인 공 지 능

[기계학습 II]

jaekoo@kookmin.ac.kr

국민대학교
KOOKMIN UNIVERSITY

1.3 데이터

1.3 데이터에 대한 이해

과학 기술의 정립 과정

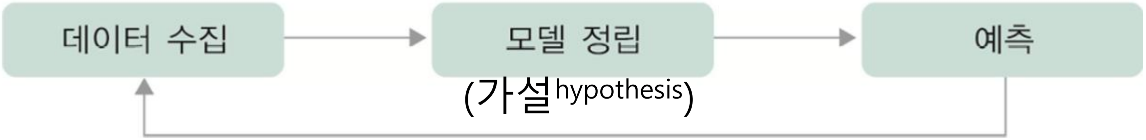


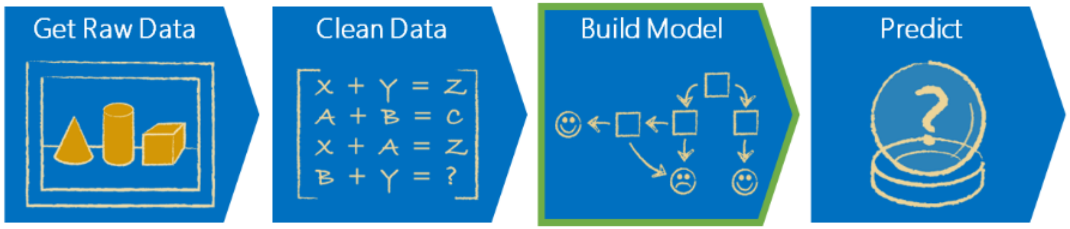
그림 1-8 과학기술의 발전 과정

- 예) Tycho Brahe는 천동설이라는 틀린 모델을 선택하여 수집한 데이터를 설명하지 못함

  Johannes Kepler는 지동설 모델을 도입하여 제1, 제2, 제 3법칙을 완성함

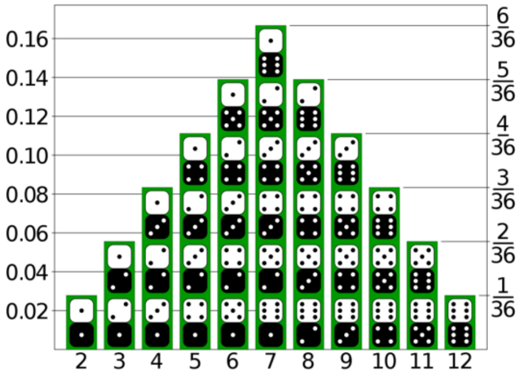
기계학습

- 기계학습은 복잡 문제/과업을 다룸
  - 지능적 범주의 행위들은 규칙의 다양한 변화 양상을 가짐
- 단순한 수학 공식으로 표현 불가능함
- 데이터를 설명할 수 있는 학습 모델을 찾아내는 과정

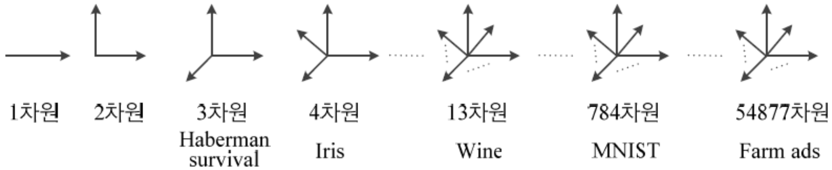


1.3.1 데이터 생성 과정

- 데이터 생성 과정을 완전히 아는 인위적 상황의 예제 (가상)
  - 예) 두 개 주사위를 던져 나온 눈의 합을 x라 할 때, $y=(x-7)^2+1$ 점을 받는 게임
    - 해당 상황은 '데이터 생성 과정을 완전히 알고 있다'고 말함
      - x를 알면 정확히 y를 예측할 수 있음
        - » 실제 주사위를 던져 $\mathbb{X}=\{3,10,8,5\}$를 얻었다면, $\mathbb{Y}=\{17,10,2,5\}$
      - x의 발생 확률 $P(x)$를 정확히 알 수 있음
      - $P(x)$를 알고 있으므로, 새로운 데이터 생성 가능



- [그림 1-6]과 같은 실제 기계 학습 문제 (현실)
  - 데이터 생성 과정을 알 수 없음
  - 단지 주어진 훈련집합 $\mathbb{X}, \mathbb{Y}$로 가설 모델을 통해 근사 추정만 가능



