머신 러닝 기본개념

Hee-il Hahn

Professor
Department of Information and Communications Engineering
Hankuk University of Foreign Studies
hihahn@hufs.ac.kr

1.1.3 기계 학습 개념

- 훈련집합
  - 가로축은 특징, 세로축은 목표치
  - 관측한 4개의 점이 훈련집합을 구성함

$$\text{훈련집합: } \mathbb{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n\}, \quad \mathbb{Y} = \{y_1, y_2, \cdots, y_n\} \tag{1.1}$$

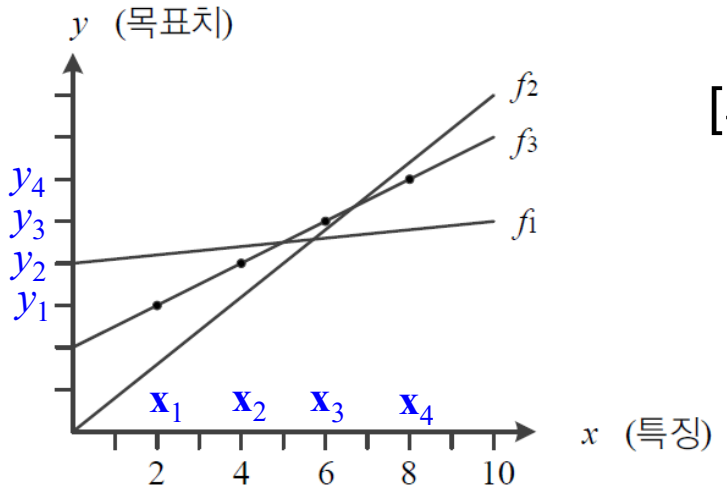


그림 1-4 간단한 기계 학습 예제

[그림 1-4] 예제의 훈련집합

$$\mathbb{X} = \{\mathbf{x}_1 = (2.0), \mathbf{x}_2 = (4.0), \mathbf{x}_3 = (6.0), \mathbf{x}_4 = (8.0)\}$$
$$\mathbb{Y} = \{y_1 = 3.0, y_2 = 4.0, y_3 = 5.0, y_4 = 6.0\}$$

1.1.3 기계 학습 개념

- 학습을 마치면,
  - 예측에 사용
  - 예) 10.0 순간의 이동체 위치를 알고자 하면, $f_3(10.0)=0.5*10.0+2.0=7.0$이라 예측함

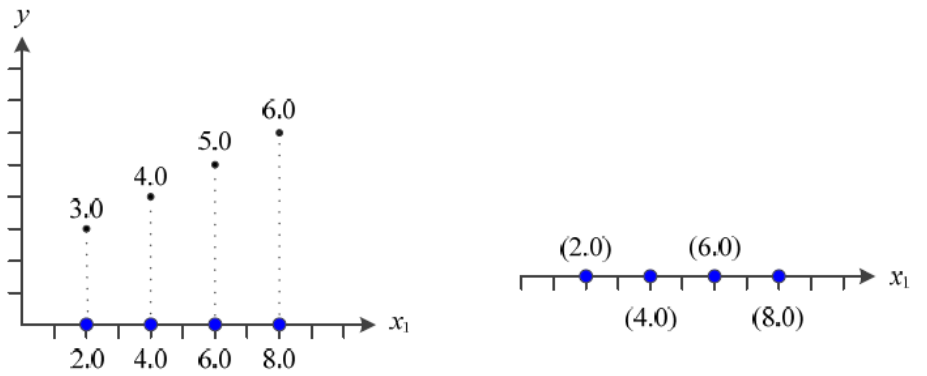
- 기계 학습의 궁극적인 목표
  - 훈련집합에 없는 새로운 샘플에 대한 오류를 최소화 (새로운 샘플 집합: 테스트 집합)
  - 테스트 집합에 대한 높은 성능을 일반화generalization 능력이라 부름

