Introduction to Predictive Analytics

예측?

데이터

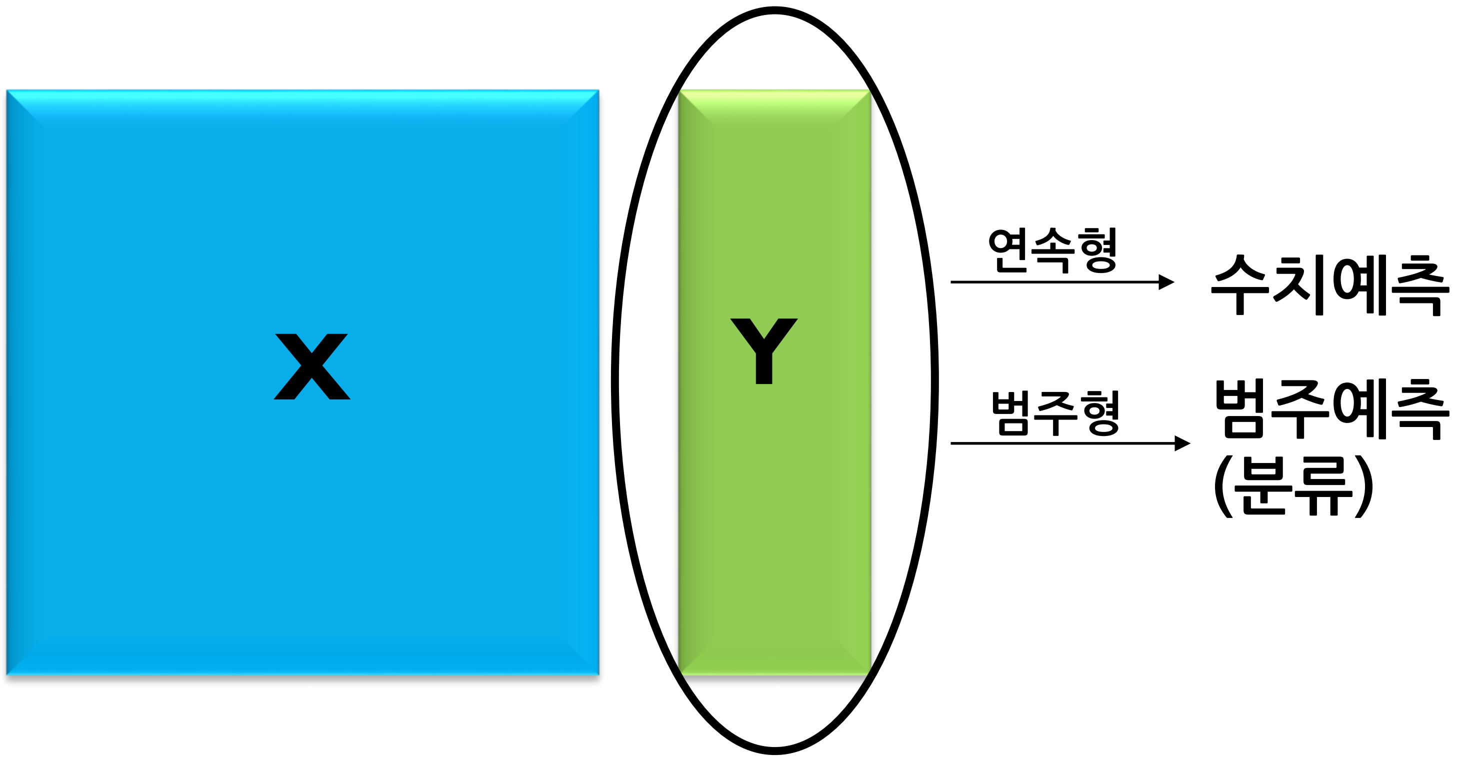
Y (결과): 종속변수, 반응변수, 출력변수

X (원인): 독립변수, 예측변수, 입력변수

변수 / 관측치	X_1	…	X_i	…	X_p
N_1	x_{11}	…	x_{1i}	…	x_{1p}
N_2	x_{21}	…	x_{2i}	…	x_{2p}
…	…	…	…	…	…
N_{n-1}	x_{n-11}	…	x_{n-1i}	…	x_{n-1p}
N_n	x_{n1}	…	x_{ni}	…	x_{np}

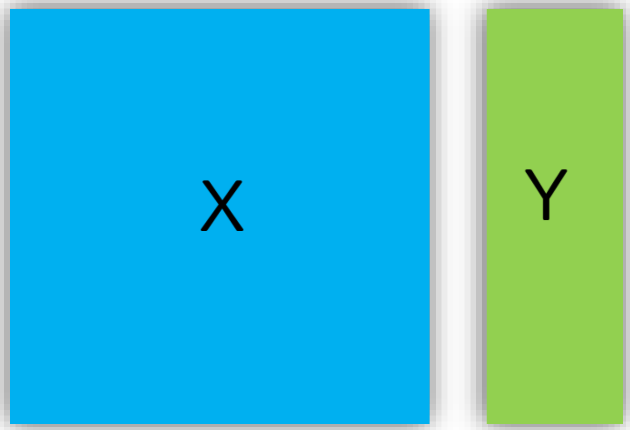
Y
y_1
y_2
…
y_{n-1}
y_n

수치예측 / 범주예측 (분류)



- 연속형 데이터: 데이터 자체를 숫자로 표현
  예)가격, 길이, 압력, 두께, …
- 범주형 데이터: 원칙적으로 숫자로 표시할 수 없는 데이터
  예) 제품불량여부 (양품/불량), 보험사기여부(정상/비정상), …

수치예측 데이터

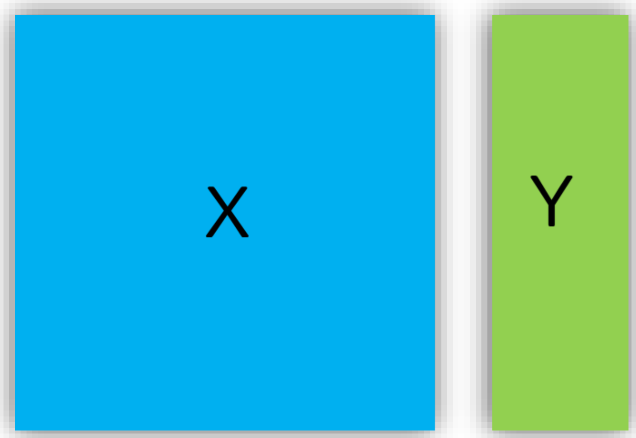


연속형 ⟶ **수치예측 (Regression)**

인자 (변수) / 관측치	X_1	…	X_i	…	X_p
N_1	x_{11}	…	x_{1i}	…	x_{1p}
N_2	x_{21}	…	x_{2i}	…	x_{2p}
…	…	…	…	…	…
N_{n-1}	x_{n-11}	…	x_{n-1i}	…	x_{n-1p}
N_n	x_{n1}	…	x_{ni}	…	x_{np}

