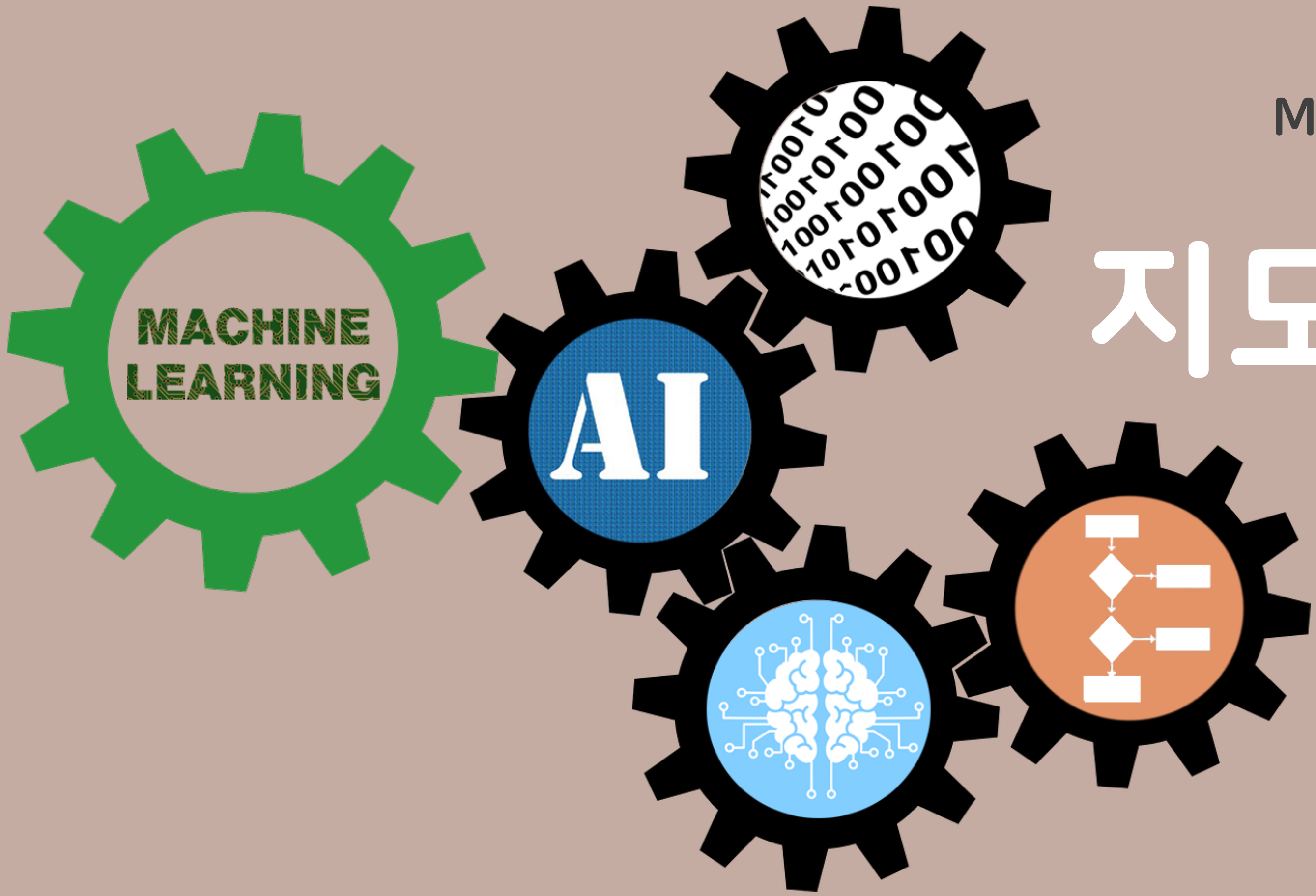


Machine Learning **3강**

지도학습: 회귀

컴퓨터과학과 이관용 교수

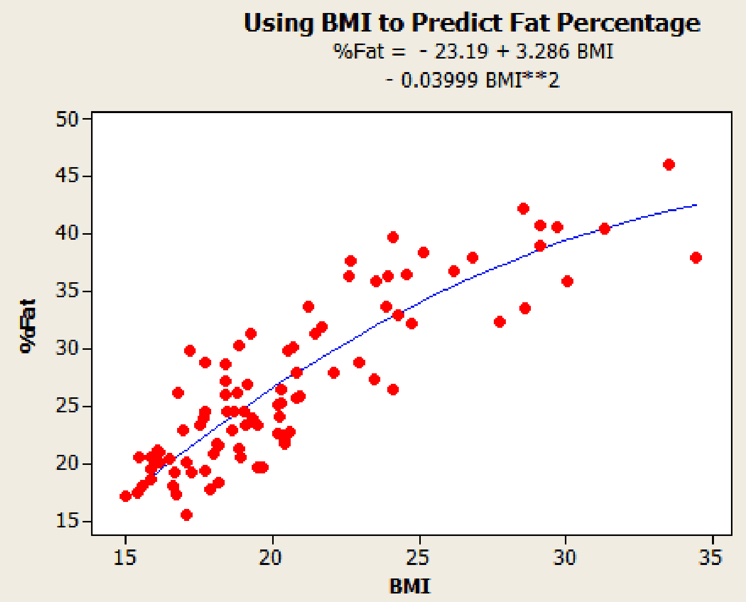
국립한국방송통신대학교
Korea National Open University

