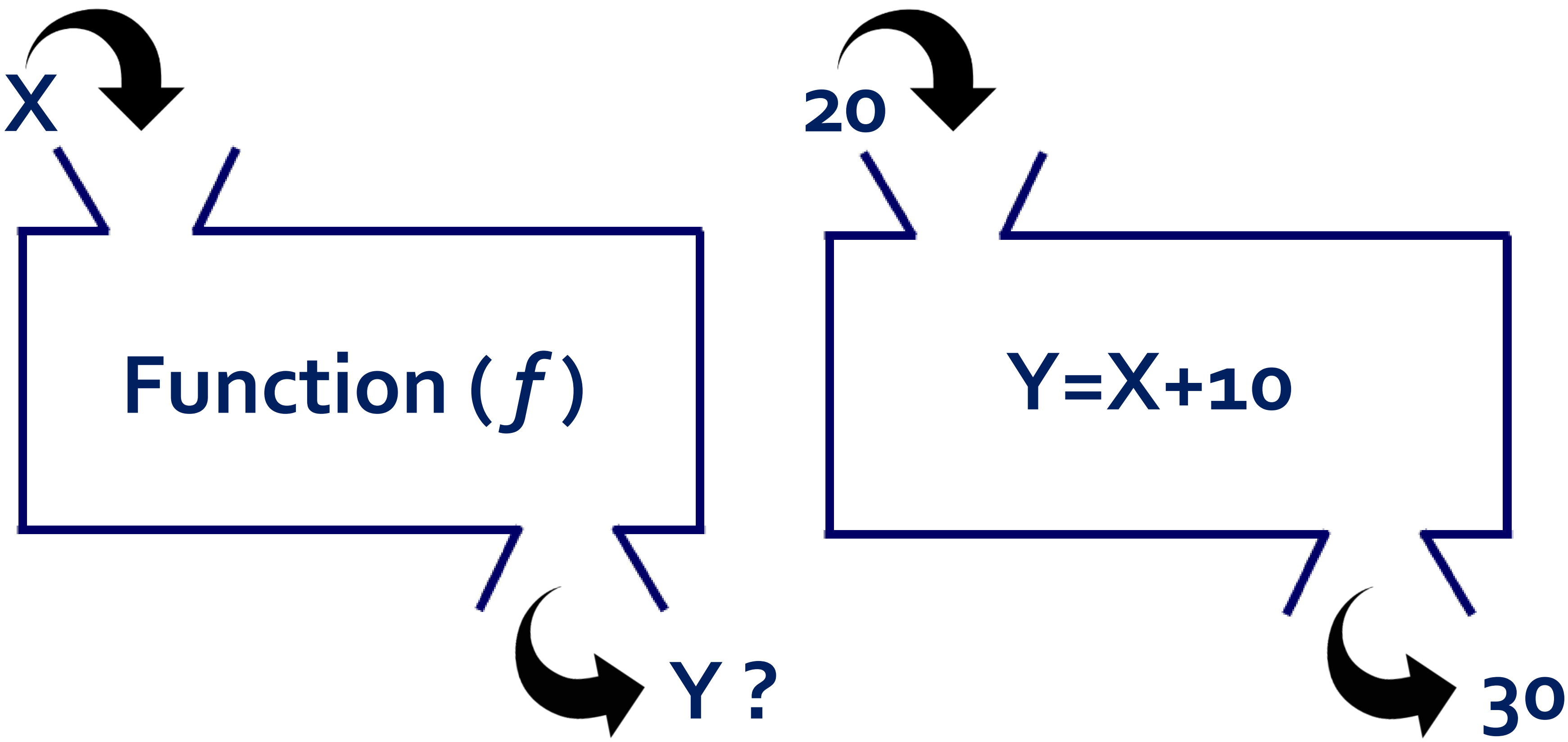
여러 개의 X와 Y의 관계를 찾는 것!