1.2 특징 공간에 대한 이해

- 1.2.1 1차원과 2차원 특징 공간
- 1.2.2 다차원 특징 공간
- 1.2.3 특징 공간 변환과 표현 학습

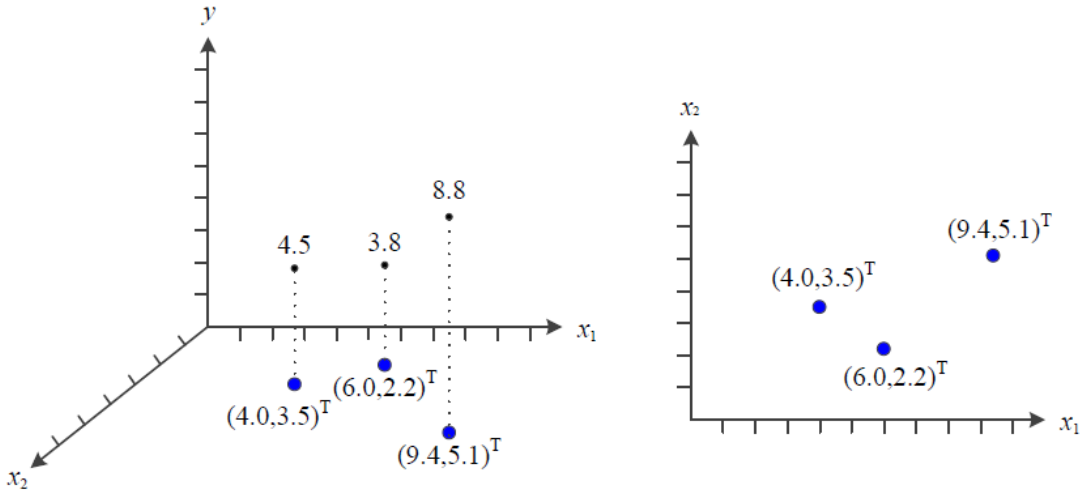
1.2.1 1차원과 2차원 특징 공간

- 1차원 특징 공간 →



(a) 1차원 특징 공간(왼쪽: 특징과 목푯값을 축으로 표시, 오른쪽: 특징만 축으로 표시)

- 2차원 특징 공간 →
  - 특징 벡터 표기
    - $\mathbf{x}=(x_1,x_2)^T$
  - 예시
    - $\mathbf{x}$=(몸무게,키)^T, y=장타율
    - $\mathbf{x}$=(체온,두통)^T, y=감기 여부

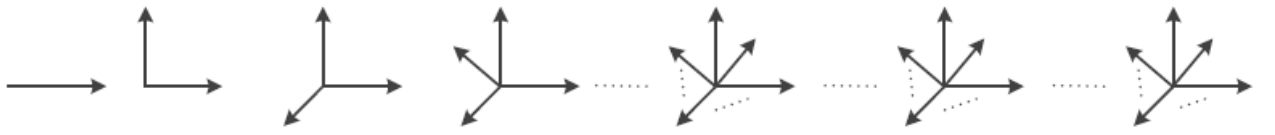


(b) 2차원 특징 공간(왼쪽: 특징 벡터와 목푯값을 축으로 표시, 오른쪽: 특징 벡터만 축으로 표시)

그림 1-5 특징 공간과 데이터의 표현

1.2.2 다차원 특징 공간

- 다차원 특징 공간 예제



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

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

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

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

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

그림 1-6 다차원 특징 공간

1.2.2 다차원 특징 공간

- *d*-차원 데이터
  - 특징 벡터 표기: $\mathbf{x}=(x_1,x_2, \ldots ,x_d)^T$
- *d*-차원 데이터를 위한 학습 모델
  - 직선 모델을 사용하는 경우 매개변수 수=d+1