학습목차

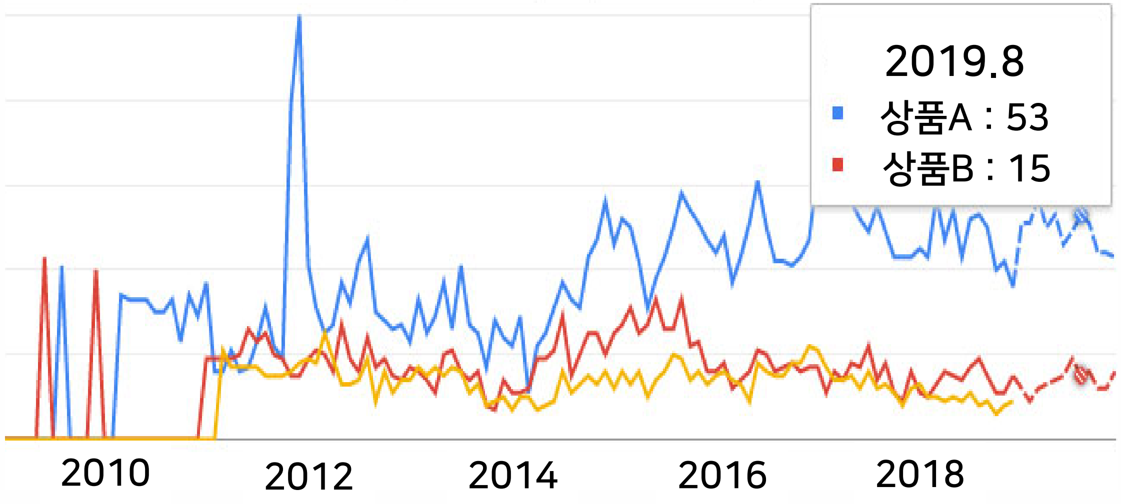
1 회귀의 개념

회귀?

- 입력변수와 출력변수 사이의 매핑 관계를 찾는 것
  - 예: 시계열 예측 → 주가 예측, 환율 예측 등



빅데이터 분석을 통한 판매 예측



  - 선형회귀, 비선형회귀, 로지스틱 회귀, SVM, 신경망(MLP, RBF, CNN, LSTM)

회귀 시스템

○ 입력·출력의 관계

학습 데이터 집합
$D = \{(\boldsymbol{x}_i, y_i)\}_{i=1...N}$

$y_i \in R$

학습 (데이터 분석)

회귀함수
$y = f(\boldsymbol{x};\, \boldsymbol{\theta})$

테스트 데이터
$\boldsymbol{x}_{new}$

$y_{new} = f(\boldsymbol{x}_{new};\, \boldsymbol{\theta})$?

예측 결과
y_{new}

○ 학습 결과 → 회귀함수

회귀 시스템

학습 목표

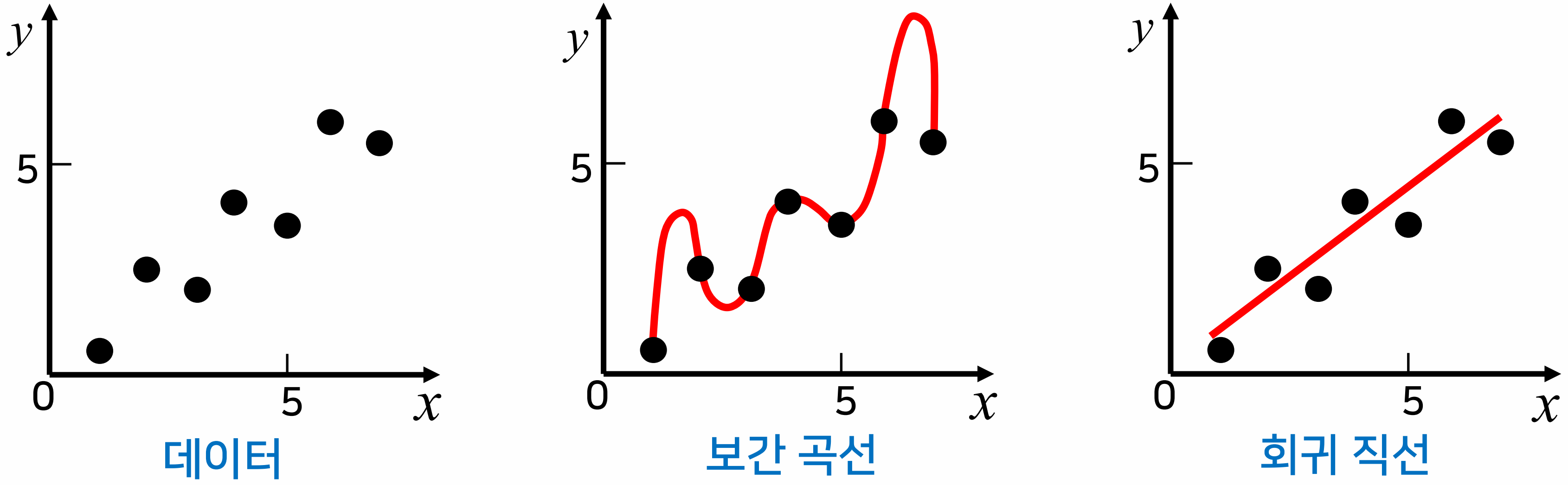
- 예측 오차를 최소화하는 최적의 회귀함수 $y = f(\boldsymbol{x};\boldsymbol{\theta})$를 찾는 것

$$E(D;\boldsymbol{\theta}) = \frac{1}{N}\sum_{(\boldsymbol{x}_i, y_i)\in D}\{y_i - f(\boldsymbol{x}_i;\boldsymbol{\theta})\}^2$$

