Lecture 4: 모델복잡도와 정규화 (교재 Chapter 3 딥러닝과 정규화)

<기계학습 개론> 2019 강의
서울대학교 컴퓨터공학부
장 병 탁

교재: 장교수의 딥러닝, 홍릉과학출판사, 2017.

Biointelligence Laboratory

School of Computer Science and Engineering

Seoul National University

BI
BIOINTELLIGENCE

목차

4.1 과다학습과 오캄의 면도날 (1/2)

- **모델 복잡도 (model complexity)**
  - 데이터나 문제의 복잡도에 비해 모델 복잡도가 크면 훈련데이터에 대한 정확도 우수
  - 그러나, 높은 복잡도 모델로 과다학습(overfitting)하면 일반화 성능(테스트데이터에 대한 성능) 저하

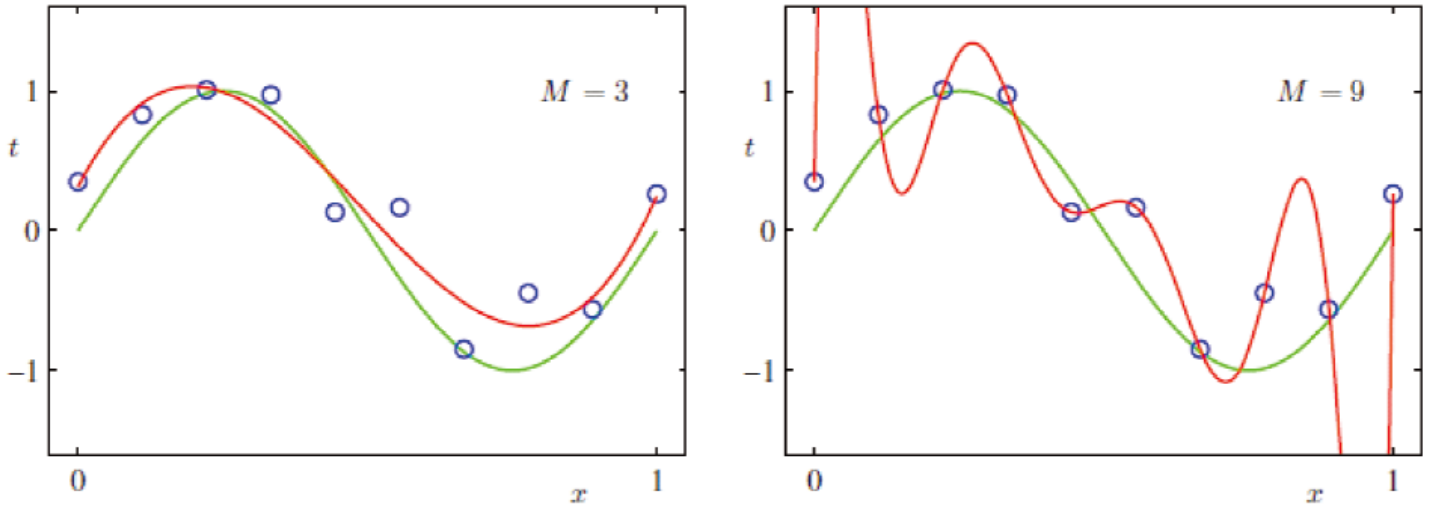

그림 3.9 다항식 신경망의 복잡도에 따른 일반화 성능 비교. 3차 다항식 모델(M=3)에 비해 9차 다항식 모델(M=9)은 과다학습 현상을 보임.