Haberman survival: $\mathbf{x}=$ (나이, 수술년도, 양성 림프샘 개수)^T

Iris: $\mathbf{x}=$ (꽃받침 길이, 꽃받침 너비, 꽃잎 길이, 꽃잎 너비)^T

Wine: $\mathbf{x}=$ (Alcohol, Malic acid, Ash, Alcalinity of ash, Magnesium, Total phenols, Flavanoids, Nonflavanoid phenols Proanthocyanins, Color intensity, Hue, OD280 / OD315 of diluted wines, Proline)^T

MNIST: $\mathbf{x}=$ (화소1, 화소2, ⋯ , 화소784)^T

Farm ads: $\mathbf{x}=$ (단어1, 단어2, ⋯, 단어54877)^T

그림 1-6 다차원 특징 공간

1.3.2 데이터의 중요성

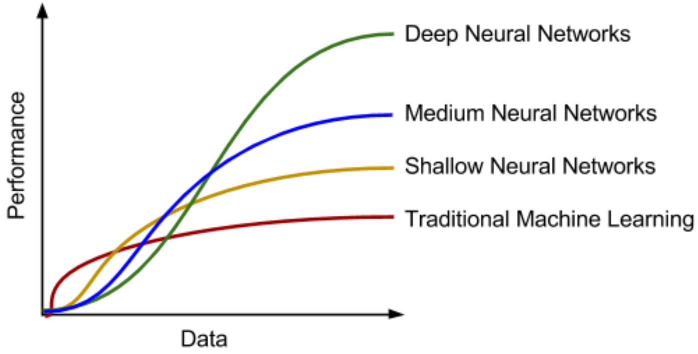
데이터의 양과 질

- 주어진 과업에 적합한 다양한 데이터를 충분한 양만큼 수집 → 과업 성능 향상
  - 예) 정면 얼굴만 가진 데이터로 인식 학습하면,

    측면 얼굴은 매우 낮은 인식 성능을 가짐

  → 주어진 과업에 관련된 데이터 확보는 아주 중요함

- 데이터의 양과 학습 모델의 성능 경향성 비교



공개 데이터

- 기계 학습의 대표적인 3가지 데이터: Iris, MNIST, ImageNet
- UCI 저장소repository (2017년11월 기준으로 394개 데이터 제공)

1.3.2 데이터의 중요성

- Iris 데이터베이스는 통계학자인 피셔 교수가 1936년에 캐나다 동부 해안의 가스페 반도에 서식하는 3종의 붓꽃(setosa, versicolor, virginica)을 50송이씩 채취하여 만들었다[Fisher1936]. 150개 샘플 각각에 대해 꽃받침 길이, 꽃받침 너비, 꽃잎 길이, 꽃잎 너비를 측정하여 기록하였다. 따라서 4차원 특징 공간이 형성되며 목푯값은 3종을 숫자로 표시함으로써 1, 2, 3 값 중의 하나이다. http://archive.ics.uci.edu/ml/datasets/Iris에 접속하여 내려받을 수 있다.

Sepal length	Sepal width	Petal length	Petal width	Species
5.2	3.5	1.4	0.2	*I. setosa*
4.9	3.0	1.4	0.2	*I. setosa*
4.7	3.2	1.3	0.2	*I. setosa*
4.6	3.1	1.5	0.2	*I. setosa*
7.0	3.2	4.7	1.4	*I. versicolor*
6.4	3.2	4.5	1.5	*I. versicolor*
6.9	3.1	4.9	1.5	*I. versicolor*
5.5	2.3	4.0	1.3	*I. versicolor*
6.3	3.3	6.0	2.5	*I. virginica*
5.8	2.7	5.1	1.9	*I. virginica*
7.1	3.0	5.9	2.1	*I. virginica*
6.3	2.9	5.6	1.8	*I. virginica*