"최소자승법" or "최소제곱법"
least square method

보간법과 회귀

○ 데이터를 가장 잘 표현하는 직선/곡선을 찾는 경우

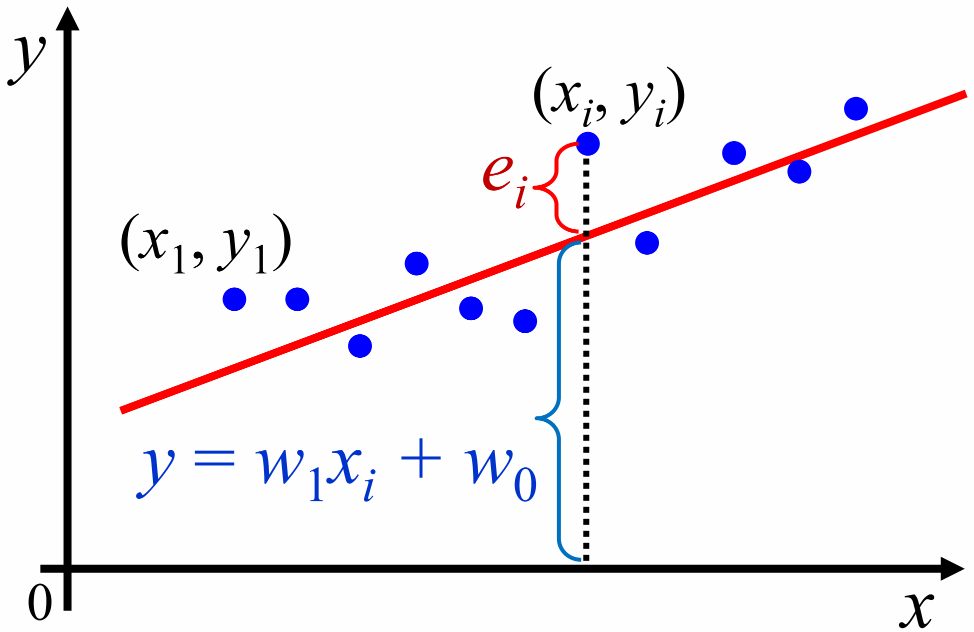


- 보간 곡선 → 제곱 오차가 0이지만 매우 복잡
- 회귀 직선 → 작은 오차,
  전체적인 데이터의 경향을 보여주는 입출력의 관계 표현에 적합

2 선형회귀

선형회귀?

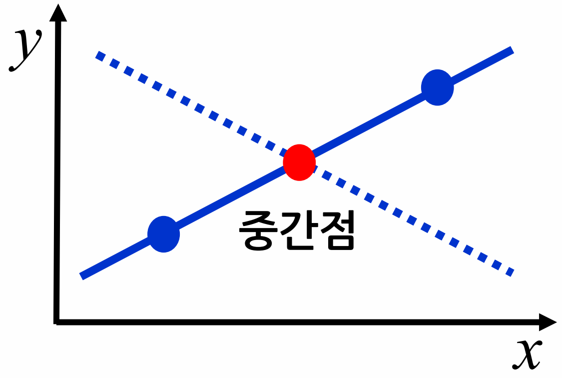
- 데이터 집합 $D = \{(x_i, y_i)\}_{i=1,\cdots,N}$ $(x_i \in R, y_i \in R)$에 대해
  (x, y) 관계를 설명할 수 있는 선형함수 $y = w_1 x + w_0 + e$를 찾는 것
  - w_1 → 기울기
  - w_0 → 절편
  - e → 오차 또는 잔차
    residual



좋은 선형회귀 모델?

- 모든 데이터에 대해서 잔차가 가능한 작아야 함
  - $e_i = y_i - (w_1 x_i + w_0)$
- 평가 기준
  - 모든 데이터에 대한 잔차의 합 → $\sum_{i=1}^{N} e_i = \sum_{i=1}^{N}\{y_i - (w_1 x_i + w_0)\}$
    - ✓ 부적합한 방법
  - 잔차의 제곱의 합

$$E(w_1, w_0) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N}\{y_i - (w_1 x_i + w_0)\}^2$$



- ✓ 주어진 데이터 집합에 대해 유일한 직선을 생성

최적의 회귀 매개변수?

○ 오차함수

$$E(w_1, w_0) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N} \{y_i - (w_1 x_i + w_0)\}^2$$

매개변수 → w_1, w_0

○ 최적의 매개변수

$$\boxed{w_0 = \bar{y} - w_1 \bar{x}}$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\boxed{w_1 = \frac{N\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}}$$

매개변수 계산 과정

$$E(w_1, w_0) = \sum_{i=1}^{N} e_i^2 = \sum_{i=1}^{N}\{y_i - (w_1 x_i + w_0)\}^2$$

각 매개변수에 대해서 편미분 수행

$$\frac{\partial E(w_1,\ w_0)}{\partial w_0} = -2\sum_{i=1}^{N}\{y_i - (w_1 x_i + w_0)\} = 0$$

$$\frac{\partial E(w_1,\ w_0)}{\partial w_1} = -2\sum_{i=1}^{N}\{y_i - (w_1 x_i + w_0)\} x_i = 0$$

연립방정식 형태로 정리

$$\sum_{i=1}^{N} y_i - \sum_{i=1}^{N} w_0 - \sum_{i=1}^{N} w_1 x_i = 0 \Rightarrow w_0 N + w_1 \sum_{i=1}^{N} x_i = \sum_{i=1}^{N} y_i$$

$$\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} w_0 x_i - \sum_{i=1}^{N} w_1 x_i^2 = 0 \Rightarrow w_0 \sum_{i=1}^{N} x_i + w_1 \sum_{i=1}^{N} x_i^2 = \sum_{i=1}^{N} y_i x_i$$