Y
20.5
22.2
…
72.3
82.8

범주예측 데이터

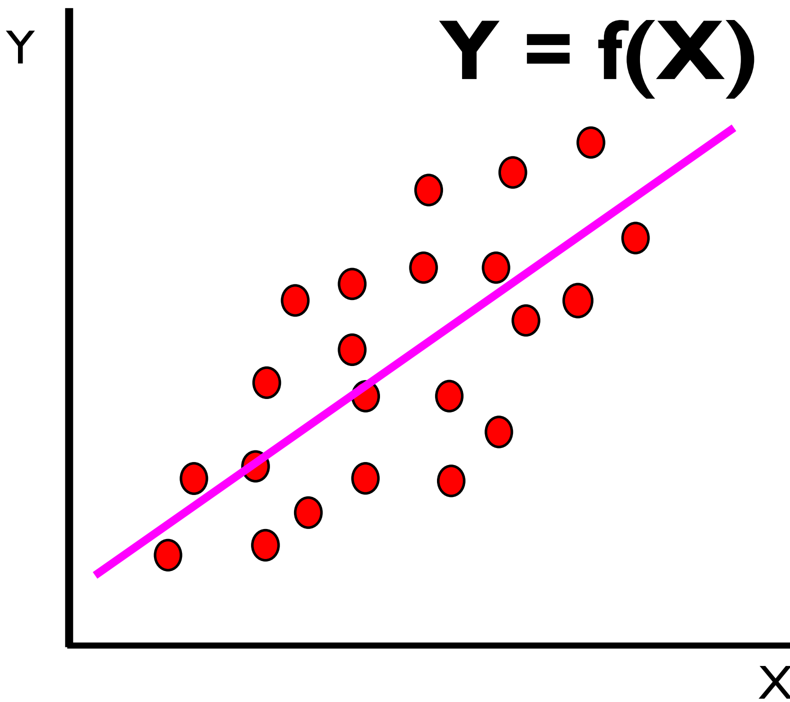


범주형 → 범주예측, 분류 (Classification)

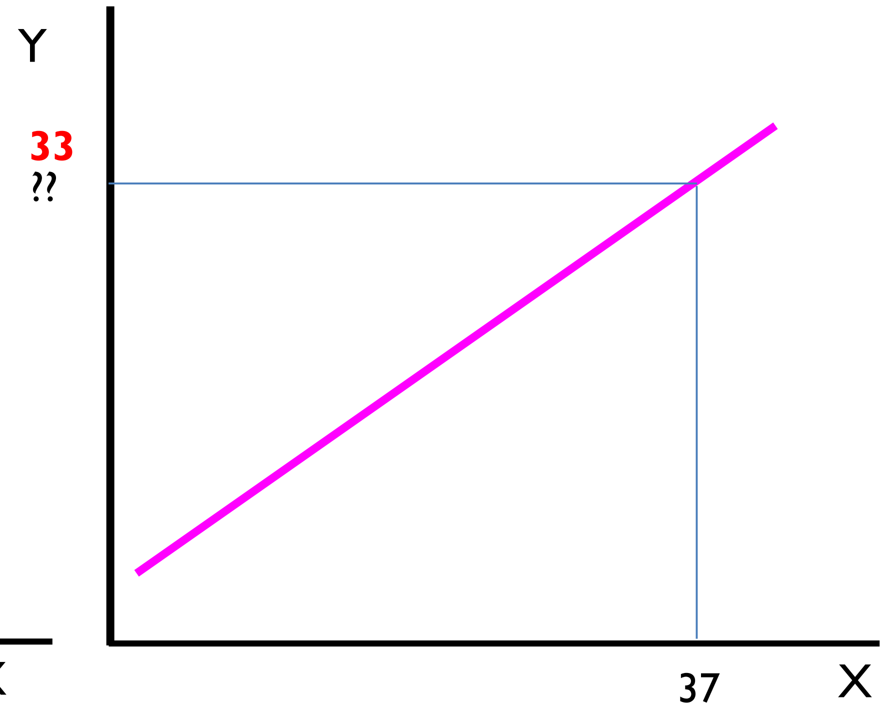
인자 (변수) / 관측치	X_1	…	X_i	…	X_p
N_1	x_{11}	…	x_{1i}	…	x_{1p}
N_2	x_{21}	…	x_{2i}	…	x_{2p}
…	…	…	…	…	…
N_{n-1}	x_{n-11}	…	x_{n-1i}	…	x_{n-1p}
N_n	x_{n1}	…	x_{ni}	…	x_{np}

Y
0 (정상)
0 (정상)
…
1(불량)
1(불량)

수치예측 모델링 개요



Training (학습)



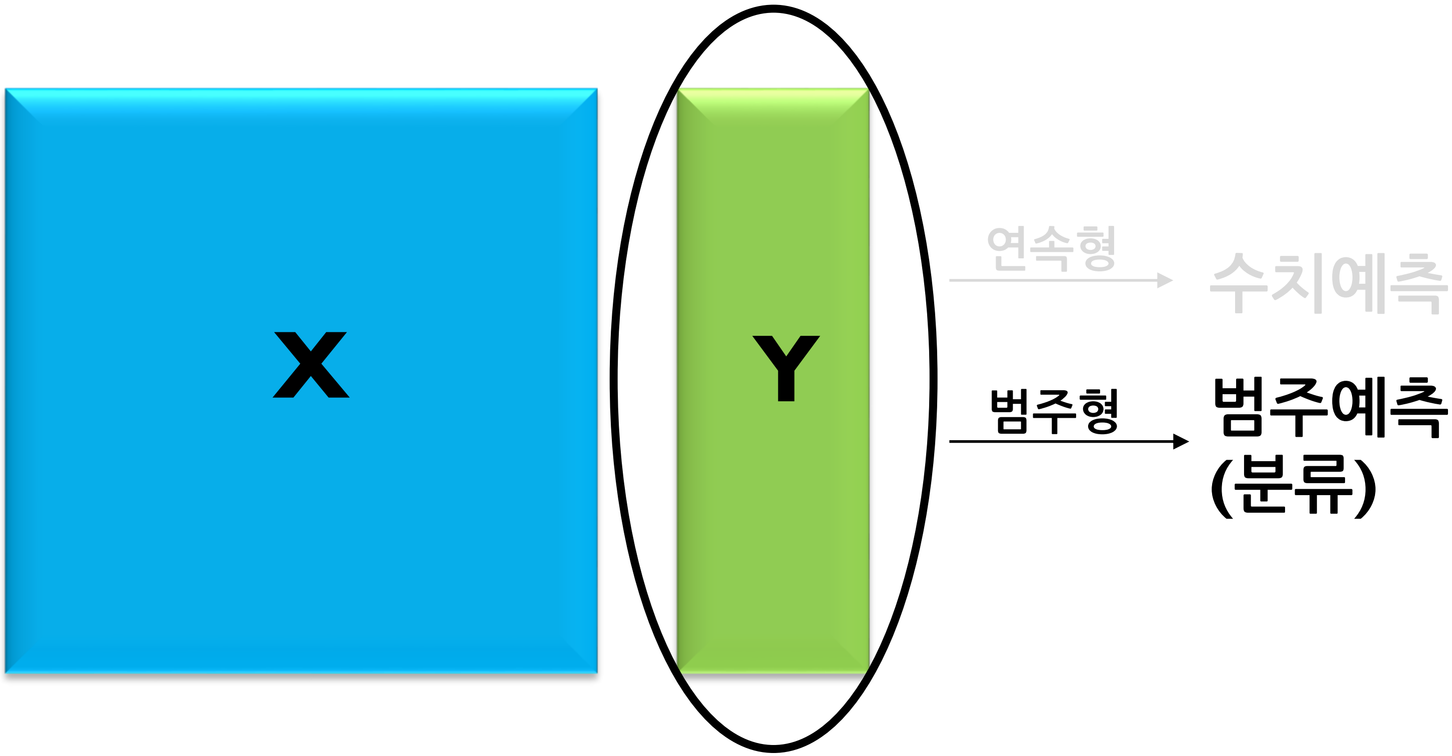
Testing, Inference (테스팅, 인퍼런스)