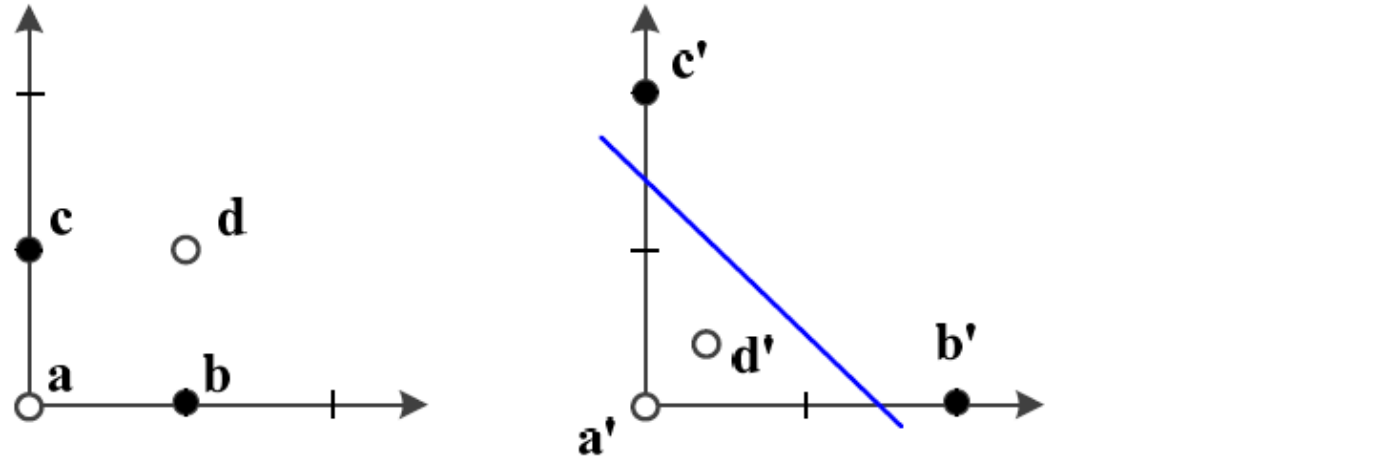
$$y = w_1 x_1 + w_2 x_2 + \cdots + w_d x_d + b \tag{1.3}$$

  - 2차 곡선 모델을 사용하면 매개변수 수가 크게 증가
    - 매개변수 수=d^2+d+1
    - 예) Iris 데이터: d=4이므로 21개의 매개변수
    - 예) MNIST 데이터: d=784이므로 615,441개의 매개변수

$$y = w_1 x_1^2 + w_2 x_2^2 + \cdots + w_d x_d^2 + w_{d+1} x_1 x_2 + \cdots + w_{d^2} x_{d-1} x_d + w_{d^2+1} x_1 + \cdots + w_{d^2+d} x_d + b \tag{1.5}$$

1.2.3 특징 공간 변환과 표현 학습

- 선형 분리 불가능linearly non-separable한 원래 특징 공간 ([그림 1-7(a)])
  - 직선 모델을 적용하면 75% 정확률이 한계