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$$

$$\bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$$

$$\boxed{w_0 = \bar{y} - w_1 \bar{x} \qquad w_1 = \frac{N\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} x_i^2 - \left(\sum_{i=1}^{N} x_i\right)^2}}$$

예제

데이터

x_i	y_i	${e_i}^2$
1	0.5	0.1687
2	2.5	0.5625
3	2.0	0.3473
4	4.0	0.3265
5	3.5	0.5896
6	6.0	0.7972
7	5.5	0.1993
Σ	24.0	2.911

$$N = 7$$

$$\sum x_i y_i = 119.5$$

$$\sum {x_i}^2 = 140$$

$$\sum x_i = 28 \qquad \bar{x} = \frac{28}{7} = 4$$

$$\sum y_i = 24 \qquad \bar{y} = \frac{24}{7} = 3.428571$$

$$w_1 = \frac{N\sum_{i=1}^{N} y_i x_i - \sum_{i=1}^{N} x_i \sum_{i=1}^{N} y_i}{N\sum_{i=1}^{N} {x_i}^2 - \left(\sum_{i=1}^{N} x_i\right)^2} = 0.8392857$$

$$w_0 = \bar{y} - w_1\bar{x} = 0.0714282$$

회귀함수

$$y = 0.8392857x + 0.0714282$$

예측과 평가

- 회귀함수를 사용한 새 데이터 x_{new}에 대한 예측

$$y_{new} = w_1 x_{new} + w_0$$

- 테스트 데이터 집합 $\{(x_j^{tst}, y_j^{tst})\}_{j=1,2,\cdots,N_{tst}}$ 에 대한 평가 기준
  - 평균 제곱 오차 Mean Squared Error: MSE

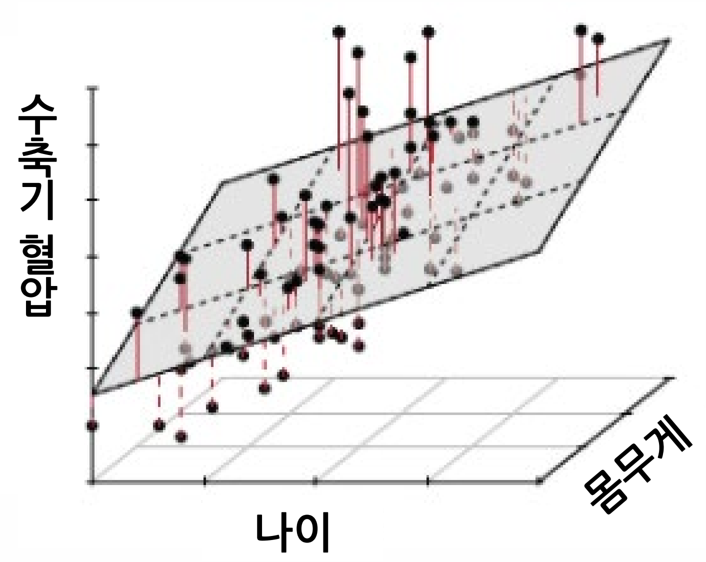
$$MSE(w_1, w_0) = \frac{1}{N_{tst}} \sum_{j=1}^{N_{tst}} \{y_j^{tst} - (w_1 x_j^{tst} + w_0)\}^2$$

  - 평균 제곱근 오차 Root Mean Square Error: RMSE

$$RMSE(w_1, w_0) = \frac{1}{N_{tst}} \sqrt{\sum_{j=1}^{N_{tst}} \{y_j^{tst} - (w_1 x_j^{tst} + w_0)\}^2}$$

다변량 선형회귀

- n차원 입력 벡터 $\boldsymbol{x} = \{x_1, x_2, \cdots, x_n\}$



- 회귀함수 → $f(\boldsymbol{x}) = w_0 + w_1x_1 + w_2x_2 + \cdots + w_nx_n$

다변량 선형회귀

행렬 형태의 표현

- $f(\boldsymbol{x}) = w_0 + w_1 x_1 + w_2 x_2 + \cdots + w_n x_n$

$$\tilde{\boldsymbol{x}} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \quad \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ \vdots \\ w_n \end{bmatrix} \quad \Rightarrow \quad f(\boldsymbol{x}) = [w_0,\ w_1,\ \cdots,\ w_n] \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \boldsymbol{w}^T \tilde{\boldsymbol{x}}$$

데이터 집합 $\{(\boldsymbol{x}_i,\ y_i)\}_{i=1,\,2,\,\cdots,\,N}$ $(\boldsymbol{x}_i \in R^n,\ y_i \in R)$ 의 경우

$$X = \begin{bmatrix} 1,\ \boldsymbol{x}_1{}^T \\ 1,\ \boldsymbol{x}_2{}^T \\ \vdots \\ 1,\ \boldsymbol{x}_N{}^T \end{bmatrix} = \begin{bmatrix} \tilde{\boldsymbol{x}}_1{}^T \\ \tilde{\boldsymbol{x}}_2{}^T \\ \vdots \\ \tilde{\boldsymbol{x}}_N{}^T \end{bmatrix},\ \boldsymbol{y} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{bmatrix} = \boldsymbol{X}\boldsymbol{w}$$

다변량 선형회귀

- 오차함수 → $E(\boldsymbol{w}) = (\boldsymbol{y} - X\boldsymbol{w})^T (\boldsymbol{y} - X\boldsymbol{w})$