4.1 과다학습과 오캄의 면도날 (2/2)

- **모델 복잡도에 따른 오류**
  - 모델 복잡도가 크면 과다학습
    - 훈련 오류 감소: Training error ↓
    - 일반화 성능 악화: Test error ↑

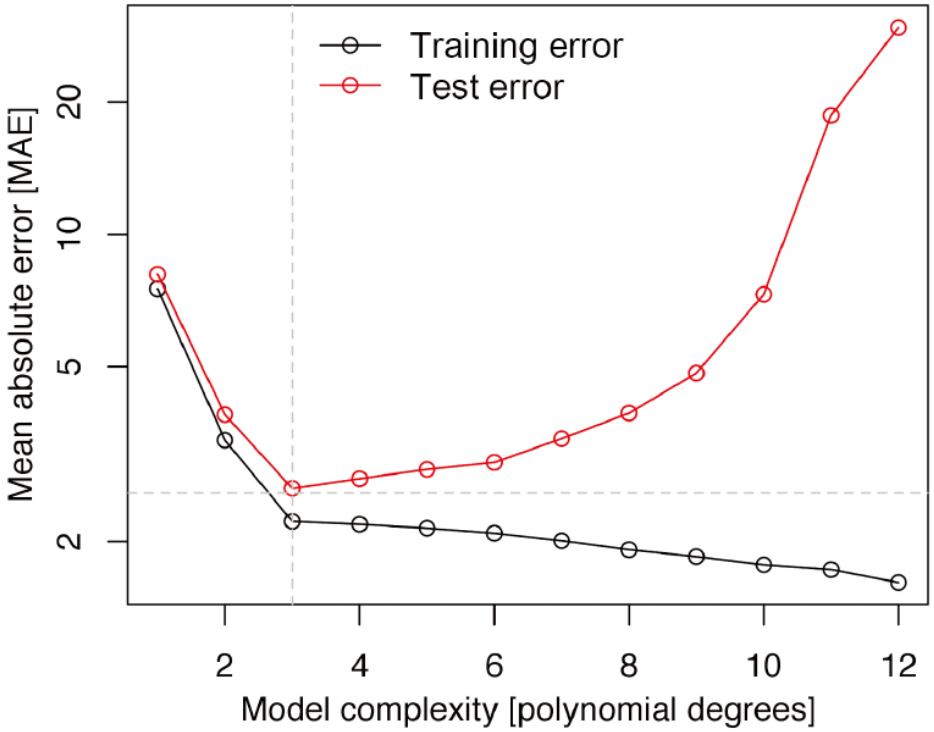

그림 3.10 모델 복잡도와 과다학습 현상의 관계

4.2 모델복잡도와 정규화 (1/2)

- 모델 선택 문제 (Model Selection Problem)
  - 주어진 데이터나 문제에 대해서 가장 최적의 모델 복잡도가 무엇일까?
  - 새로운 데이터에 대해서도 올바른 답을 줄 수 있는 일반화 (generalization) 능력

- 오캄의 면도날 원리 (Occam's razor)
  - 불필요한 복잡도는 면도날로 제거
  - 측정한 데이터에 대한 현상을 똑같이 잘 설명하는 두 개의 모델, 즉, 간단한 모델과 복잡한 모델이 있다면 이 중에서 간단한 모델을 선호하라는 원리

4.2 모델복잡도와 정규화 (2/2)

- **모델 복잡도 측정**
  - 매개변수의 개수에 비례
  - 매개변수값의 절대값의 크기에 비례

$$C(\mathbf{w}) = \|\mathbf{w}\|^2 = \sum_{k=1}^{K} w_k^2$$

- **모델 복잡도 탐색**
  - 가중치 감소법 (weight decay): 가중치 값을 조절할 때마다 일정 비율을 감소
  - 학습의 반복 횟수를 너무 크게 하지 않음
    - 반복횟수가 클수록 가중치 값들의 절대값이 커지는 경향이 있음
  - 목적함수 변경

$$F(\mathbf{w}|D) = \beta E(D|\mathbf{w}) + \alpha C(\mathbf{w})$$

4.3 구조위험최소화(SRM)와 MDL (1/3)