X변수들을 조합(결합)하여 Y를 표현

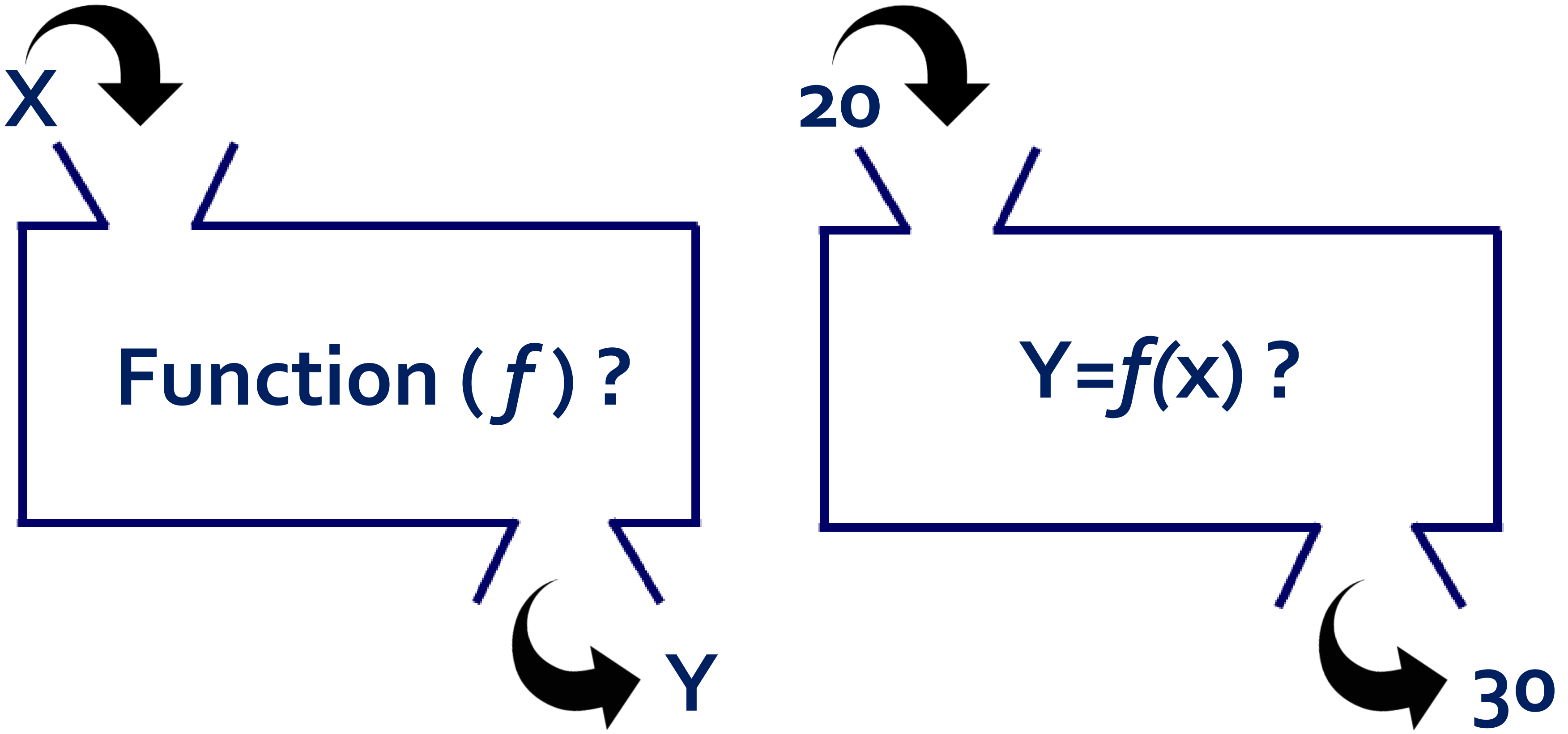
조합하는 방법은 무수히 많음

수학적으로는, $Y = f(X_1, X_2, \ldots, X_p)$

X와 Y의 관계 찾기

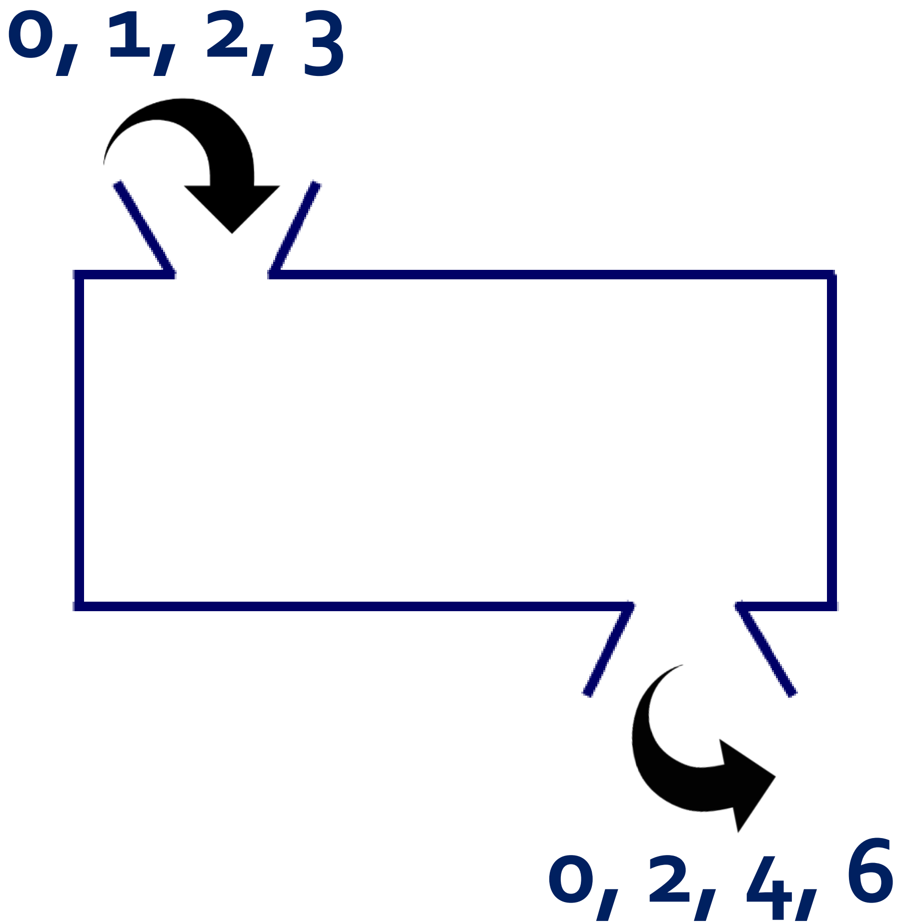