- 최적의 파라미터 $\boldsymbol{w}$

$$\frac{\partial E(\boldsymbol{w})}{\partial \boldsymbol{w}} = 2X^T(\boldsymbol{y} - X\boldsymbol{w}) = 2X^T\boldsymbol{y} - 2X^TX\boldsymbol{w} = 0 \rightarrow X^TX\boldsymbol{w} = X^T\boldsymbol{y}$$

양변에 $(\mathbf{X}^T\mathbf{X})^{-1}$ 를 곱하면

$$\boldsymbol{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\boldsymbol{y}$$

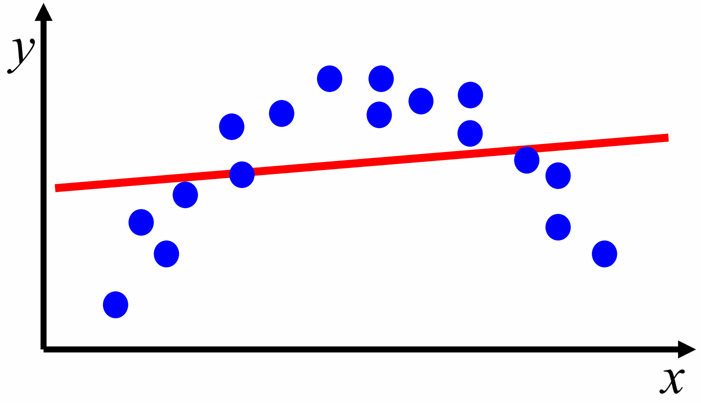
- 새 데이터 $\boldsymbol{x}_{new}$에 대한 예측

$$f(\boldsymbol{x}_{new}) = \boldsymbol{w}^T\widetilde{\boldsymbol{x}}_{new}$$

3 선형회귀의 확장

선형회귀의 한계

- x와 y의 관계를 선형 매핑으로 표현할 수 없는 경우



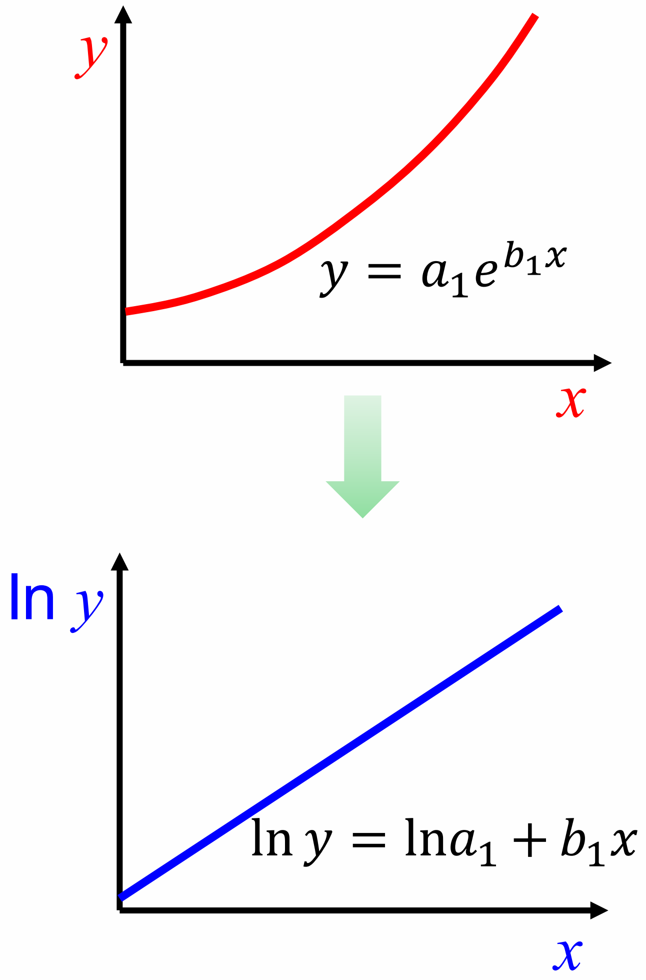
- 선형화 linearization 과정을 거친 후 선형회귀 적용
  - x와 y를 $\tilde{x}$와 $\tilde{y}$로 적절히 변형한 후 선형 매핑 관계 $\tilde{y} = m\tilde{x} + b$를 찾는 방식