수치예측 예제 – 중고차 가격 예측

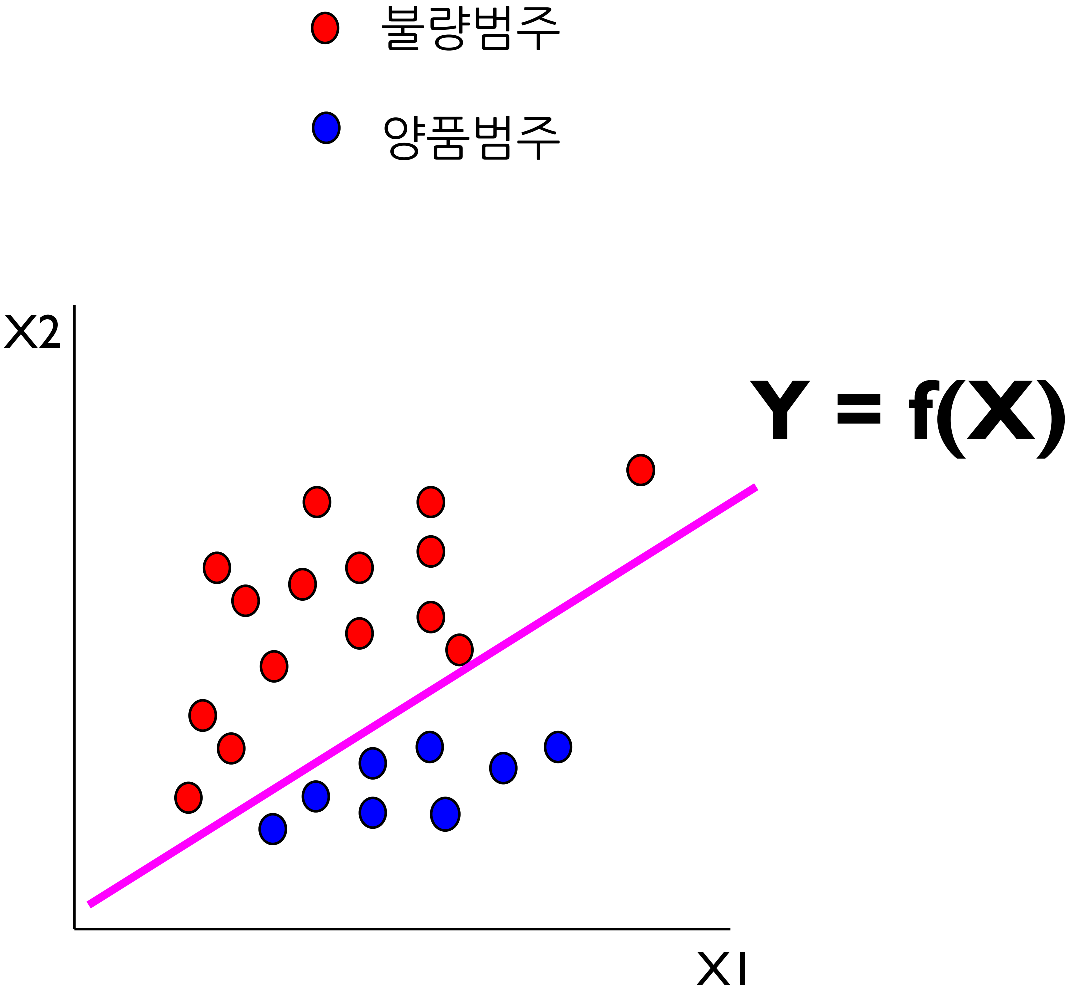
	X			Y
모델	주행거리	마력	용량 (CC)	가격
TOYOTA Corolla 2.0 D4D HATCHB TERRA 2/3-Doors	46986	90	2000	13500
TOYOTA Corolla 1800 T SPORT VVT I 2/3-Doors	19700	192	1800	21500
TOYOTA Corolla 1.9 D HATCHB TERRA 2/3-Doors	71138	69	1900	12950
TOYOTA Corolla 1.8 VVTL-i T-Sport 3-Drs 2/3-Doors	31461	192	1800	20950
TOYOTA Corolla 1.8 16V VVTLI 3DR T SPORT BNS 2/3-Doors	43610	192	1800	19950
TOYOTA Corolla 1.6 VVTI Linea Terra Comfort 2/3-Doors	21716	110	1600	17950
TOYOTA Corolla 1.6 16v L.SOL 2/3-Doors	25563	110	1600	16750
TOYOTA Corolla 1.6 16V VVT I 3DR TERRA 2/3-Doors	64359	110	1600	16950
TOYOTA Corolla 1.6 16V VVT I 3DR SOL AUT4 2/3-Doors	43905	110	1600	16950
TOYOTA Corolla 1.6 16V VVT I 3DR SOL 2/3-Doors	56349	110	1600	15950
TOYOTA Corolla 1.4 VVTI Linea Terra 2/3-Doors	9750	97	1400	12950
TOYOTA Corolla 1.4 16V VVT I 3DR 2/3-Doors	27500	97	1400	14750
TOYOTA Corolla 1.4 16V VVT I 3DR 2/3-Doors	49059	97	1400	13950
TOYOTA Corolla 1.4 16V VVT I 3DR 2/3-Doors	44068	97	1400	16750
TOYOTA Corolla 1.4 16V VVT I 3DR 2/3-Doors	46961	97	1400	13950
TOYOTA Corolla 2.0 D4D 90 5DR TERRA COMFORT 4/5-Doors	110404	90	2000	16950
TOYOTA Corolla 2.0 D4D 90 5DR TERRA COMFORT 4/5-Doors	100250	90	2000	16950
TOYOTA Corolla 2.0 D4D 90 5DR SOL 4/5-Doors	84000	90	2000	19000
TOYOTA Corolla 2.0 D4D 90 5DR TERRA 4/5-Doors	79375	90	2000	17950
TOYOTA Corolla 1.4 16V VVT I 5DR TERRA COMFORT 4/5-Doors	75048	97	1400	15800
TOYOTA Corolla 1.4 16V VVT I 5DR TERRA COMFORT 4/5-Doors	**132151**	**110**	**1600**	**??????**

18200

수치예측 / 범주예측 (분류)



범주예측 모델링 개요



범주예측 예제 – 불량 예측

- 배터리 공정에서 설비 파라미터 측정값들을 이용하여, 배터리가 양품인지 불량품인지 여부를 예측

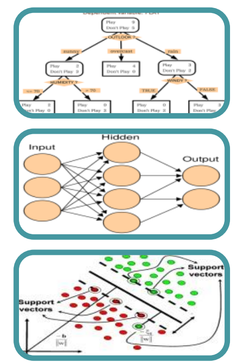
f

배터리설비 파라미터 배터리 상태

	X_1	X_2	…	X_p	Y
제품$_1$	a_{11}	a_{12}	…	a_{1p}	1
제품$_2$	a_{21}	a_{22}	…	a_{2p}	1
⋮	⋮	⋮	⋱	⋮	⋮
제품$_n$	a_{n1}	a_{n2}	…	a_{np}	0

양품

불량품



새로운 배터리

?