- **모델 선택 문제** 즉 최적의 모델 복잡도를 갖는 학습 모델을 찾는 문제를 체계적으로 접근하는 한 가지 방법은 정규화 이론을 이용

- **구조위험최소화법 (Structural Risk Minimization, SRM)**

  - 목적함수 정규화 (regularization)

$$R_{reg}[f,\gamma] = R_{train}[f] + \gamma \|\mathbf{w}\|^2$$

  - 모델의 복잡도가 증가함에 따라 오류가 증가
  - 복잡도와 훈련오류를 모두 고려한 최적의 모델을 찾을 수 있음

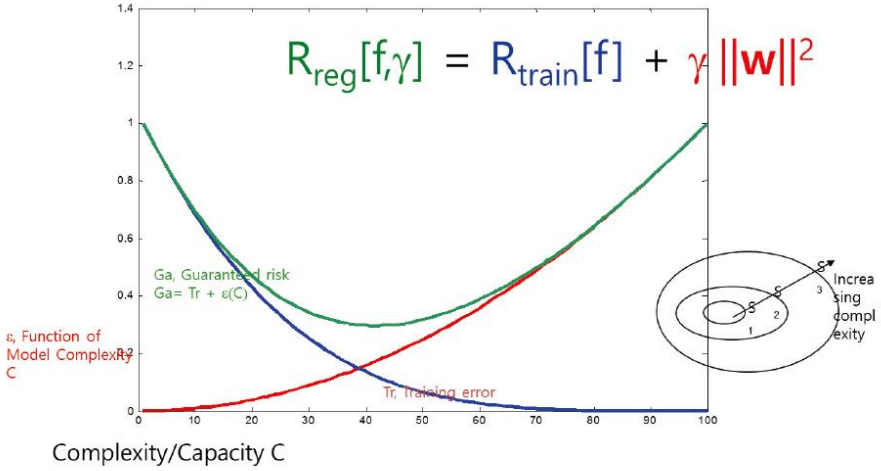

그림 3.11 정규화된 목적함수와 오류함수 및 모델복잡도의 관계

4.3 구조위험최소화(SRM)와 MDL (2/3)

- **최소묘사길이 (Minimum Description Length, MDL)**
  - 통신망을 통해서 메시지를 부호하고 압축해서 전송하고 수신단에서는 복호화해서 원래 메시지를 복원하는 방법
  - 목적함수

$$L(A|D) = L(D|A) + L(A)$$

  - $L(D \mid A)$: 모델 A에 대해서 데이터 D를 부호화하는 최소 코드의 길이
  - $L(A)$: 모델 자체를 부호화하는 최소 코드의 길이
  - 정규화 목적함수와 비교하여 동일하게 오류와 복잡도를 같이 최소화하는 방법

$$R_{train}[f] \cong L(D|A)$$

$$\|\mathbf{w}\|^2 \cong L(A)$$

$$R_{reg}[f,\gamma] = R_{train}[f] + \gamma \|\mathbf{w}\|^2$$

4.3 구조위험최소화(SRM)와 MDL (3/3)

❑ 정보이론 및 확률과의 관계

- 확률이 높은 사건 보다는 확률이 낮은 사건을 관측함으로써 얻는 정보가 더 큼

$$I(x) = \frac{1}{\log P(x)} = -\log P(x)$$

- 코드의 길이는 그 부호가 나타날 확률에 반비례

$$L(D|A) = -\log_2 P(D|A)$$

$$L(A) = -\log_2 P(A)$$

$$L(A|D) = -\log_2 P(D|A) \; - \; \log_2 P(A)$$

4.4 베이즈규칙과 MAP (1/2)

최대사후확률법 (maximum a posteriori)

- 베이즈 규칙 (Bayes' rule)

$$P(A|D) = \frac{P(D|A)P(A)}{P(D)}$$

$P(A)$: 사전확률 (prior probability)
$P(D|A)$: 우도 (likelihood)
$P(D)$: 주변우도 (marginal likelihood)
$P(A|D)$: 사후확률 (posterior probability)