(a) 원래 특징 공간 (b) 분류에 더 유리하도록 변환된 새로운 특징 공간

그림 1-7 특징 공간 변환

1.4 간단한 기계 학습의 예

선형 회귀 문제

- [그림 1-4]: 식 (1.2)의 직선 모델을 사용하므로 두 개의 매개변수 $\Theta = (w, b)^{\mathrm{T}}$

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

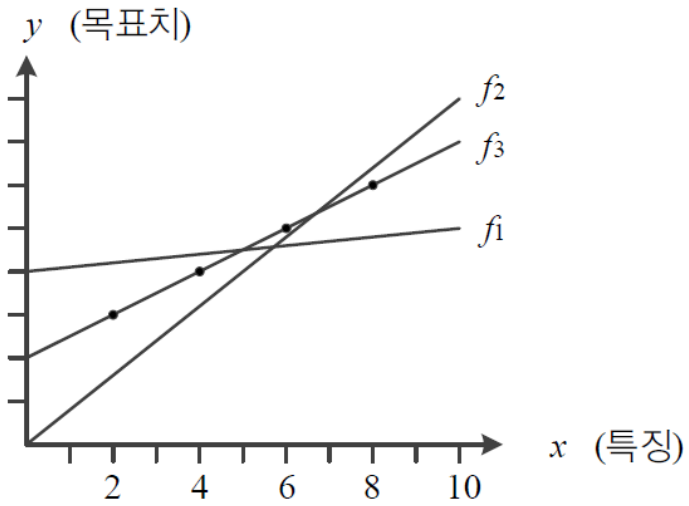


그림 1-4 간단한 기계 학습 예제

1.4 간단한 기계 학습의 예

- 목적 함수objective function (또는 비용 함수cost function)
  - 식 (1.8)은 선형 회귀를 위한 목적 함수
    - $f_\Theta(\mathbf{x}_i)$는 예측함수의 출력, y_i는 예측함수가 맞추어야 하는 목푯값이므로 $f_\Theta(\mathbf{x}_i)$-y_i는 오차
    - 식 (1.8)을 평균제곱오차MSE(mean squared error)라 부름

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

  - 처음에는 최적 매개변수 값을 알 수 없으므로 난수로 $\Theta_1 = (w_1, b_1)^T$ 설정 → $\Theta_2 = (w_2, b_2)^T$ 로 개선 → $\Theta_3 = (w_3, b_3)^T$ 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 $\Theta_1 = (0.1, 4.0)^T$라 가정

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

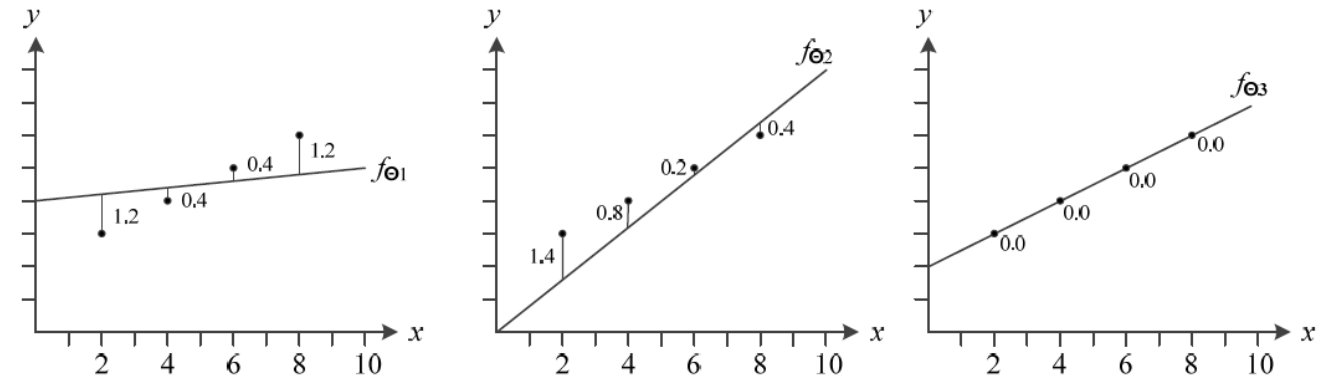
1.4 간단한 기계 학습의 예

[예제 1-1] 훈련집합

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

$$\begin{aligned}
\mathbf{x}_1, y_1 &\rightarrow \left(f_{\Theta_2}(2.0) - 3.0\right)^2 = \left((0.8 * 2.0 + 0.0) - 3.0\right)^2 = 1.96 \\
\mathbf{x}_2, y_2 &\rightarrow \left(f_{\Theta_2}(4.0) - 4.0\right)^2 = \left((0.8 * 4.0 + 0.0) - 4.0\right)^2 = 0.64 \\
\mathbf{x}_3, y_3 &\rightarrow \left(f_{\Theta_2}(6.0) - 5.0\right)^2 = \left((0.8 * 6.0 + 0.0) - 5.0\right)^2 = 0.04 \\
\mathbf{x}_4, y_4 &\rightarrow \left(f_{\Theta_2}(8.0) - 6.0\right)^2 = \left((0.8 * 8.0 + 0.0) - 6.0\right)^2 = 0.16
\end{aligned} \quad \rightarrow \quad J(\Theta_2) = 0.7$$