양품

불량품

배터리 공정 데이터

모델구축

불량 배터리 예측

범주예측 예제 – 고객 이탈 예측

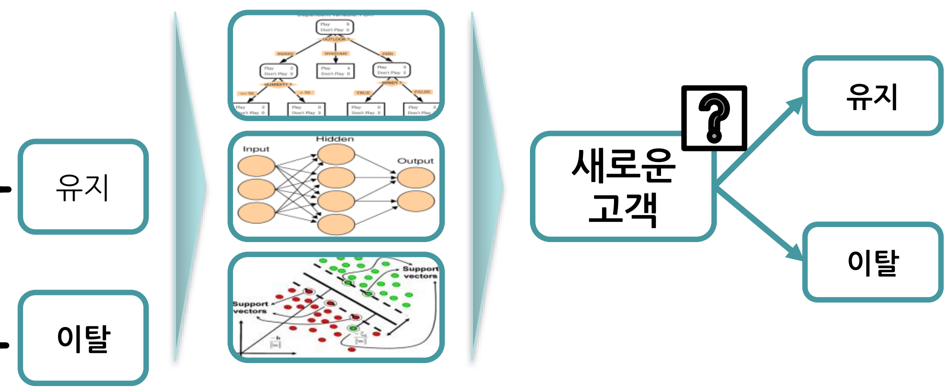
- 고객의 정보(성별, 연령, 직업, 연봉 등)를 이용하여,
  고객 이탈 여부를 예측

f

고객정보 이탈여부

	X_1	X_2	…	X_p	Y
고객$_1$	a_{11}	a_{12}	…	a_{1p}	1
고객$_2$	a_{21}	a_{22}	…	a_{2p}	1
⋮	⋮	⋮	⋱	⋮	⋮
고객$_n$	a_{n1}	a_{n2}	…	a_{np}	0

고객 데이터



모델구축

고객이탈예측

범주예측 예제 – 보험 사기 여부 예측

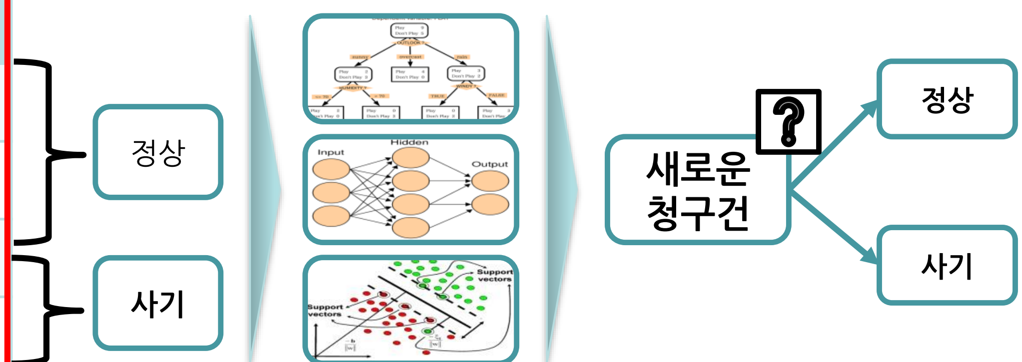
- 각 청구 건에 대한 내역 분석을 통해 청구 건에 대한 사기 여부 예측

f

청구내역 사기여부

	X_1	X_2	…	X_p	Y
청구건$_1$	a_{11}	a_{12}	…	a_{1p}	1
청구건$_2$	a_{21}	a_{22}	…	a_{2p}	1
⋮	⋮	⋮	⋱	⋮	⋮
청구건$_n$	a_{n1}	a_{n2}	…	a_{np}	0

청구 관련 데이터



모델구축

보험사기예측

데이터

Y (결과): 종속변수, 반응변수, 출력변수

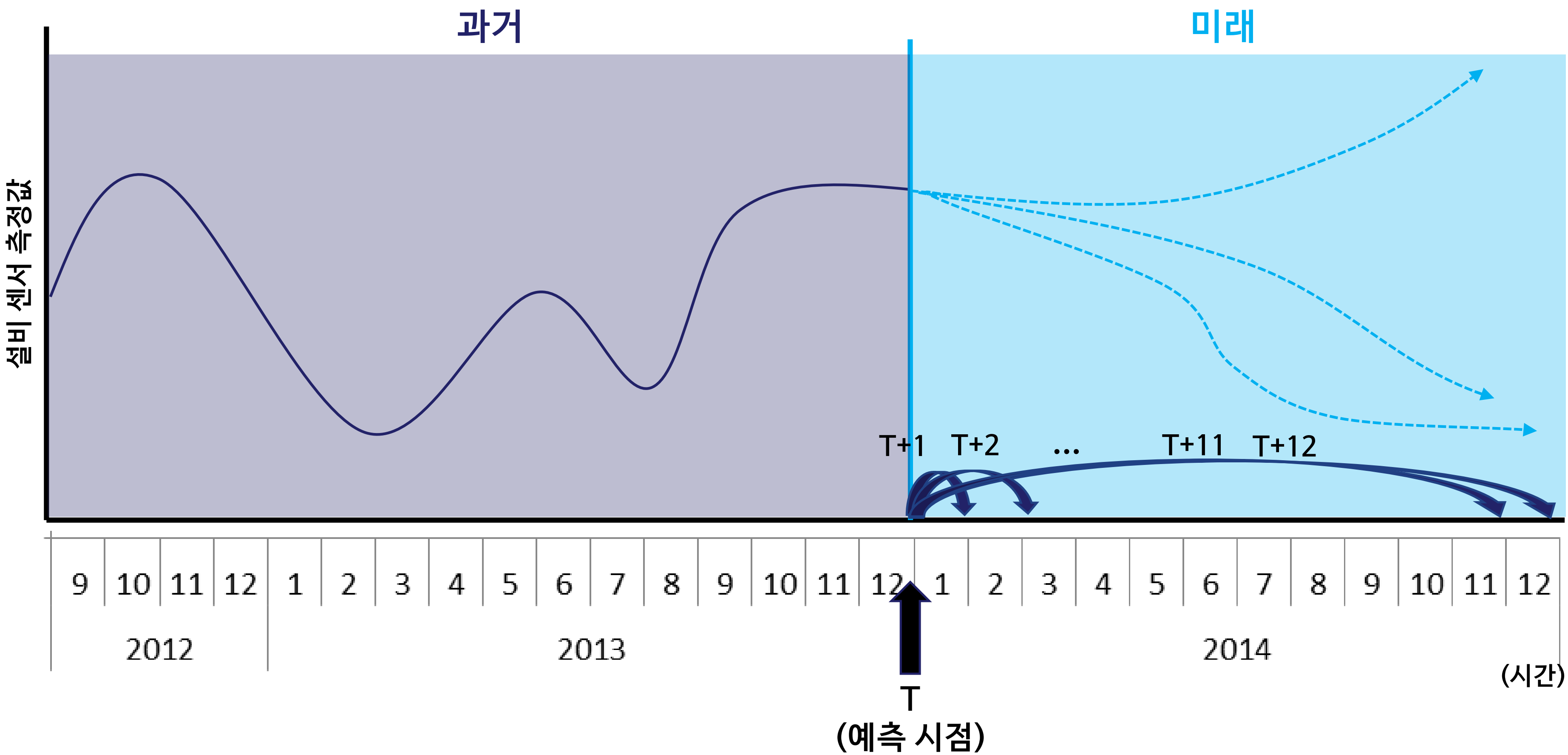
X (원인): 독립변수, 예측변수, 입력변수

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-11}	…	x_{n-1i}	…	x_{n-1p}	72.3
N_n	x_{n1}	…	x_{ni}	…	x_{np}	82.8