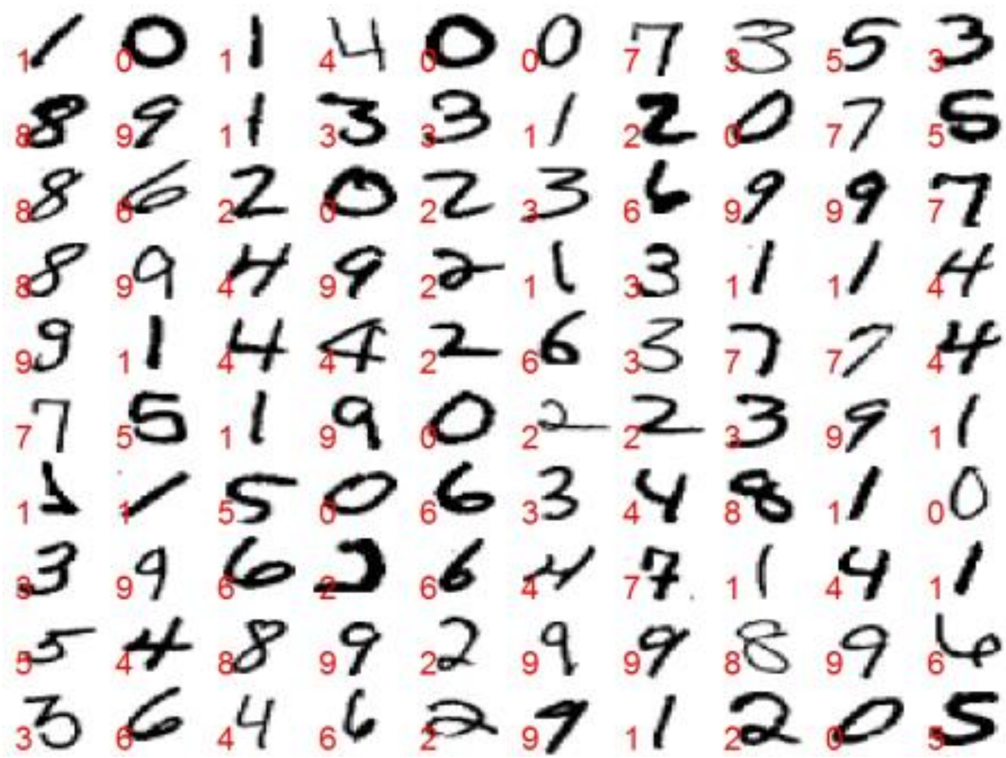

Setosa


Versicolor


Virginica

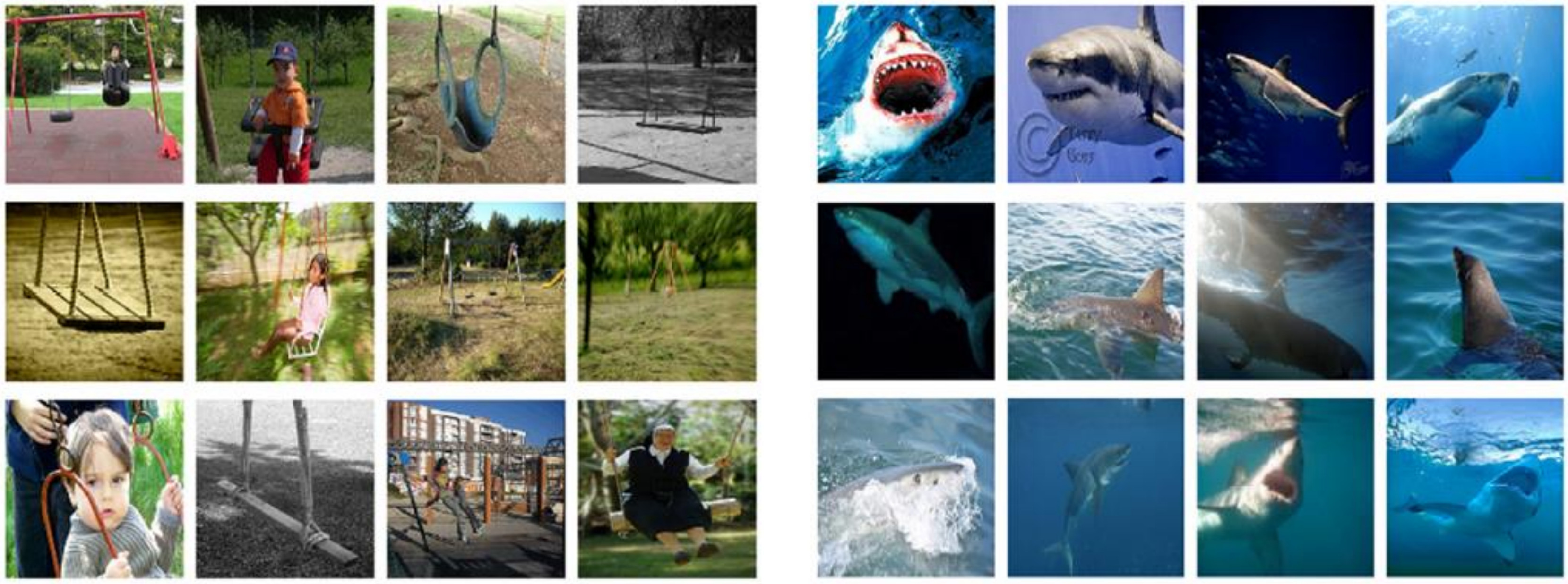
1.3.2 데이터의 중요성

- MNIST 데이터베이스는 미국표준국(NIST)에서 수집한 필기 숫자 데이터베이스로, 훈련집합 60,000자, 테스트집합 10,000자를 제공한다. http://yann.lecun.com/exdb/mnist에 접속하면 무료로 내려받을 수 있으며, 1988년부터 시작한 인식률 경쟁 기록도 볼 수 있다. 2017년 8월 기준으로는 [Ciresan2012] 논문이 0.23%의 오류율로 최고 자리를 차지하고 있다. 테스트집합에 있는 10,000개 샘플에서 단지 23개만 틀린 것이다.