X와 Y의 관계 찾기



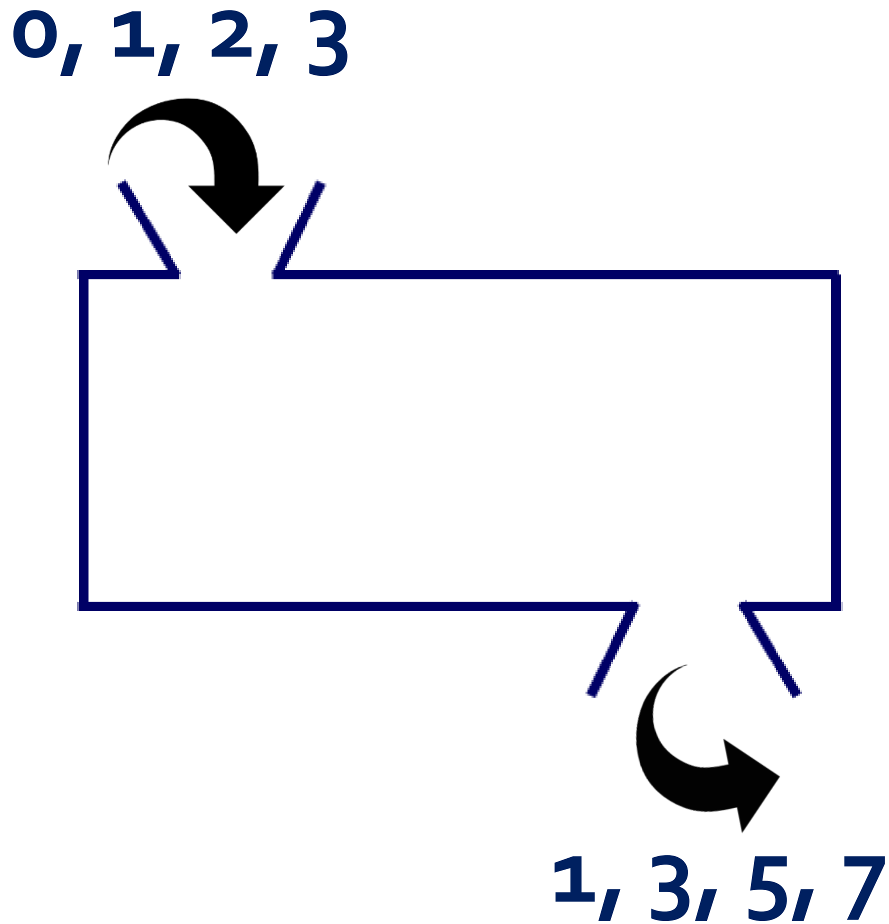
X와 Y의 관계 찾기

X	Y
0	0
1	2
2	4
3	6



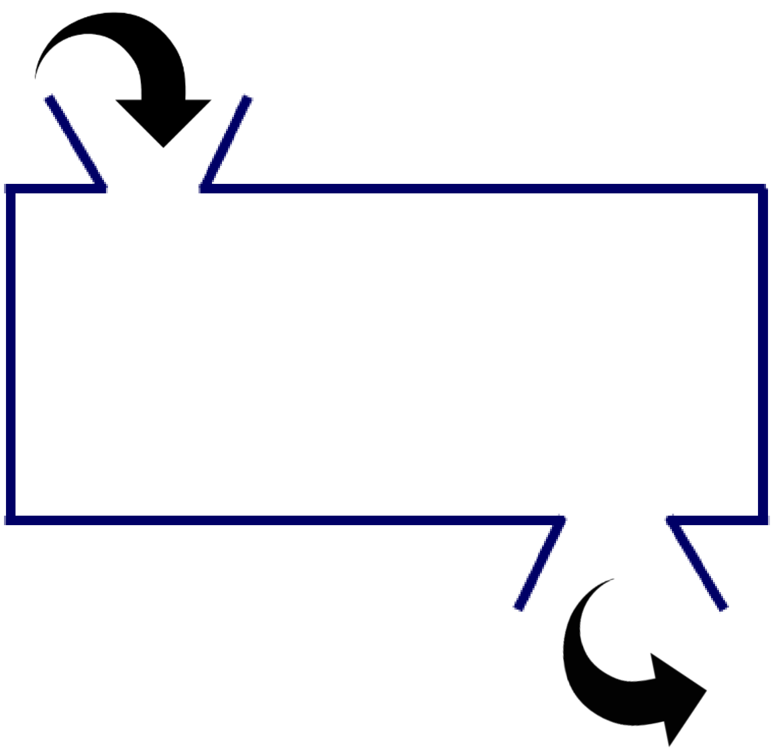
X와 Y의 관계 찾기

X	Y
0	1
1	3