단변량 시계열 예측

	X 변수					Y 변수
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

단변량 시계열 예측



과거
미래
설비 센서 측정값
T+1
T+2
...
T+11
T+12
9
10
11
12
1
2
3
4
5
6
7
8
9
10
11
12
1
2
3
4
5
6
7
8
9
10
11
12
2012
2013
2014
(시간)
T
(예측 시점)

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)

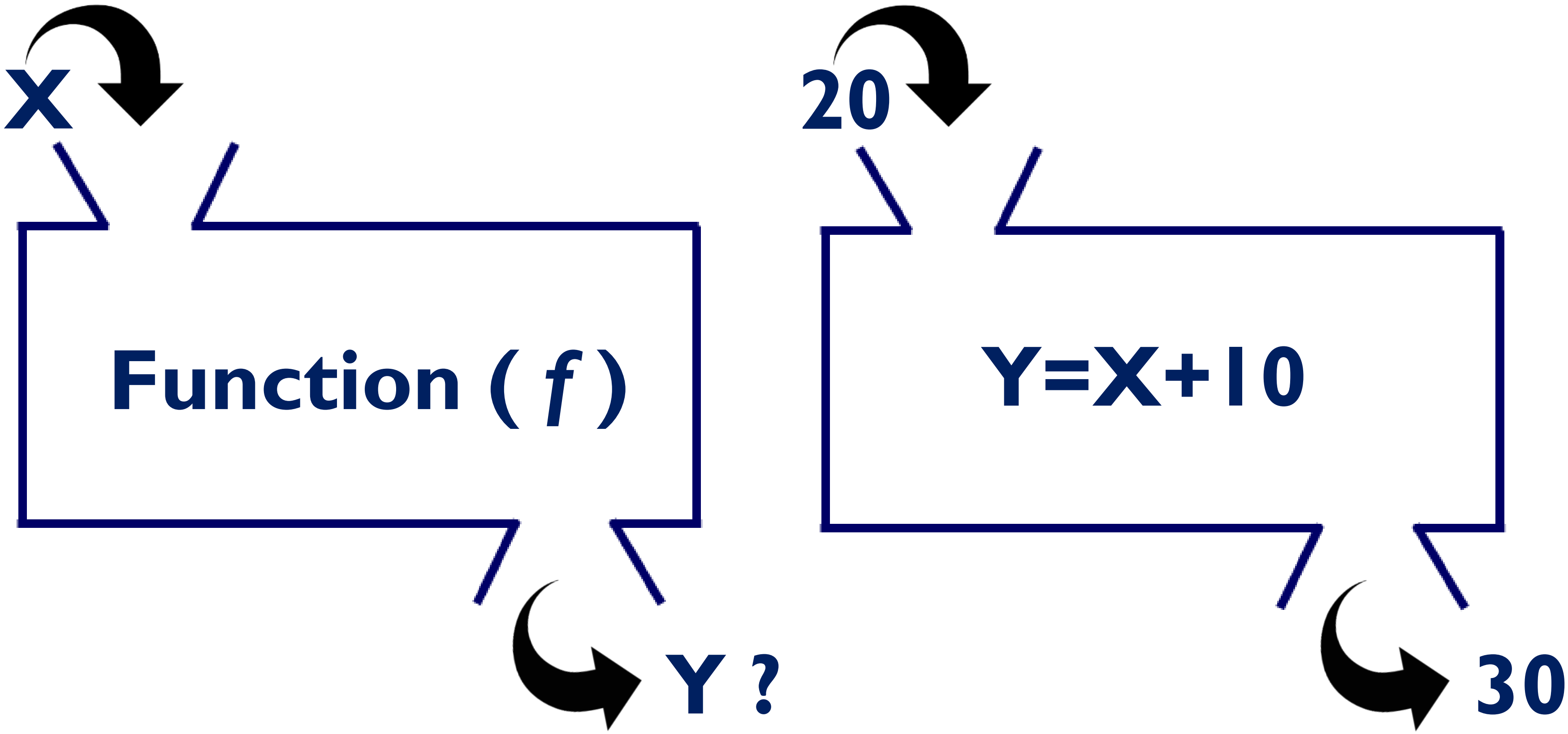
Y를 설명하는 X변수는 보통 여러 개

여러 개의 X와 Y의 관계를 찾는 것!