- 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



(a) 초기 매개변수 Θ_1 (b) Θ_1을 개선하여 Θ_2가 됨 (c) Θ_2를 개선하여 최적의 Θ_3을 찾음

그림 1-11 기계 학습에서 목적함수의 역할

1.4 간단한 기계 학습의 예

- 기계 학습이 할 일을 공식화하면,

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

  - 기계 학습은 작은 개선을 반복하여 최적해를 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 $\hat{\Theta}$

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

Overfitting and Underfitting

- Overfitting(과대적합)
  - 가진 정보를 모두 사용해서 너무 복잡한 모델을 만드는 것
  - 모델이 학습데이터의 각 샘플에 너무 가깝게 맞춰져서 새로운 데이터에 일반화되기 어려울 때 발생
- Underfitting(과소적합)
  - 반대로 모델이 너무 간단하면 데이터의 면면과 다양성을 잡아내지 못해 학습데이터에도 잘 맞지 않을 수 있다.

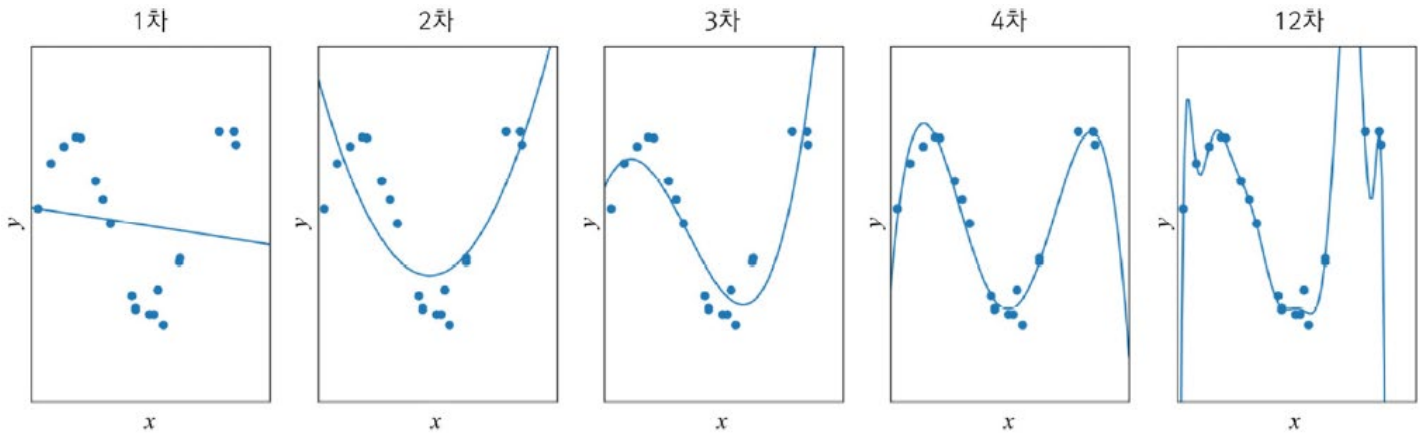


그림 1-13 과소적합과 과잉적합 현상

Overfitting and Underfitting - cont.