2	5
3	7



X와 Y의 관계 찾기

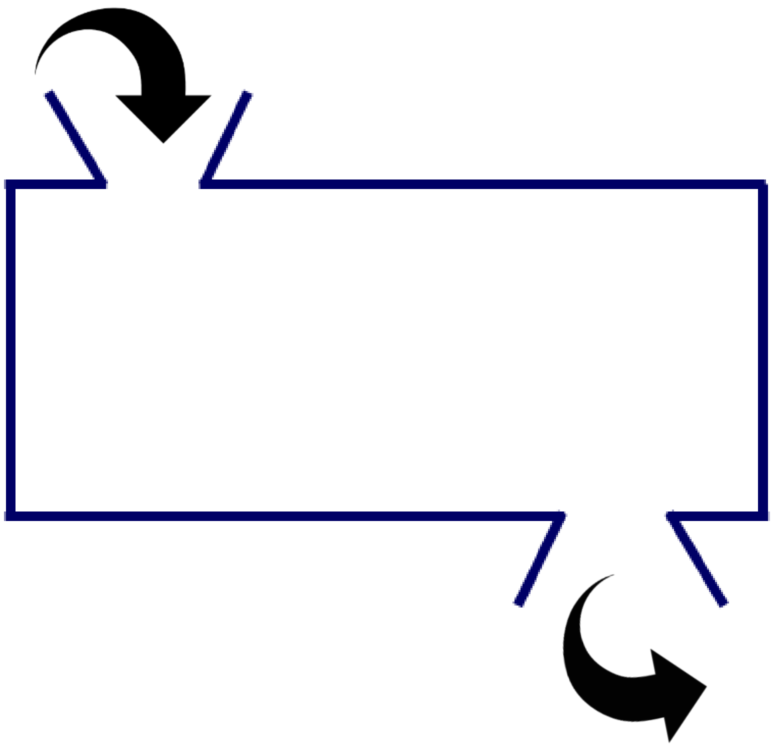
X	Y
0	2
1	2.5
2	3
3	3.5



X와 Y의 관계 찾기

X_1	X_2	Y
0	2	2
1	3	4
2	4	6
3	5	8

(0, 2), (1,3), (2,4), (3,5)