선형화_1

○ 지수 형태

$$y = a_1 e^{b_1 x}$$

$$\ln y = \ln a_1 + b_1 x$$

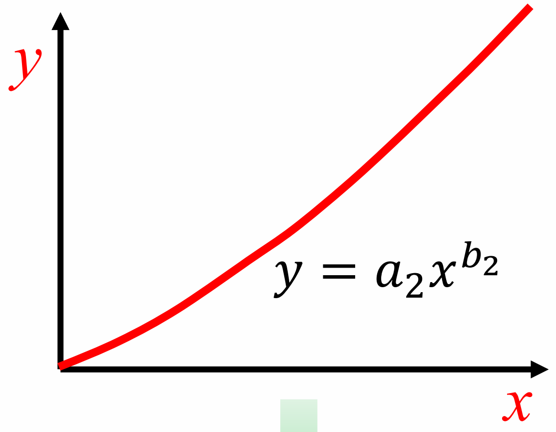


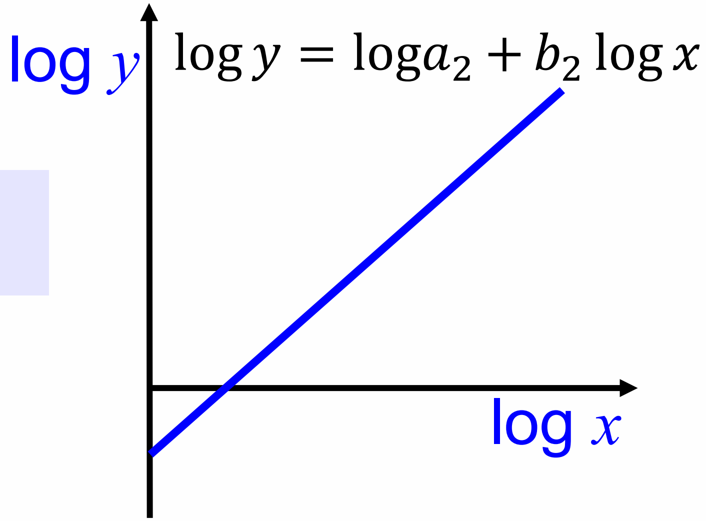
선형화_2

○ 단순 거듭제곱 형태

$$y = a_2 x^{b_2}$$

$$\log y = \log a_2 + b_2 \log x$$



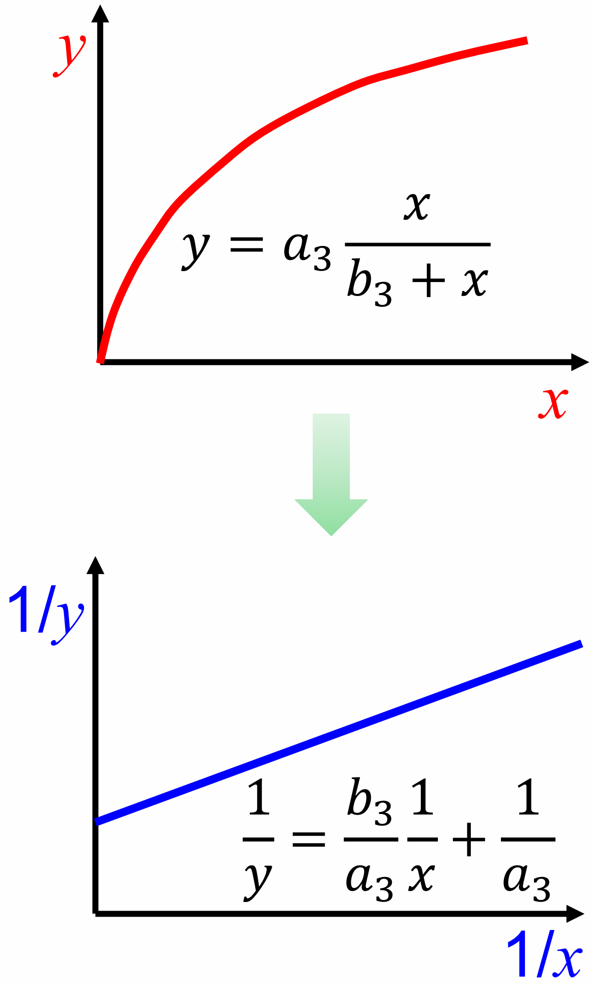


선형화_3

○ 포화된 증가 형태

$$y = a_3 \frac{x}{b_3 + x}$$

$$\frac{1}{y} = \frac{b_3}{a_3}\frac{1}{x} + \frac{1}{a_3}$$



다른 접근법

보다 복잡한 형태의 곡선으로의 매핑을 위한 방법

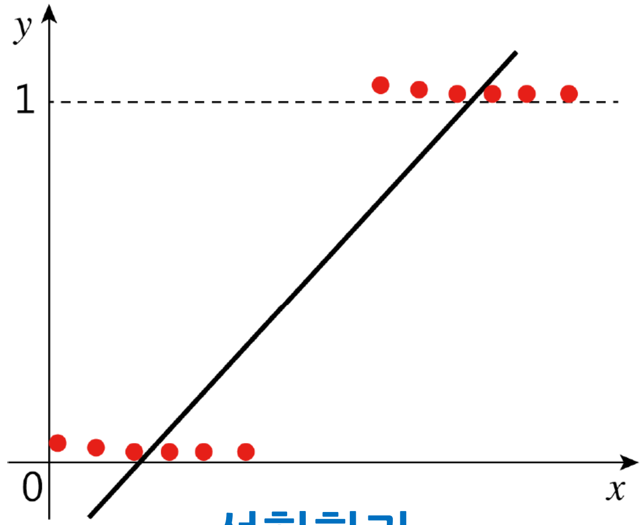
- 다항 회귀 polynomial regression
  - ✓ 고차다항식 사용 → $y(x) = w_0 + w_1 x + w_1 x^2 + \cdots + w_n x^n$
- 비선형 입력 변환함수를 사용한 선형회귀
  - ✓ $y(x) = w_0 + w_1 \phi_1(x) + w_2 \phi_2(x) + \cdots + w_n \phi_n(x)$
  - ✓ $\phi_i(x)$ → 비선형 기저 함수
- 비선형회귀 nonlinear regression
  - ✓ 신경망과 같은 복잡한 비선형함수를 사용하는 방법
  - ✓ 커널을 이용하여 고차원 공간으로 매핑하는 SVM 적용

4 로지스틱 회귀

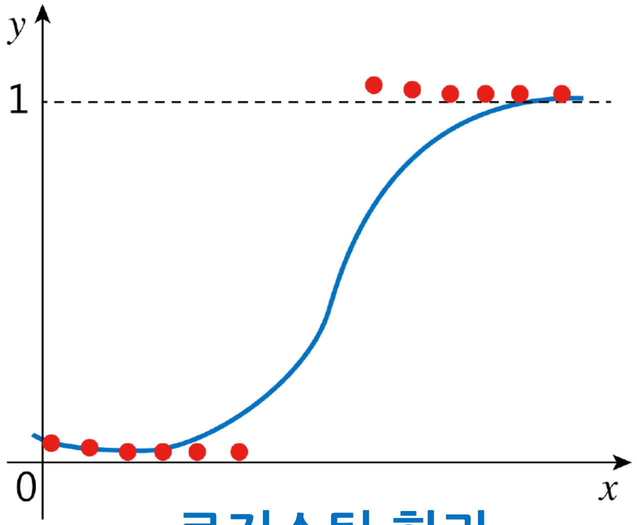
로지스틱 회귀?