- Overfitting(과대적합)
  - 12차 다항식 곡선을 채택한다면 훈련집합에 대해 거의 완벽하게 근사화함
  - 하지만 '새로운' 데이터를 예측한다면 큰 문제 발생
  - x_0에서 빨간 막대 근방을 예측해야 하지만 빨간 점을 예측
  - 이유는 '용량이 크기' 때문. 학습 과정에서 잡음까지 수용 → 과잉적합 현상

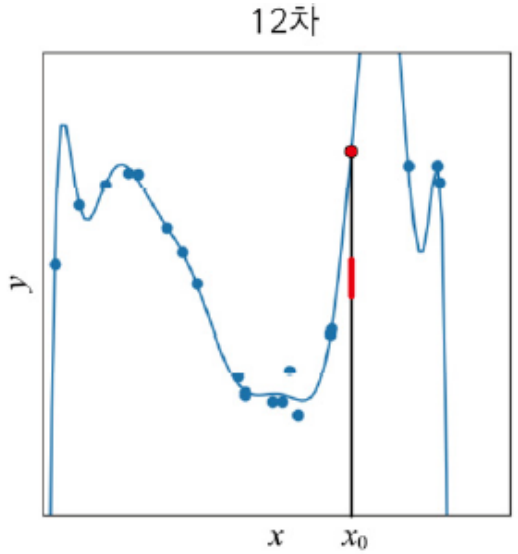


그림 1-14 과잉적합되었을 때 부정확한 예측 현상

Bias and Variance

- Bias
  - $Bias(\hat{\theta}_m) = \mathbb{E}(\hat{\theta}_m) - \theta$, $\hat{\theta}_m$: estimator, θ: true value
  - Said to be unbiased if $bias(\hat{\theta}_m) = 0$.

  - Example: Gaussian distribution estimator of the Mean

    A set of samples $\{x^{(1)} \quad \dots \quad x^{(m)}\}$: iid (independently and identically distributed)

$$p(x^{(i)}; \mu, \sigma) = \sqrt{\frac{1}{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}\left(x^{(i)} - \mu\right)^2\right)$$

$$\hat{\mu}_m = \frac{1}{m}\sum_i^m x^{(i)}$$

$$\begin{aligned} Bias(\hat{\mu}_m) &= \mathbb{E}(\hat{\mu}_m) - \mu \\ &= \mathbb{E}\left(\frac{1}{m}\sum_i^m x^{(i)}\right) - \mu \\ &= \frac{1}{m}\sum_i^m \mathbb{E}\left(x^{(i)}\right) - \mu \\ &= \frac{1}{m}\sum_i^m \mu - \mu = 0 \end{aligned}$$

Bias and Variance – cont.

– Example: Estimators of the variance of Gaussian distribution

A set of samples $\{x^{(1)} \quad \dots \quad x^{(m)}\}$: iid (independently and identically distributed)

i) $\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2, \qquad \hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}$

$$Bias(\hat{\sigma}_m^2) = \mathbb{E}(\hat{\sigma}_m^2) - \sigma^2 = \frac{m-1}{m}\sigma^2 = -\frac{1}{m}\sigma^2 \Rightarrow Biased\ estimator$$

$$\mathbb{E}(\hat{\sigma}_m^2) = \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right)$$

$$= \mathbb{E}\left(\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \mu + \mu - \hat{\mu}_m\right)^2\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(\left(x^{(i)} - \mu + \mu - \hat{\mu}_m\right)^2\right)$$

$$= \frac{1}{m}\sum_{i=1}^{m}\mathbb{E}\left(\left(x^{(i)} - \mu\right)^2 + 2\left(x^{(i)} - \mu\right)(\mu - \hat{\mu}_m) + (\mu - \hat{\mu}_m)^2\right)$$

$$= \frac{1}{m}(m\sigma^2 - 2\sigma^2 + \sigma^2) = \frac{m-1}{m}\sigma^2$$

$$\text{Un}biased\ estimator: \quad \hat{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2$$

Bias and Variance – cont.