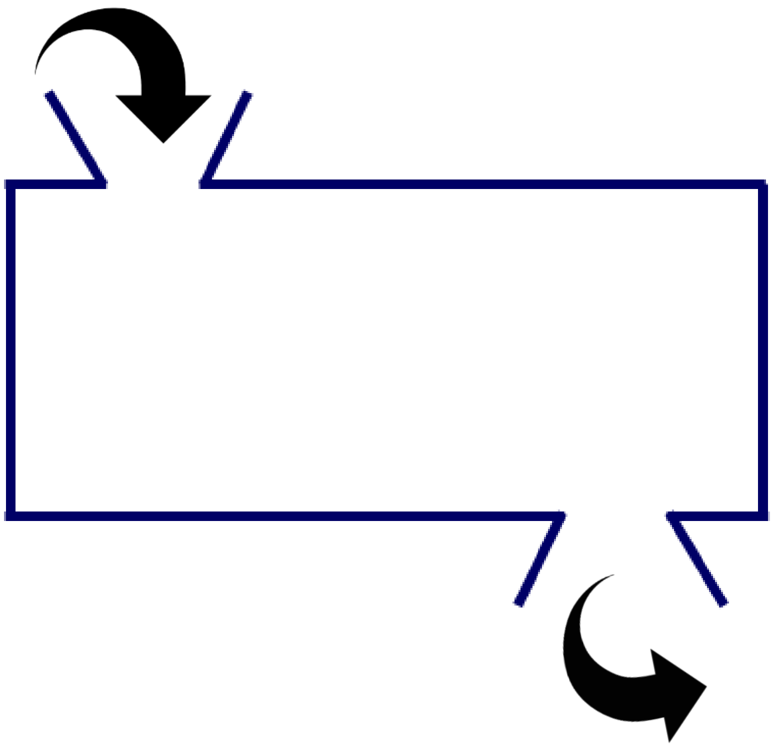


2, 4, 6, 8

X와 Y의 관계 찾기

X_1	X_2	Y
0	2	6
1	3	9.5
2	4	13
3	5	16.5

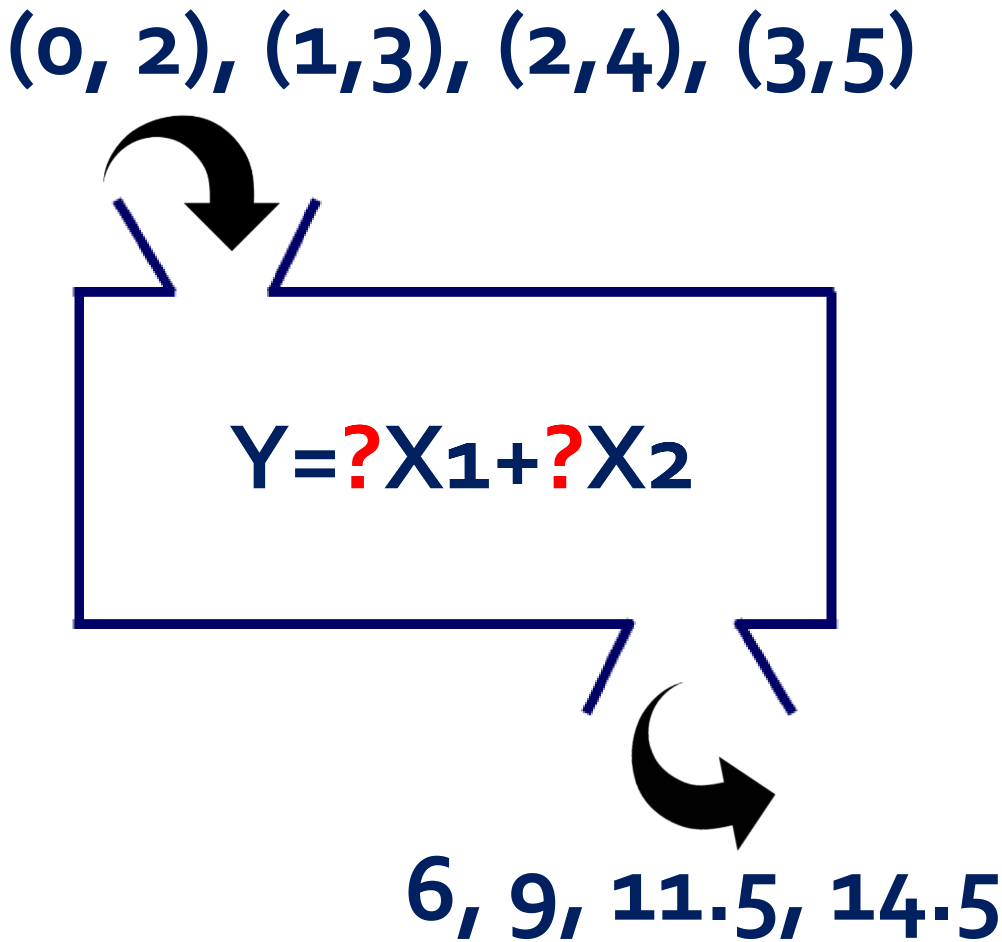
(0, 2), (1,3), (2,4), (3,5)