1.3.2 데이터의 중요성

- ImageNet 데이터베이스는 정보검색 분야에서 만든 WordNet의 단어 계층 분류를 그대로 따랐고, 부류마다 수백에서 수천 개의 영상을 수집하였다[Deng2009]. 총 21,841개 부류에 대해 총 14,197,122개의 영상을 보유하고 있다. 그중에서 1,000개 부류를 뽑아 ILSVRC ImageNet Large Scale Visual Recognition Challenge라는 영상인식 경진대회를 2010년부터 매년 개최하고 있다. 대회 결과에 대한 자세한 내용은 4.4절을 참조하라. http:// image-net.org에서 내려받을 수 있다.



(a) 'swing' 부류 (b) 'Great white shark' 부류

그림 4-20 ImageNet의 예제 영상

1.3.3 데이터베이스 크기와 기계 학습 성능

- 데이터의 적은 양 → 차원의 저주와 관련
  - 예) MNIST: 28*28 단순히 흑백으로 구성된다면 서로 다른 총 샘플 수는 2^{784}가지이지만, MNIST는 고작 6만 개 샘플

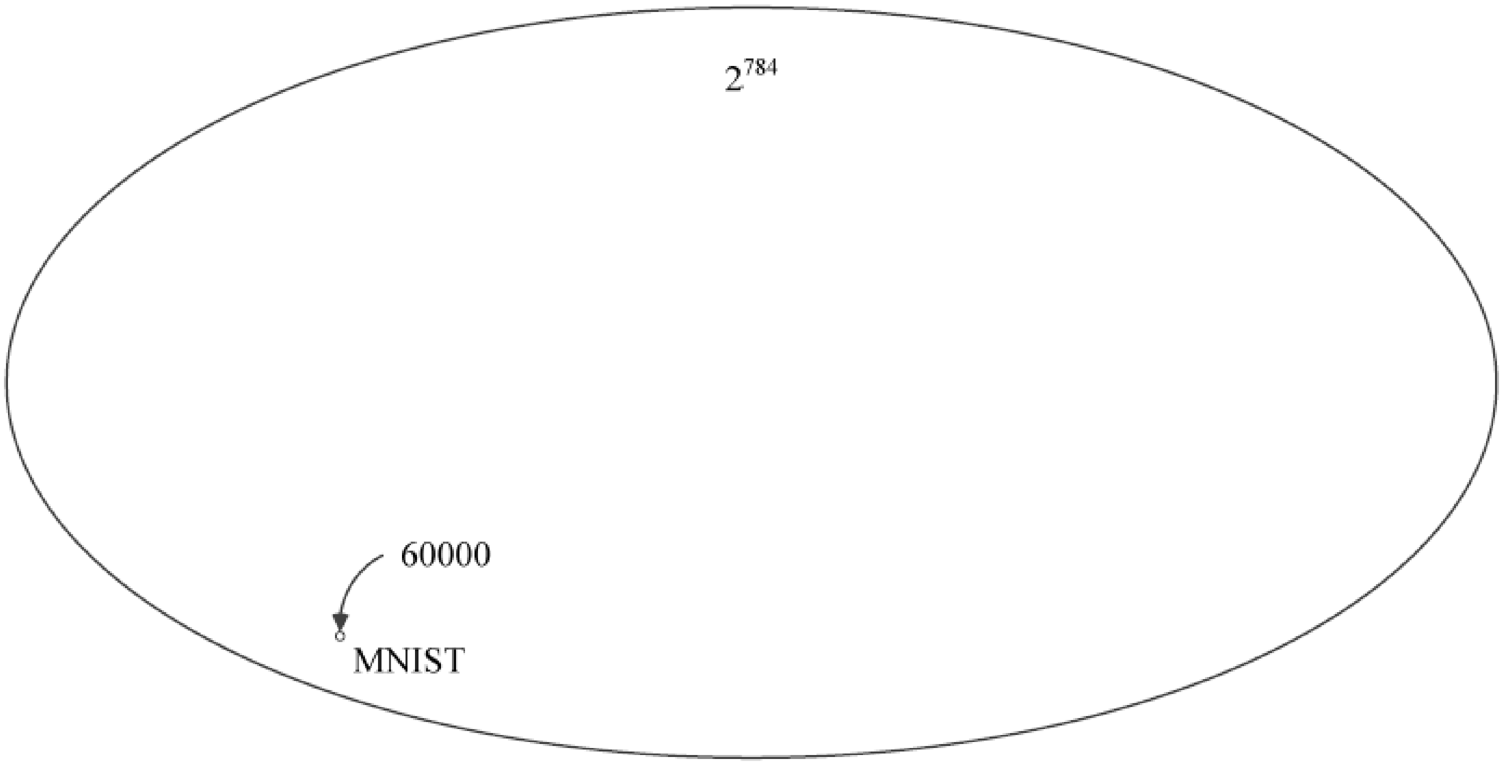