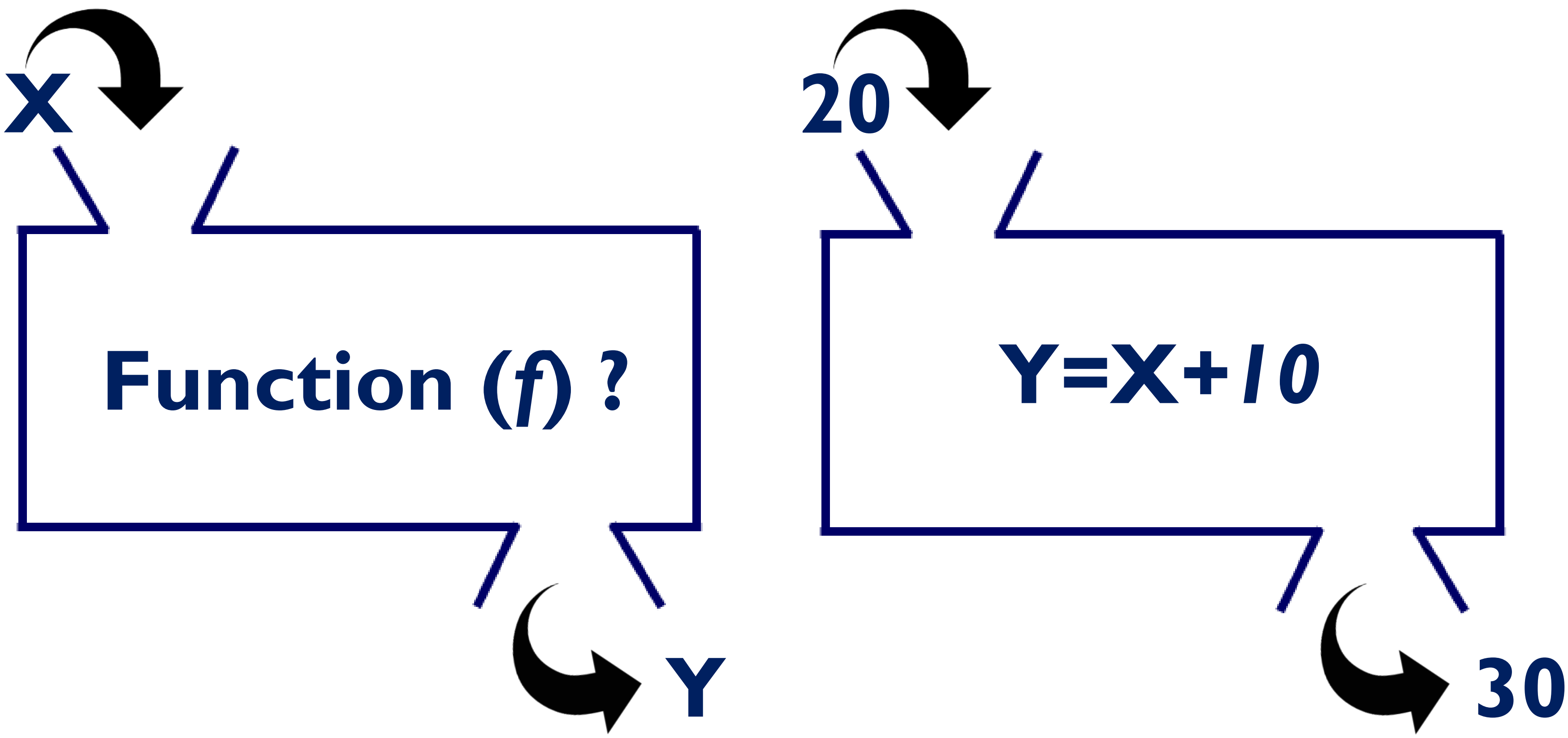
X변수들을 조합(결합)하여 Y를 표현

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

X와 Y의 관계 찾기

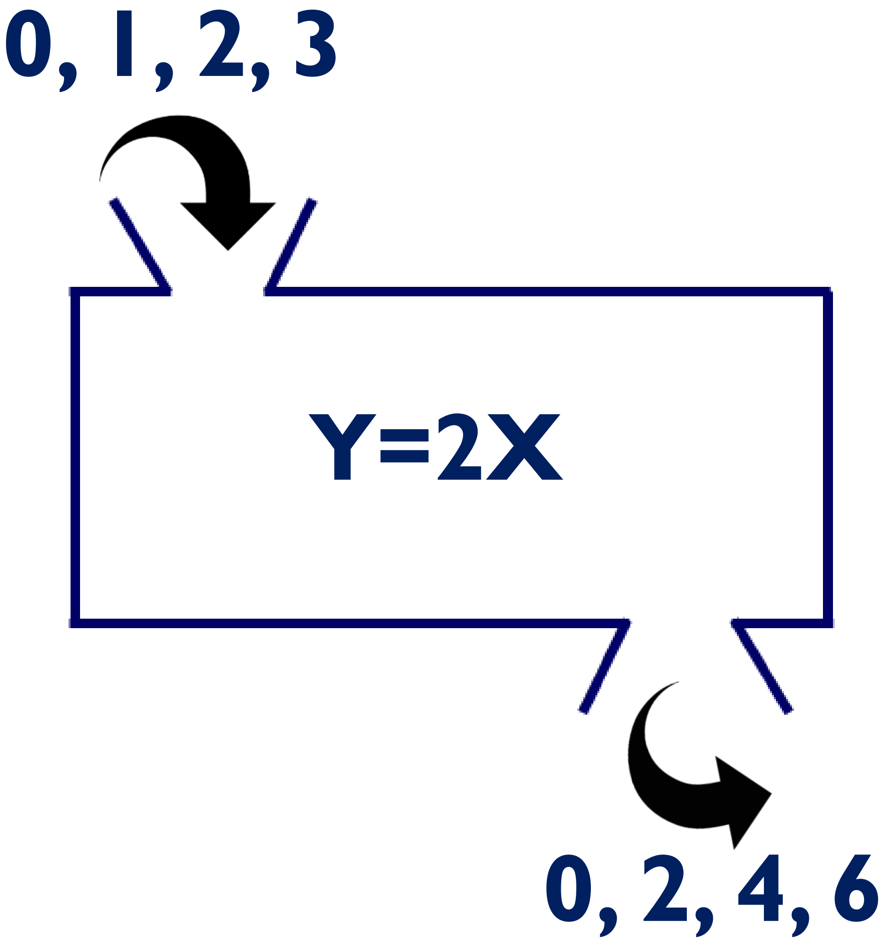


X와 Y의 관계 찾기



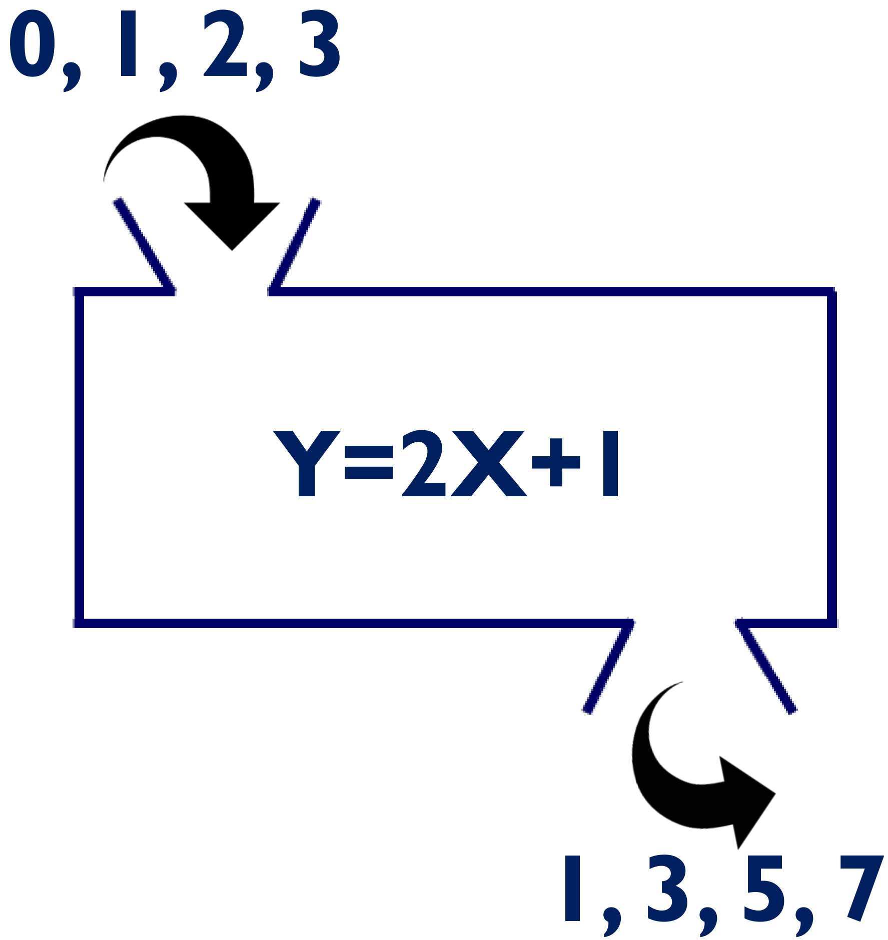
X와 Y의 관계 찾기

X	Y
0	0
1	2
2	4
3	6



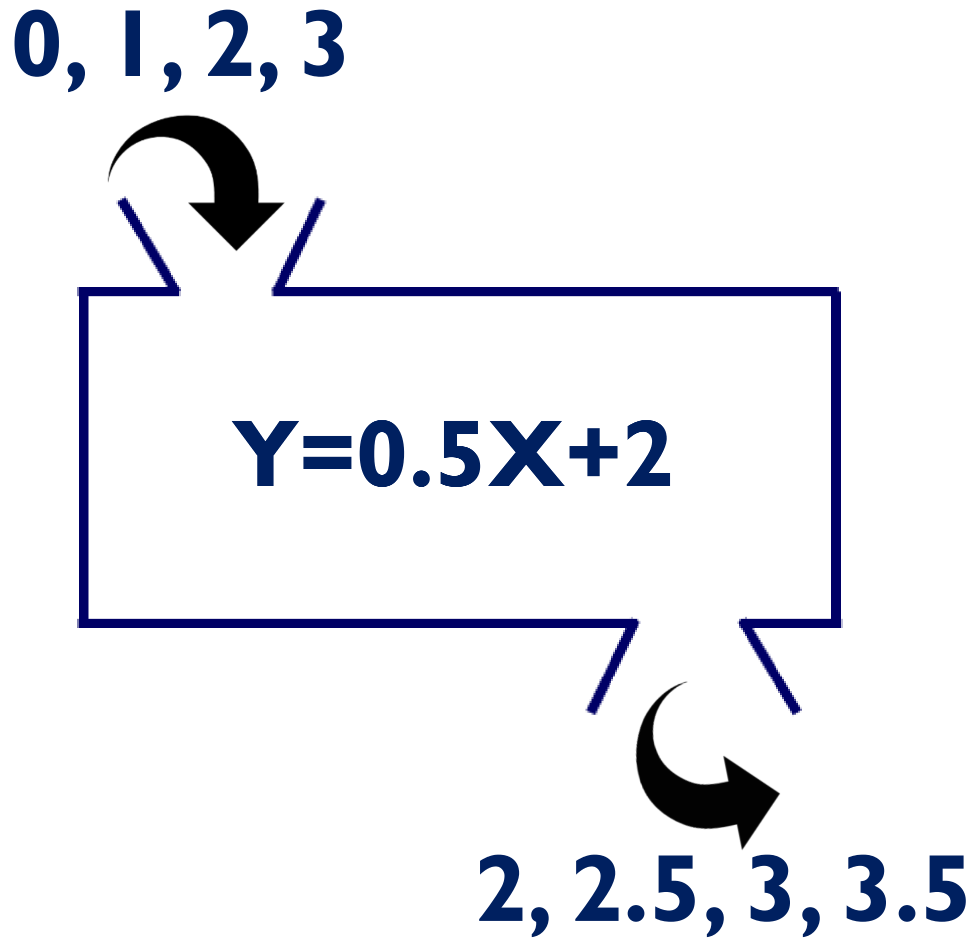
X와 Y의 관계 찾기

X	Y
0	1
1	3
2	5
3	7



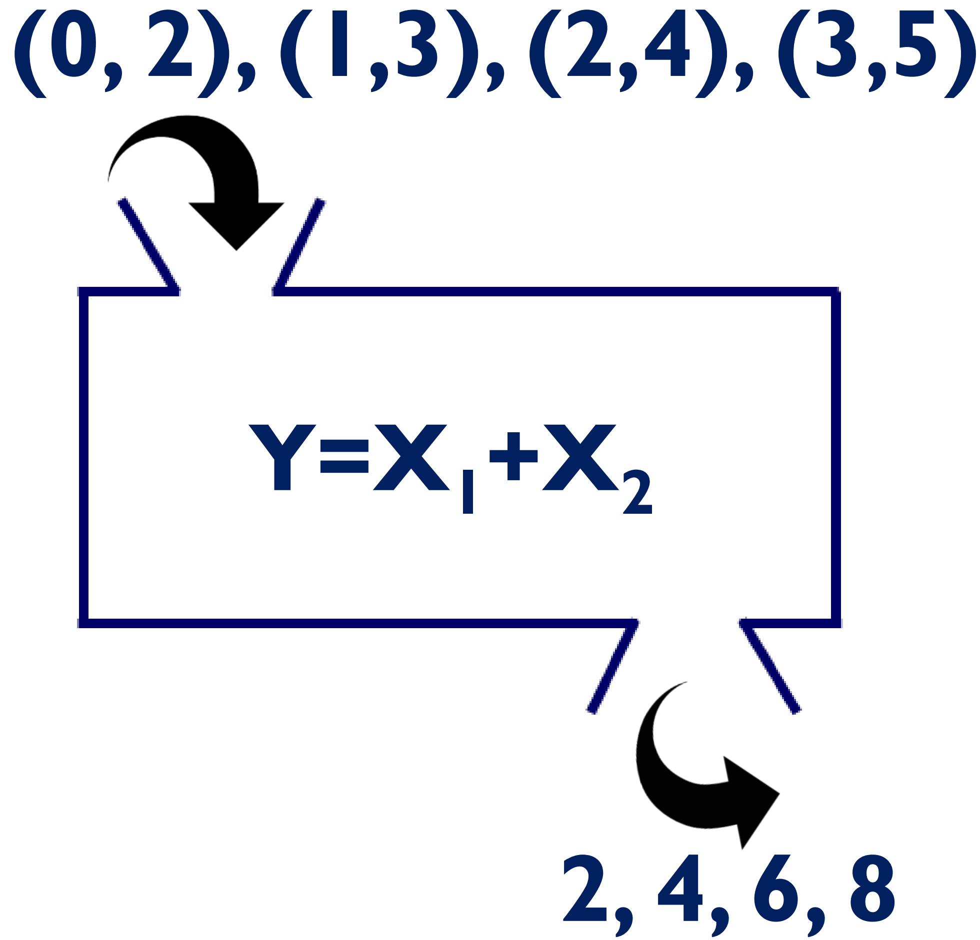
X와 Y의 관계 찾기

X	Y
0	2
1	2.5
2	3
3	3.5



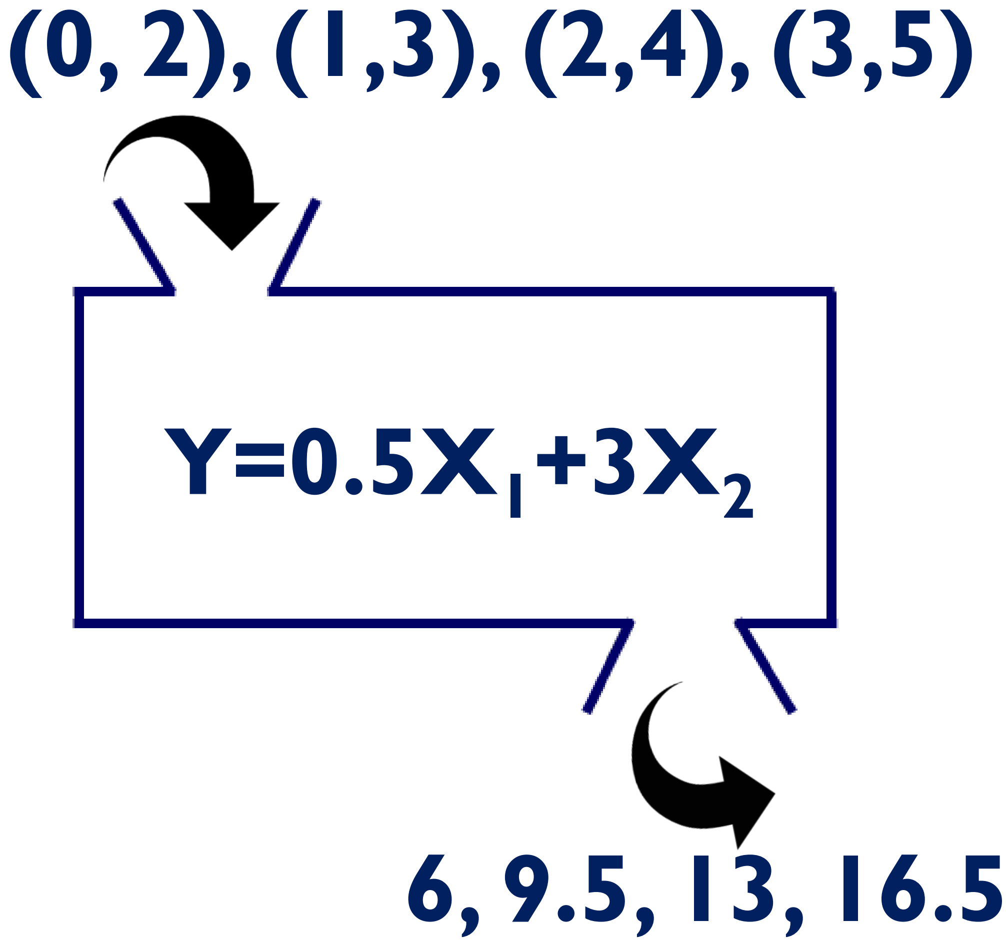
X와 Y의 관계 찾기

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



X와 Y의 관계 찾기

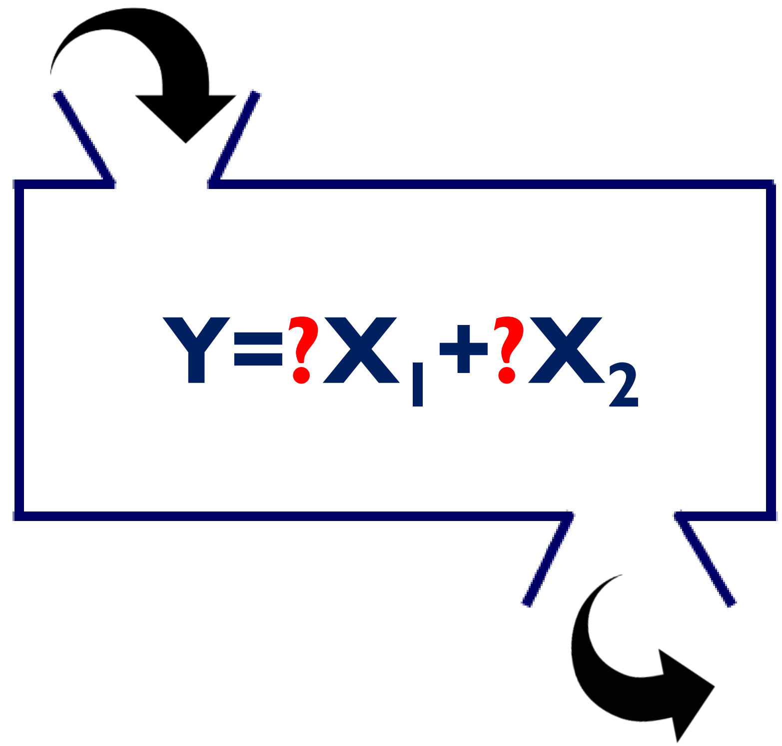
X_1	X_2	Y
0	2	6
1	3	9.5
2	4	13
3	5	16.5



X와 Y의 관계 찾기

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

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



6, 9, 11.5, 14.5

X와 Y의 관계 찾기

	X_1	X_2	X_3	Y
모델	주행거리	마력	용량	가격
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

$$Y=\underbrace{?X_1+?X_2+?X_3}_{\text{X로 설명되는 부분}}\ \underbrace{+\varepsilon}_{\text{그렇지 않은 부분}}$$

머신러닝 모델 핵심

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

$$Y = w_1 X_1 + w_2 X_2 + \varepsilon$$

$$w_1? \quad w_2?$$

Given X_1, X_2, Y (데이터)

머신러닝 모델 핵심 - 파라미터

$$Y = w_1 X_1 + w_2 X_2 + \varepsilon$$

파라미터 (母數)(媒介變數)

데이터가 주어졌을 때 모델의 파라미터 찾기!

파라미터 추정

$$Y = w_1 X_1 + w_2 X_2 + \varepsilon$$
$$= f(X) + \varepsilon$$

$$\varepsilon = Y - f(X)$$ ➡ 오차

Loss function
(손실함수)

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

파라미터 추정

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