6, 9.5, 13, 16.5

X와 Y의 관계 찾기

X_1	X_2	Y
0	2	6
1	3	9
2	4	11.5
3	5	14.5



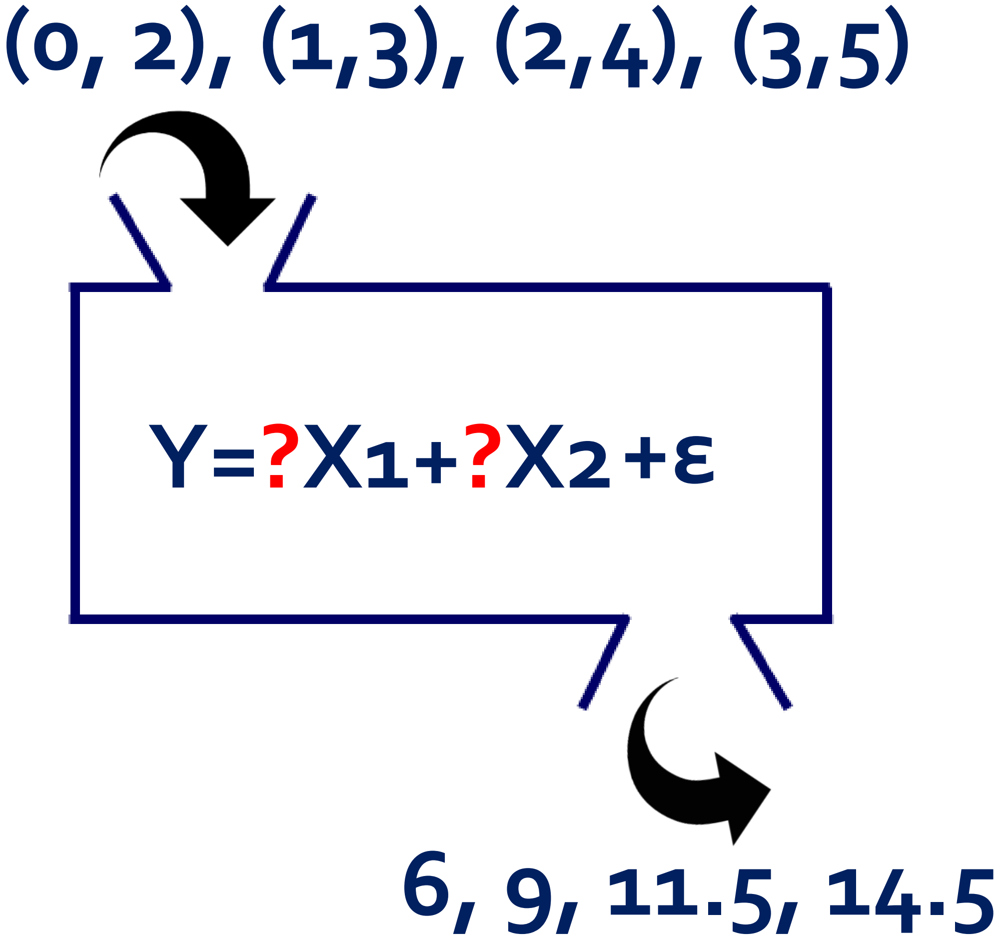
X와 Y의 관계 찾기

모델	주행거리 (X_1)	마력 (X_2)	용량 (X_3)	가격 (Y)
TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	46,986	90	2,000	13,500
TOYOTA Corolla 1800 T SPORT VVT I 2/3-Doors	19,700	192	1,800	21,500
TOYOTA Corolla 1.9 D HATCHB TERRA 2/3-Doors	71,138	69	1,900	12,950
TOYOTA Corolla 1.8 VVTL-i T-Sport 3-Drs 2/3-Doors	31,461	192	1,800	20,950
TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT BNS 2/3-Doors	43,610	192	1,800	19,950
TOYOTA Corolla 1.6 VVTI Linea Terra Comfort 2/3-Doors	21,716	110	1,600	17,950
TOYOTA Corolla 1.6 16v L.SOL 2/3-Doors	25,563	110	1,600	16,750
TOYOTA Corolla 1.6 16V VVT I 3DR TERRA 2/3-Doors	64,359	110	1,600	16,950
TOYOTA Corolla 1.6 16V VVT I 3DR SOL AUT4 2/3-Doors	43,905	110	1,600	16,950
TOYOTA Corolla 1.6 16V VVT I 3DR SOL 2/3-Doors	56,349	110	1,600	15,950
TOYOTA Corolla 1.4 VVTI Linea Terra 2/3-Doors	9,750	97	1,400	12,950
TOYOTA Corolla 1.4 16V VVT I 3DR 2/3-Doors	27,500	97	1,400	14,750
TOYOTA Corolla 1.4 16V VVT I 3DR 2/3-Doors	49,059	97	1,400	13,950
TOYOTA Corolla 1.4 16V VVT I 3DR 2/3-Doors	44,068	97	1,400	16,750
TOYOTA Corolla 1.4 16V VVT I 3DR 2/3-Doors	46,961	97	1,400	13,950
TOYOTA Corolla 2.0 D4D 90 5DR TERRA COMFORT 4/5-Doors	110,404	90	2,000	16,950
TOYOTA Corolla 2.0 D4D 90 5DR TERRA COMFORT 4/5-Doors	100,250	90	2,000	16,950
TOYOTA Corolla 2.0 D4D 90 5DR SOL 4/5-Doors	84,000	90	2,000	19,000
TOYOTA Corolla 2.0 D4D 90 5DR TERRA 4/5-Doors	79,375	90	2,000	17,950
TOYOTA Corolla 1.4 16V VVT I 5DR TERRA COMFORT 4/5-Doors	75,048	97	1,400	15,800

$$Y = ?X_1 + ?X_2 + ?X_3 + \varepsilon$$

X와 Y의 관계 찾기

X_1	X_2	Y
0	2	6
1	3	9
2	4	11.5
3	5	14.5



머신러밍 모델 핵심

$$Y = ?X_1 + ?X_2 + \varepsilon$$

$$Y = w_1X_1 + w_2X_2 + \varepsilon$$

$$w_1? \quad w_2?$$

Given X_1, X_2, Y (데이터)

$$Y = W_1 X1 + W_2 X2 + \varepsilon$$

파라미터 (母數) (媒介變數)

데이터가 주어졌을 때 모델의 파라미터 찾기!