그림 1-9 방대한 특징 공간과 희소한 데이터베이스

1.3.3 데이터베이스 크기와 기계 학습 성능

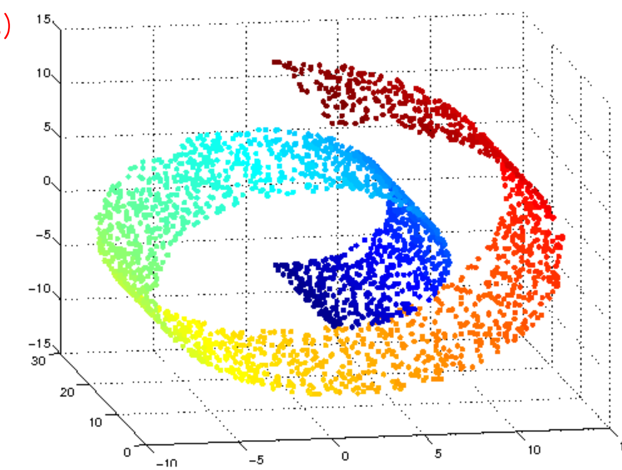
적은 양의 데이터베이스로 어떻게 높은 성능을 달성하는가?

- 방대한 공간에서 실제 데이터가 발생하는 곳은 매우 작은 부분 공간임

→ 데이터 희소data sparsity 특성 가정

  -  ,  와 같은 데이터 발생 확률은 거의 0 가까움

- 매니폴드(많이+끼다) 가정manifold assumption (or manifold hypothesis)



  - 고차원의 데이터는

    관련된 낮은 차원의 매니폴드에 가깝게 집중되어 있음

  -  와 같이 일정한 규칙에 따라 매끄럽게 변화

1.3.4 데이터 가시화

- 4차원 이상의 초공간hyperplane은 한꺼번에 가시화visualization 불가능
- 여러 가지 가시화 기법
  - 2개씩 조합하여 여러 개의 그래프 그림

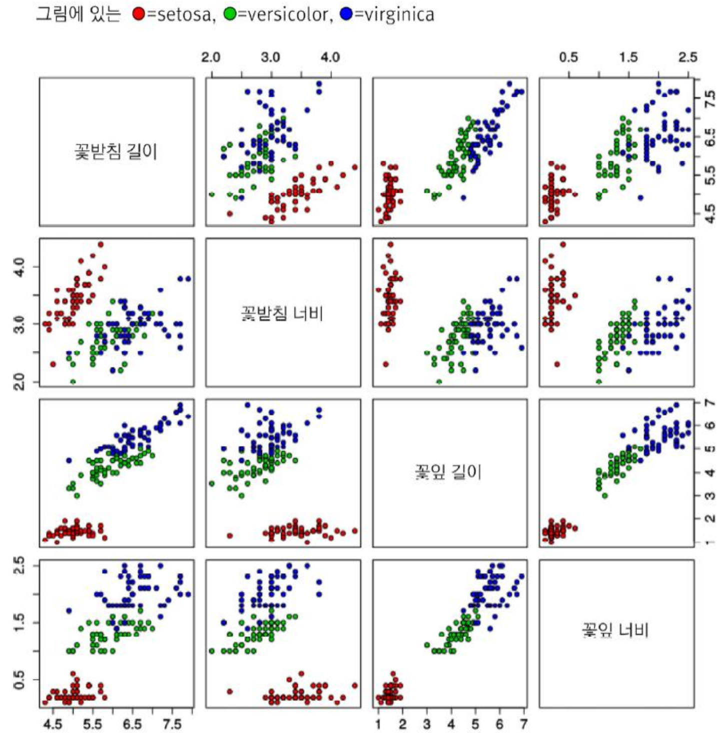


그림 1-10 고차원 특징 공간의 가시화

  - 고차원 공간을 저차원으로 변환하는 기법들 (6.6.1절)