○ 범주형 데이터의 회귀 logistic regression

- 선형회귀분석의 종속변수(출력)를 범주형으로 확장한 것
- 분류 문제에 적용 가능
- 입력값이 각 클래스에 속하는 확률값을 회귀분석으로 예측



선형회귀



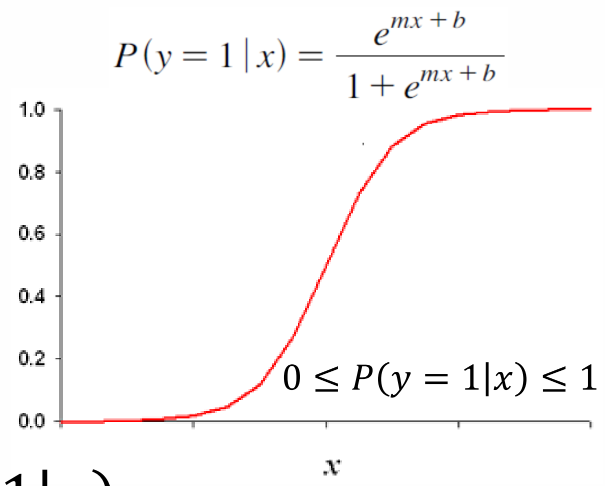
로지스틱 회귀

로지스틱 함수

로지스틱 함수

- $x \in (-\infty, \infty)$를 항상 $(0, 1)$ 범위로 매핑하는 S자형 함수

$$\varphi(x) = \frac{1}{1 + \exp\{-x\}} = \frac{\exp\{x\}}{1 + \exp\{x\}}$$

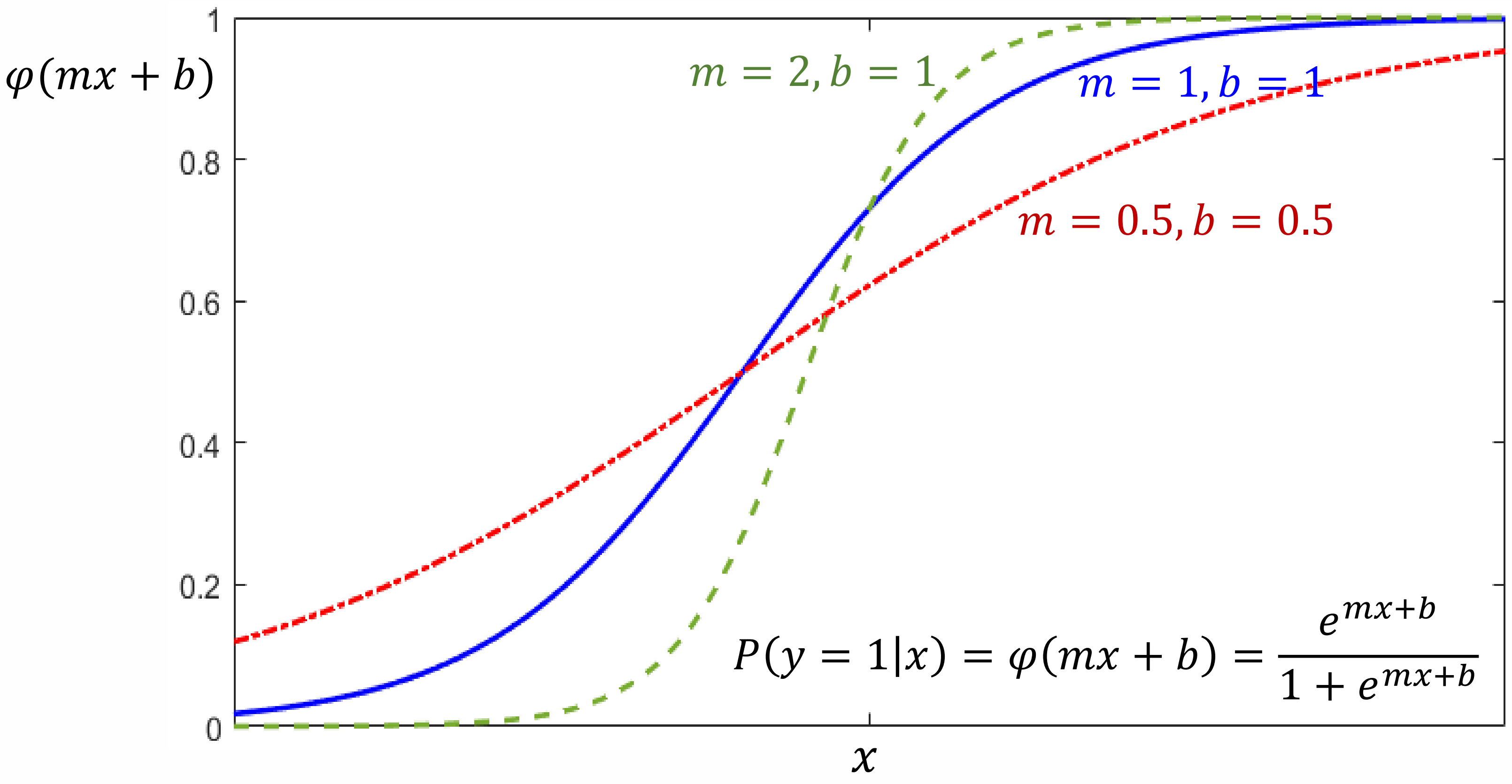


로지스틱 함수를 이용한 분류

- 함수의 출력값 ⇒ 클래스 레이블에 대한 사후확률 $P(y = 1|x)$
- $P(y = 1|x) \le 0.5 \quad \rightarrow \quad x \in C_1$
- $P(y = 1|x) > 0.5 \quad \rightarrow \quad x \in C_2$
- 로지스틱 함수를 이용한 사후확률 추정 $P(y = 1|x) = \dfrac{exp\{mx + b\}}{1 + exp\{mx + b\}}$