파라미터 추정

$$Y = w_1 X1 + w_2 X2 + \varepsilon$$

$$= f(X) + \varepsilon$$

$$\varepsilon = Y - f(X)$$ ➡ 오차

Loss function
(손실함수)

$$Y - f(X) = 0, \varepsilon = 0$$

파라미터 추정

$$\varepsilon = Y - f(X)$$

Loss function
(손실함수)

$$f(X) = w_1 X1 + w_2 X2 + \varepsilon$$

X1	X2	Y
0	2	2
1	3	4
2	4	6
3	5	8

$$\varepsilon = Y - (w_1 X1 + w_2 X2)$$

$$\varepsilon_i = Y_i - (w_1 X1_i + w_2 X2_i),\ i = 1, 2, \ldots, n$$

파라미터 추정

$$\varepsilon_i = Y_i - (w_1 X1_i + w_2 X2_i),\ i=1,2,\ldots,n$$

$$\sum_{i=1}^{n} \{Y_i - (w_1 X_{1i} + w_2 X_{2i})\} = 0$$

$$\sum_{i=1}^{n} \{Y_i - (w_1 X_{1i} + w_2 X_{2i})\}^2$$

Cost function
(비용함수)

파라미터 추정

$$\sum_{i=1}^{n}\{Y_i - (w_1 X_{1i} + w_2 X_{2i})\}^2$$

Cost function
(비용함수)

비용함수를 최소로 하는 w_1와 w_2를 찾자!

$$\min_{w_1, w_2} \sum_{i=1}^{n}\{Y_i - (w_1 X_{1i} + w_2 X_{2i})\}^2$$

파라미터 추정

$$\min_{w_1, w_2} \sum_{i=1}^{n} \{Y_i - (w_1 X_{1i} + w_2 X_{2i})\}^2$$

답: $\hat{w}_1, \hat{w}_2$

$$\hat{f}(X) = \hat{w}_1 X_{1i} + \hat{w}_2 X_{2i}$$

모델 결정 → 파라미터 추정

$$\min_{w_1, w_2} \sum_{i=1}^{n} \{Y_i - (w_1 X_{1i} + w_2 X_{2i})\}^2$$

$$f(X)$$

$$f(X) = w_0 + w_1 X_1 + w_2 X_2$$

다중선형회귀 모델

$$f(X) = \frac{1}{1 + e^{-(w_0 + w_1 X_1 + w_2 X_2)}}$$

로지스틱회귀 모델

$$f(X) = \sum_{m=1}^{n} k(m) I\{(x_1, x_2) \in R_m\}$$

의사결정나무 모델

$$f(X) = \frac{1}{1 + exp\left(-\left(w_0 + w_1\left(\frac{1}{1 + e^{-(w_{01} + w_{11} X_1 + w_{21} X_2)}}\right)\right) + w_2\left(\frac{1}{1 + e^{-(w_{02} + w_{12} X_1 + w_{22} X_2)}}\right)\right)}$$

뉴럴네트워크 모델

모델 결정 → 파라미터 추정

$$\min_{W} \sum_{i=1}^{n} \{Y_i - f(X)\}^2$$

$$f(X) = w_0 + w_1 X_1 + w_2 X_2$$

다중선형회귀 모델

$$\min_{w_0, w_1, w_2} \sum_{i=1}^{n} \{Y_i - (w_0 + w_1 X_{1i} + w_2 X_{2i})\}^2$$

$$\hat{f}(X) = \hat{w}_0 + \hat{w}_1 X_1 + \hat{w}_2 X_2$$

모델 결정 → 파라미터 추정

$$\min_{W} \sum_{i=1}^{n} \{Y_i - f(X)\}^2$$

$$f(X) = \frac{1}{1 + e^{-(w_0 + w_1 X_1 + w_2 X_2)}}$$

로지스틱회귀 모델

$$\min_{w_0, w_1, w_2} \sum_{i=1}^{n} \left\{ Y_i - \left(\frac{1}{1 + e^{-(w_0 + w_1 X_1 + w_2 X_2)}} \right) \right\}^2$$