1.4 간단한 기계 학습의 예

1.4 간단한 기계 학습의 예

- 선형 회귀linear regression 문제
  - [그림 1-4]: 식 (1.2)의 직선 모델(가설)을 사용하므로 두 개의 매개변수 $\Theta = (w, b)^{\mathrm{T}}$

$$y = wx + b \tag{1.2}$$

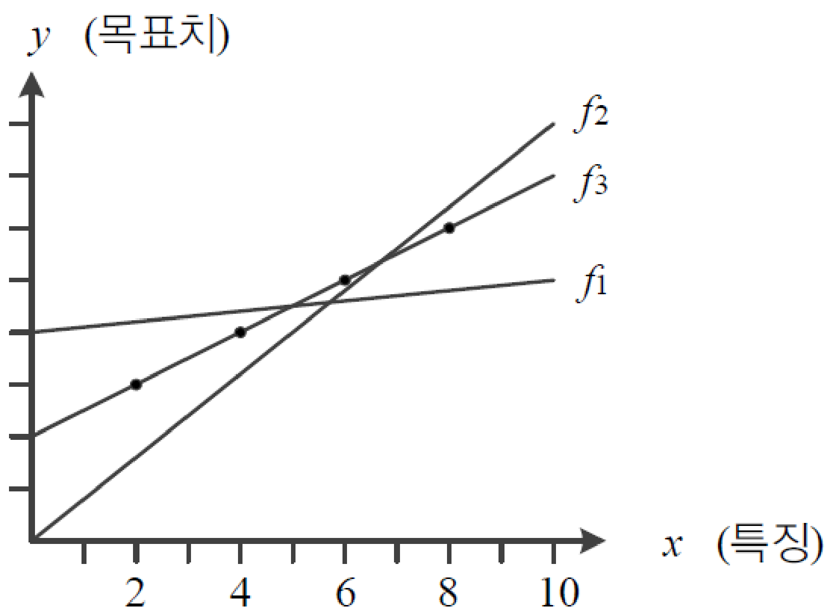


그림 1-4 간단한 기계 학습 예제

1.4 간단한 기계 학습의 예

- **목적 함수**objective function (또는 **비용 함수**cost function)
  - 식 (1.8)은 선형 회귀를 위한 목적 함수
    - 식 (1.8)을 평균제곱오차MSE(mean squared error)라 부름

$$J(\Theta) = \frac{1}{n}\sum_{i=1}^{n}(f_\Theta(\mathbf{x}_i) - y_i)^2 \tag{1.8}$$

    - $f_\Theta(\mathbf{x}_i)$는 예측함수의 예측 출력, y_i는 예측함수가 맞추어야 하는 실제 목표치
    - $f_\Theta(\mathbf{x}_i) - y_i$는 오차error 혹은 손실loss

  - 처음에는 최적 매개변수 값을 알 수 없으므로 임의의 난수로 $\Theta_1 = (w_1, b_1)^\mathrm{T}$ 설정

    → $\Theta_2 = (w_2, b_2)^\mathrm{T}$ 로 개선 → $\Theta_3 = (w_3, b_3)^\mathrm{T}$ 로 개선 → Θ_3는 최적해 $\hat{\Theta}$
    - 이때 $J(\Theta_1) > J(\Theta_2) > J(\Theta_3)$

1.4 간단한 기계 학습의 예

[예제 1-1]

- 훈련집합

  $\mathbb{X} = \{x_1 = (2.0), x_2 = (4.0),\ x_3 = (6.0),\ x_4 = (8.0)\}$,

  $\mathbb{Y} = \{y_1 = 3.0,\ y_2 = 4.0,\ y_3 = 5.0,\ y_4 = 6.0\}$

- 초기 직선의 매개변수 $\Theta_1 = (0.1, 4.0)^{\mathrm{T}}$라 가정

$$
\begin{aligned}
\mathbf{x}_1, y_1 &\longrightarrow \left(f_{\Theta_1}(2.0) - 3.0\right)^2 = \left((0.1 * 2.0 + 4.0) - 3.0\right)^2 = 1.44 \\
\mathbf{x}_2, y_2 &\longrightarrow \left(f_{\Theta_1}(4.0) - 4.0\right)^2 = \left((0.1 * 4.0 + 4.0) - 4.0\right)^2 = 0.16 \\
\mathbf{x}_3, y_3 &\longrightarrow \left(f_{\Theta_1}(6.0) - 5.0\right)^2 = \left((0.1 * 6.0 + 4.0) - 5.0\right)^2 = 0.16 \\
\mathbf{x}_4, y_4 &\longrightarrow \left(f_{\Theta_1}(8.0) - 6.0\right)^2 = \left((0.1 * 8.0 + 4.0) - 6.0\right)^2 = 1.44
\end{aligned}
\longrightarrow \quad J(\Theta_1) = 0.8
$$