– Example: Estimators of the variance of Gaussian distribution

$$\mathbb{E}\left(\left(x^{(i)}-\mu\right)(\mu-\hat{\mu}_m)\right) = -\mathbb{E}\left(\left(x^{(i)}-\mu\right)\sum_{j=1}^{m}\frac{x^{(j)}-\mu}{m}\right) = -\frac{1}{m}\mathbb{E}\left(\left(x^{(i)}-\mu\right)^2\right) = -\frac{1}{m}\sigma^2$$

$$\because\ \mathbb{E}\left(\left(x^{(i)}-\mu\right)\left(x^{(j)}-\mu\right)\right) = \begin{cases}\sigma^2 & i=j\\ 0 & i\neq j\end{cases}$$

$$\mathbb{E}\left((\mu-\hat{\mu}_m)^2\right) = \mathbb{E}\left(\left(\sum_{j=1}^{m}\frac{x^{(j)}-\mu}{m}\right)^2\right)$$

$$= \frac{1}{m^2}\left(\sum_{i=1}^{m}\mathbb{E}\left(\left(x^{(i)}-\mu\right)^2\right) + 2\sum_{i\neq j}\mathbb{E}\left(\left(x^{(i)}-\mu\right)\left(x^{(j)}-\mu\right)\right)\right)$$

$$= \frac{1}{m^2}\sum_{i=1}^{m}\mathbb{E}\left(\left(x^{(i)}-\mu\right)^2\right)$$

$$= \frac{m}{m^2}\sigma^2 = \frac{1}{m}\sigma^2$$

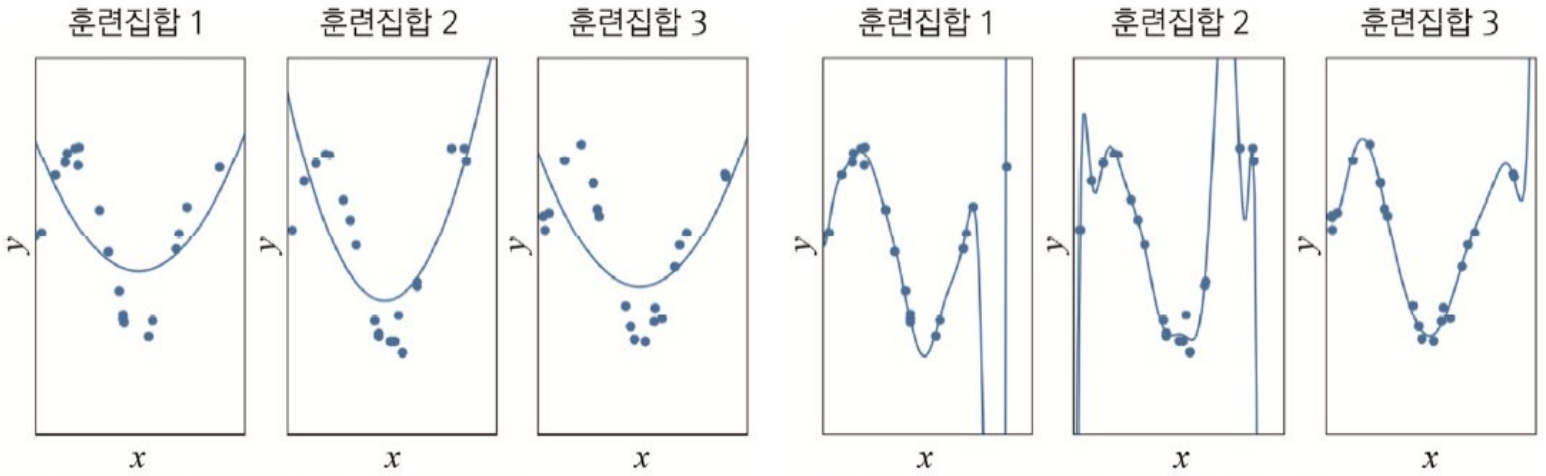
Bias and Variance – cont.