$$\hat{f}(X) = \frac{1}{1 + e^{-(\hat{w}_0 + \hat{w}_1 X_1 + \hat{w}_2 X_2)}}$$

모델 결정 → 파라미터 추정

$$\min_{W} \sum_{i=1}^{n} \{Y_i - f(X)\}^2$$

뉴럴네트워크 모델

$$f(X) = \frac{1}{1 + exp\left(-\left(w_0 + w_1\left(\frac{1}{1+e^{-(w_{01}+w_{11}X_1+w_{21}X_2)}}\right)\right) + w_2\left(\frac{1}{1+e^{-(w_{02}+w_{12}X_1+w_{22}X_2)}}\right)\right)}$$

$$\min_{w_0,\ldots,w_{22}} \sum_{i=1}^{n} \left\{Y_i - \left(\frac{1}{1 + exp\left(-\left(w_0 + w_1\left(\frac{1}{1+e^{-(w_{01}+w_{11}X_1+w_{21}X_2)}}\right)\right) + w_2\left(\frac{1}{1+e^{-(w_{02}+w_{12}X_1+w_{22}X_2)}}\right)\right)}\right)\right\}^2$$

$$\hat{f}(X) = \frac{1}{1 + exp\left(-\left(\hat{w}_0 + \hat{w}_1\left(\frac{1}{1+e^{-(\hat{w}_{01}+\hat{w}_{11}X_1+\hat{w}_{21}X_2)}}\right)\right) + \hat{w}_2\left(\frac{1}{1+e^{-(\hat{w}_{02}+\hat{w}_{12}X_1+\hat{w}_{22}X_2)}}\right)\right)}$$

모델 결정 → 파라미터 추정

$$f(X) = w_0 + w_1 X_1 + w_2 X_2 \qquad \hat{f}(X) = \hat{w}_0 + \hat{w}_1 X_1 + \hat{w}_2 X_2$$

다중선형회귀 모델

Least square estimation algorithm

$$f(X) = \frac{1}{1 + e^{-(w_0 + w_1 X_1 + w_2 X_2)}} \qquad \hat{f}(X) = \frac{1}{1 + e^{-(\hat{w}_0 + \hat{w}_1 X_1 + \hat{w}_2 X_2)}}$$

로지스틱회귀 모델

Conjugate gradient algorithm

$$f(X) = \frac{1}{1 + exp\left(-\left(w_0 + w_1\left(\frac{1}{1 + e^{-(w_{01} + w_{11} X_1 + w_{21} X_2)}}\right)\right) + w_2\left(\frac{1}{1 + e^{-(w_{02} + w_{12} X_1 + w_{22} X_2)}}\right)\right)}$$

뉴럴네트워크 모델

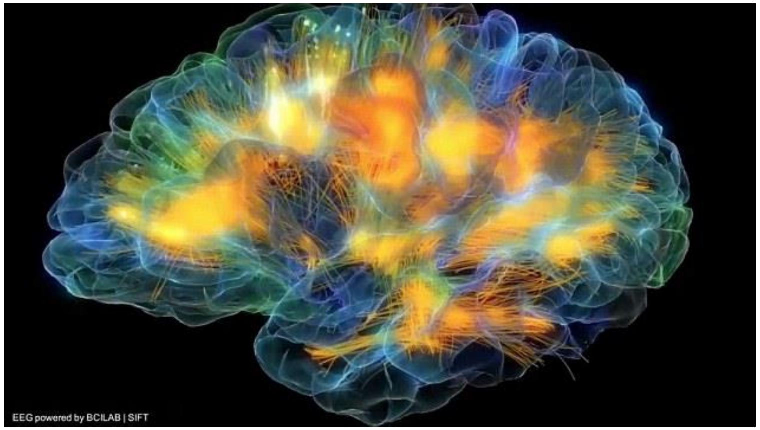
Backpropagation algorithm

$$\hat{f}(X) = \frac{1}{1 + exp\left(-\left(\hat{w}_0 + \hat{w}_1\left(\frac{1}{1 + e^{-(\hat{w}_{01} + \hat{w}_{11} X_1 + \hat{w}_{21} X_2)}}\right)\right) + \hat{w}_2\left(\frac{1}{1 + e^{-(\hat{w}_{02} + \hat{w}_{12} X_1 + \hat{w}_{22} X_2)}}\right)\right)}$$

모델 결정 → 파라미터 추정

모델 $Y = f(X)$

알고리즘



$$Y = w_1X_1 + W_2X_2 + w_3X_3 + w_4X_4 + W_5X_5 + W_6X_6 + w_7X_7$$

머신러닝 모델 학습 과정 요약

1. 모델 결정하기 (Y를 표현하기 위한 X들을 조합 방식 결정)

2. 모델을 구성하는 파라미터 찾기 (모델의 핵심!!)

어떻게?

가지고 있는 데이터를 이용하여

무엇을 추구하며?

실제 데이터의 값과 최대한 같게 나오도록!

EOD