1.4 간단한 기계 학습의 예

[예제 1-1] (계속)

- Θ_1을 개선하여 $\Theta_2 = (0.8, 0.0)^T$가 되었다고 가정

$$\mathbf{x}_1, \mathrm{y}_1 \rightarrow \left(f_{\Theta_2}(2.0) - 3.0\right)^2 = \left((0.8 * 2.0 + 0.0) - 3.0\right)^2 = 1.96$$
$$\mathbf{x}_2, \mathrm{y}_2 \rightarrow \left(f_{\Theta_2}(4.0) - 4.0\right)^2 = \left((0.8 * 4.0 + 0.0) - 4.0\right)^2 = 0.64$$
$$\mathbf{x}_3, \mathrm{y}_3 \rightarrow \left(f_{\Theta_2}(6.0) - 5.0\right)^2 = \left((0.8 * 6.0 + 0.0) - 5.0\right)^2 = 0.04$$
$$\mathbf{x}_4, \mathrm{y}_4 \rightarrow \left(f_{\Theta_2}(8.0) - 6.0\right)^2 = \left((0.8 * 8.0 + 0.0) - 6.0\right)^2 = 0.16$$

$$\longrightarrow \quad J(\Theta_2) = 0.7$$

- 다음으로 Θ_2를 개선하여 $\Theta_3 = (0.5, 2.0)^T$가 되었다고 가정

- 이때 $J(\Theta_3) = 0.0$이 되어 Θ_3은 최적값 $\hat{\Theta}$ 이 됨

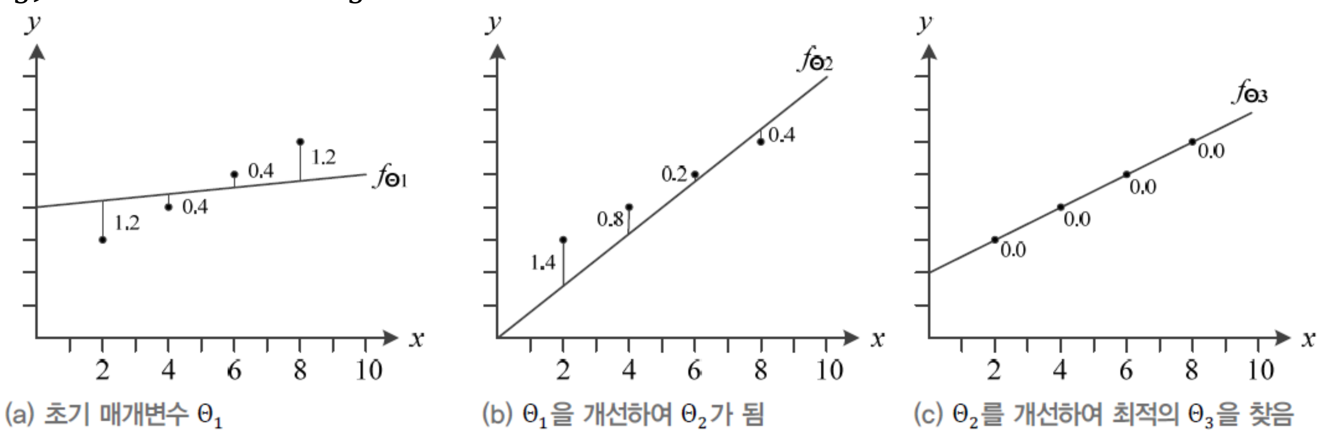
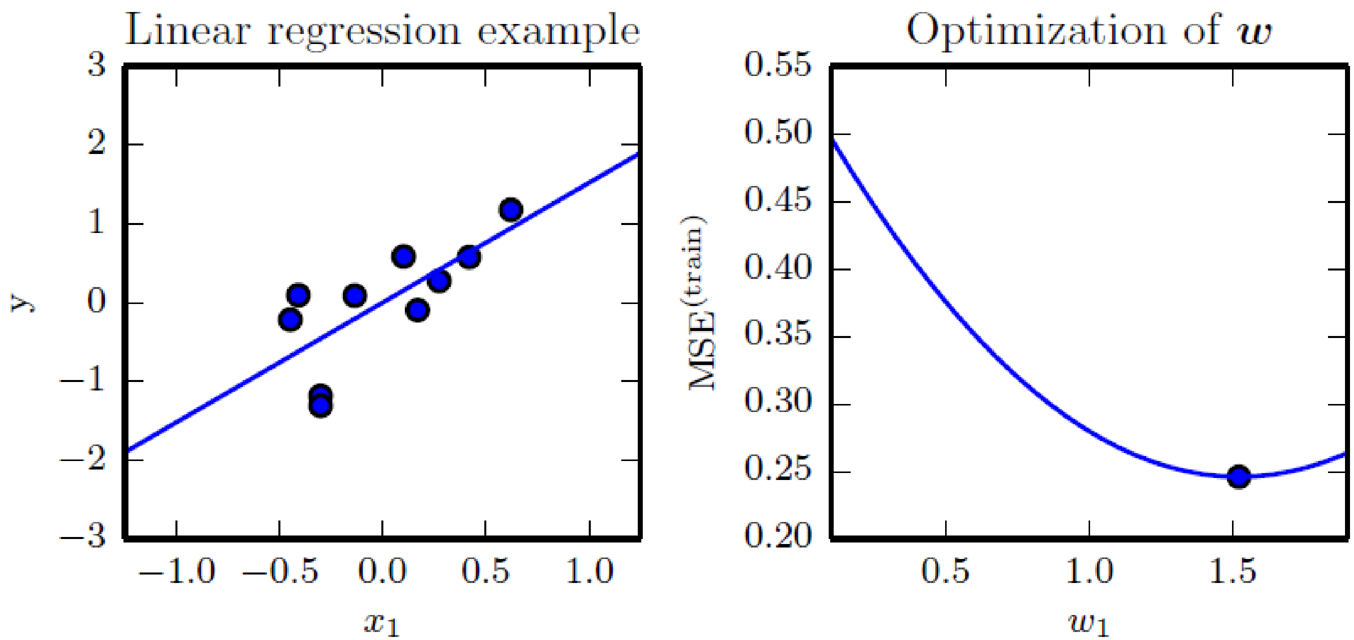