로지스틱 함수

- 파라미터 m과 b값에 따른 함수의 형태



오즈비, 로짓 함수, 결정 경계

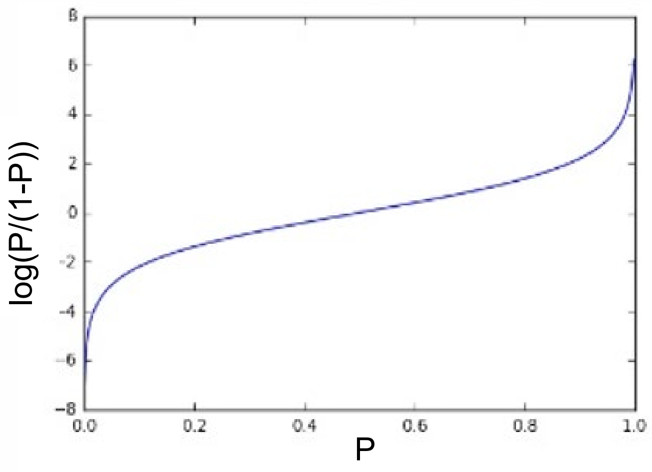
○ 오즈비 odds ratio, 승산비

$$\text{odds} = \frac{P(y=1|x)}{1-P(y=1|x)} = \exp\{mx+b\} \qquad (0 \le \text{odds} \le \infty)$$

○ 로짓 함수 logit function

- 오즈비에 대해 로그를 취한 것

$$\text{logit}(P) = \log\left\{\frac{P(y=1|x)}{1-P(y=1|x)}\right\} = mx+b$$



○ 로지스틱 회귀의 결정경계

$$\text{logit}(P) = \log\left\{\frac{P(y=1|x)}{1-P(y=1|x)}\right\} = mx+b = 0$$

로지스틱 회귀의 매개변수 추정

- 데이터 → $D = \{(x_i, y_i)\}_{i=1,\cdots,N}$ $(y_i \in \{0,1\})$
- $p(y|x)$의 확률함수 → 베르누이 분포를 따름

$$p(y|x) = \{P(y=1|x)\}^y \{1 - P(y=1|x)\}^{1-y}$$

$$= \{\frac{exp\{mx+b\}}{1+exp\{mx+b\}}\}^y \{1 - \frac{exp\{mx+b\}}{1+exp\{mx+b\}}\}^{1-y}$$

- 목적함수 → D에 대한 로그 우도 log likelihood

$$l(m,b) = \log \prod_{i=1}^{N} p(y_i|x_i)$$

$$= \sum_{i=1}^{N} \left(y_i \log\left\{\frac{exp\{mx_i+b\}}{1+exp\{mx_i+b\}}\right\} + (1-y_i)\log\left\{1 - \frac{exp\{mx_i+b\}}{1+exp\{mx_i+b\}}\right\}\right)$$

로지스틱 회귀의 매개변수 추정

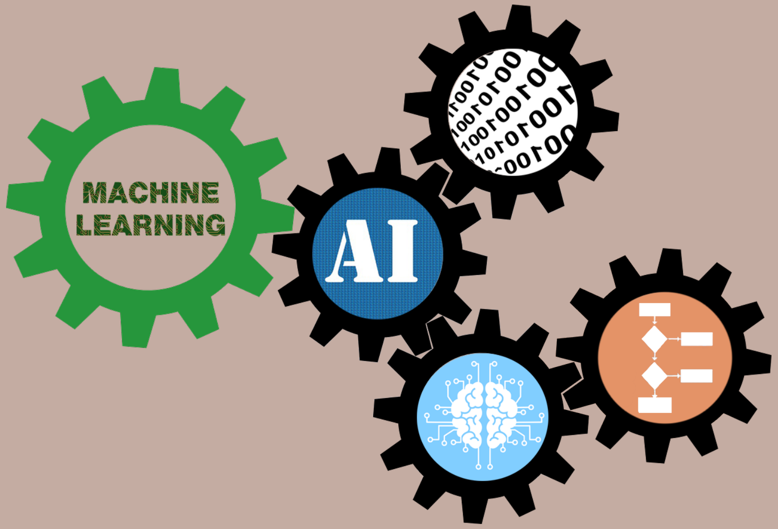
- 추정 → 최대 우도 추정법 maximum likelihood estimation

$$\frac{\partial l(m,b)}{\partial m} = 0, \quad \frac{\partial l(m,b)}{\partial b} = 0$$

  - 수치적 최적화 방법으로 반복적 추정을 통해 최적화

- 파라미터 m과 b의 추정 후 새로운 데이터의 분류 과정

$$\begin{cases} x_{new} \in C_1 & \text{if } mx_{new} + b < 0 \\ x_{new} \in C_2 & \text{if } mx_{new} + b \geq 0 \end{cases}$$



다음시간안내

제4강

비지도학습: 군집화