Loss function (손실함수)

$$f(X) = w_1 X_1 + w_2 X_2 + \varepsilon$$

$$\varepsilon = Y - (w_1 X_1 + w_2 X_2)$$

$$\varepsilon_i = Y_i - (w_1 X_{1i} + w_2 X_{2i}),\ i = 1, 2, \ldots, n$$

파라미터 추정

$$\varepsilon_i = Y_i - (w_1 X_{1i} + w_2 X_{2i}),\ i = 1, 2, \ldots, n$$

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

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

Cost function
(비용함수)

파라미터 추정

$$\sum_{i=1}^{n}\{Y_i-(w_1X_{1i}+w_2X_{2i})\}^2$$

Cost function
(비용함수)

비용함수를 최소로 하는 w_1와 w_2를 찾자!

$$\min_{w_1,w_2}\sum_{i=1}^{n}\{Y_i-(w_1X_{1i}+w_2X_{2i})\}^2$$

파라미터 추정

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

답: $\hat{w}_1, \hat{w}_2$

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

비용함수

- **Regression (Y가 연속형)**

Mean squared error (MSE)

$$C(Y, f(X)) = \sum_{i=1}^{n} \{Y_i - f(X_i)\}^2$$

- **Classification (Y가 범주형)**

Cross Entropy

$$C(Y, f(X)) = \sum_{i=1}^{N} \{-Y_i \cdot \log(f(X_i)) - (1 - Y_i) \cdot \log(1 - f(X_i))\}$$

모델 결정

$$C(Y, \boxed{f(X)})$$

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

선형회귀 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

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

$$f(X_i) = w_0 + w_1 X_{1i} + w_2 X_{2i}$$ 선형회귀 모델

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

Least square estimation algorithm
(최소제곱법)

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

모델 결정 → 파라미터 추정

$$\min_{W} \sum_{i=1}^{N} \{-Y_i \cdot \log(f(X_i)) - (1 - Y_i) \cdot \log(1 - f(X_i))\}$$