- MSE와 Bias, Variance의 관계

$$
\begin{aligned}
\mathrm{MSE} &= \mathbb{E}\left(\left(\hat{\theta}_m - \theta\right)^2\right) \\
&= \mathbb{E}\left(\left(\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m) + \mathbb{E}(\hat{\theta}_m) - \theta\right)^2\right) \\
&= \mathbb{E}\left(\left(\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m)\right)^2 + 2\left(\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m)\right)\left(\mathbb{E}(\hat{\theta}_m) - \theta\right) + \left(\mathbb{E}(\hat{\theta}_m) - \theta\right)^2\right) \\
&= \mathbb{E}\left(\left(\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m)\right)^2\right) + 2\mathbb{E}\left(\left(\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m)\right)\left(\mathbb{E}(\hat{\theta}_m) - \theta\right)\right) + \mathbb{E}\left(\left(\mathbb{E}(\hat{\theta}_m) - \theta\right)^2\right) \\
&= \mathbb{E}\left(\left(\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m)\right)^2\right) + 2Bias(\hat{\theta}_m)\mathbb{E}\left(\hat{\theta}_m - \mathbb{E}(\hat{\theta}_m)\right) + \mathbb{E}\left(\left(\mathbb{E}(\hat{\theta}_m) - \theta\right)^2\right) \\
&= Var(\hat{\theta}_m) + Bias(\hat{\theta}_m)^2
\end{aligned}
$$

Bias and Variance – cont.

- 훈련집합을 여러 번 수집하여 1차~12차에 적용하는 실험
  - 2차는 매번 큰 오차 → 바이어스가 큼. 하지만 비슷한 모델을 얻음 → 낮은 분산
  - 12차는 매번 작은 오차 → 바이어스가 작음. 하지만 크게 다른 모델을 얻음 → 높은 분산
  - 일반적으로 용량이 작은 모델은 바이어스는 크고 분산은 작음. 복잡한 모델은 바이어스는 작고 분산은 큼
  - 바이어스와 분산은 트레이드오프 관계



(a) 2차 모델(바이어스는 크고, 분산은 작음) (b) 12차 모델(바이어스는 작고, 분산은 큼)

그림 1-15 모델의 바이어스와 분산 특성

Bias and Variance – cont.

- 기계 학습의 목표
  - 낮은 바이어스와 낮은 분산을 가진 예측기 제작이 목표. 즉 왼쪽 아래 상황

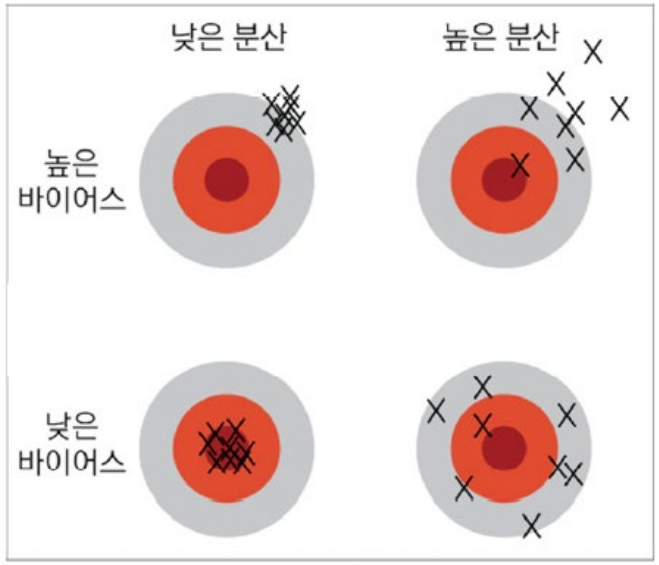


그림 1-16 바이어스와 분산

  - 하지만 바이어스와 분산은 트레이드오프 관계
  - 따라서 바이어스 희생을 최소로 유지하며 분산을 최대로 낮추는 전략 필요