그림 1-11 기계 학습에서 목적함수의 역할

1.4 간단한 기계 학습의 예

- 선형 회귀 문제와 매개변수 최적화 관계의 예



1.4 간단한 기계 학습의 예

- 기계 학습이 할 일을 공식화하면,

$$\hat{\Theta} = \underset{\Theta}{\operatorname{argmin}} J(\Theta) \tag{1.9}$$

  - 기계 학습은 작은 개선을 반복하여 최적의 해를 찾아가는 수치적 방법으로 식 (1.9)를 품

- 알고리즘 형식으로 쓰면,

알고리즘 1-1 기계 학습 알고리즘

입력 : 훈련집합 $\mathbb{X}$와 $\mathbb{Y}$
출력 : 최적의 매개변수 $\hat{\Theta}$

난수를 생성하여 초기 해 Θ_1을 설정한다.
t=1
while ($J(\Theta_t)$가 0.0에 충분히 가깝지 않음) // 수렴 여부 검사
    $J(\Theta_t)$가 작아지는 방향 $\Delta\Theta_t$를 구한다. // $\Delta\Theta_t$는 주로 미분을 사용하여 구함
    $\Theta_{t+1} = \Theta_t + \Delta\Theta_t$
    t=t+1
$\hat{\Theta} = \Theta_t$

1.4 간단한 기계 학습의 예

좀더 현실적인 상황

- 지금까지는 데이터가 선형을 이루는 아주 단순한 상황을 고려함
- 실제 세계는 선형이 아니며 잡음이 섞임 → 비선형 모델이 필요

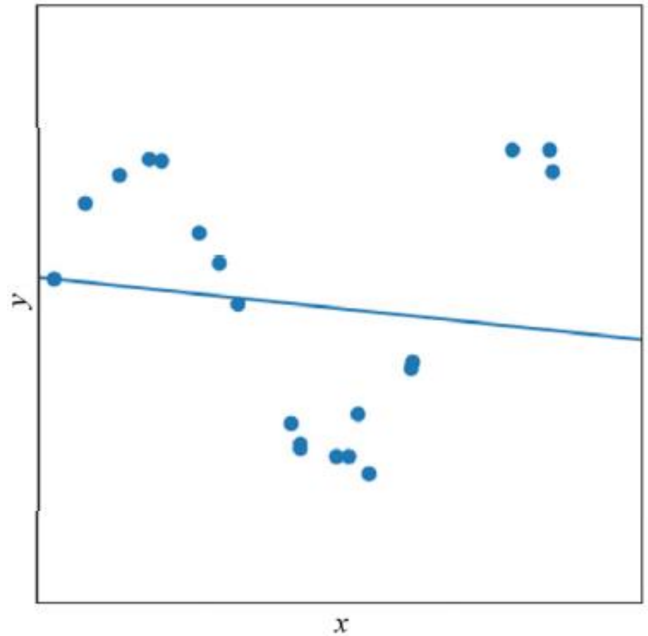


그림 1-12 선형 모델의 한계