$$f(X_i) = \frac{1}{1 + e^{-(w_0 + w_1 X_{1i} + w_2 X_{2i})}}$$ 로지스틱회귀 모델

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

Conjugate gradient algorithm

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

모델 결정 → 파라미터 추정

$$\min_{W}\sum_{i=1}^{N}\{-Y_i\cdot\log(f(X_i))-(1-Y_i)\cdot\log(1-f(X_i))\}$$

뉴럴네트워크 모델

$$f(X_i)=\frac{1}{1+exp\left(-\left(w_0+w_1\left(\frac{1}{1+e^{-(w_{01}+w_{11}X_{1i}+w_{21}X_{2i})}}\right)\right)+w_2\left(\frac{1}{1+e^{-(w_{02}+w_{12}X_{1i}+w_{22}X_{2i})}}\right)\right)}$$

$$\min_{w_0,\dots,w_{22}}\sum_{i=1}^{N}\{-Y_i\cdot\log\left(\frac{1}{1+exp\left(-\left(w_0+w_1\left(\frac{1}{1+e^{-(w_{01}+w_{11}X_{1i}+w_{21}X_{2i})}}\right)\right)+w_2\left(\frac{1}{1+e^{-(w_{02}+w_{12}X_{1i}+w_{22}X_{2i})}}\right)\right)}\right)-(1-Y_i)\cdot\log\left(1-\frac{1}{1+exp\left(-\left(w_0+w_1\left(\frac{1}{1+e^{-(w_{01}+w_{11}X_{1i}+w_{21}X_{2i})}}\right)\right)+w_2\left(\frac{1}{1+e^{-(w_{02}+w_{12}X_{1i}+w_{22}X_{2i})}}\right)\right)}\right)\}$$

Backpropagation algorithm
(오차역전파)

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

모델 결정 → 파라미터 추정

$$f(X) = w_0 + w_1X_1 + w_2X_2 \qquad \hat{f}(X) = \hat{w}_0 + \hat{w}_1X_1 + \hat{w}_2X_2$$

다중선형회귀 모델

Least square estimation algorithm

$$f(X) = \frac{1}{1+e^{-(w_0+w_1X_1+w_2X_2)}} \qquad \hat{f}(X) = \frac{1}{1+e^{-(\hat{w}_0+\hat{w}_1X_1+\hat{w}_2X_2)}}$$

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

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

EOD