- $P(D)$는 실제로 계산 불가능

$$P(D) = \int_A P(D|A)P(A)dA \triangleq \sum_A P(D|A)P(A)$$

4.4 베이즈규칙과 MAP (2/2)

- 분포를 추정하는 대신에 최대값을 추정 (maximum a posteriori estimation)

$$A^* = \underset{A}{\operatorname{argmax}}\ P(A|D) = \underset{A}{\operatorname{argmax}}\ \{P(D|A)P(A)\}$$

$$= \underset{A}{\operatorname{argmax}}\ \{\log P(D|A) + \log P(A)\}$$

$$= \underset{A}{\operatorname{argmax}}\ \{-\log P(D|A) + \log P(A)\}$$

- MAP은 MDL에 의해서 학습하는 것과 동등함

$$A^* = \underset{A}{\operatorname{argmin}}\ L(A|D) = \underset{A}{\operatorname{argmin}}\ \{-\log_2 P(D|A) - \log_2 P(A)\}$$

4.5 편향분산 분석 (1/3)

감독학습의 회귀분석 문제

- 훈련데이터

$$D_N = \{(\boldsymbol{x}^{(d)}, y^{(d)})\}_{d=1}^{N}$$

- 출력

$$y^{(d)} = f(\boldsymbol{x}^{(d)})$$

$$(\boldsymbol{x}^{(d)}, y^{(x)}) \sim P_f(\boldsymbol{x}, y)$$

- 손실함수 (loss function)

$$L(y, f_A(\boldsymbol{x})) = (y - f_A(\boldsymbol{x}))^2$$

- 최적화

$$E_N(A) = \frac{1}{N}\left\{\sum_{d=1}^{N} L(y^{(d)}, f_A(\boldsymbol{x}^{(d)}))\right\}$$

4.5 편향분산 분석 (2/3)

- 훈련데이터 외에 관측되지 않은 데이터 손실 최소화

$$R(A) = \int_{X,Y} P_f(\boldsymbol{x}, y) L\big(y, f_A(\boldsymbol{x})\big) d\boldsymbol{x} dy$$

- 그러나, $P_f(\boldsymbol{x}, y) = P_f(y|\boldsymbol{x}) P_f(\boldsymbol{x})$ 는 알려져 있지 않음
- 알고있는 정보는 훈련데이터 D

$$\mathrm{E}[(y - f_A(\boldsymbol{x}; D))^2 | \boldsymbol{x}, D]$$

$$= \mathrm{E}\big[\big(y - \mathrm{E}(y|\boldsymbol{x})\big)^2 \big| \boldsymbol{x}, D\big] + \big(\mathrm{E}(y|\boldsymbol{x}) - f_A(\boldsymbol{x}; D)\big)^2$$

- f 에 대한 평균제곱오차

$$\mathrm{E}_D\left[\big(\mathrm{E}(y|\boldsymbol{x}) - f_A(\boldsymbol{x}; D)\big)^2\right]$$

$$= \big(\mathrm{E}_D[f_A(\boldsymbol{x}; D) - \mathrm{E}(y|\boldsymbol{x})]\big)^2 + \mathrm{E}_D[(f_A(\boldsymbol{x}; D) - \mathrm{E}_D[f_A(\boldsymbol{x}; D)])^2]$$

4.5 편향분산 분석 (3/3)

- 두 종류의 오차
  - 편향(bias) 오차
  - 분산(variance) 오차
- 모델 선택
  - 단순한 모델:
    편향 오차 ↑
    분산 오차 ↓
  - 복잡한 모델:
    편향 오차 ↓
    분산 오차 ↑

$$\underbrace{E_D[f(x)-y]^2}_{\text{Expected value of the loss over datasets D of the same size}} = \underbrace{[E_D f(x)-y]^2}_{\text{Bias}^2} + \underbrace{E_D[f(x)-E_D f(x)]^2}_{\text{Variance}}$$

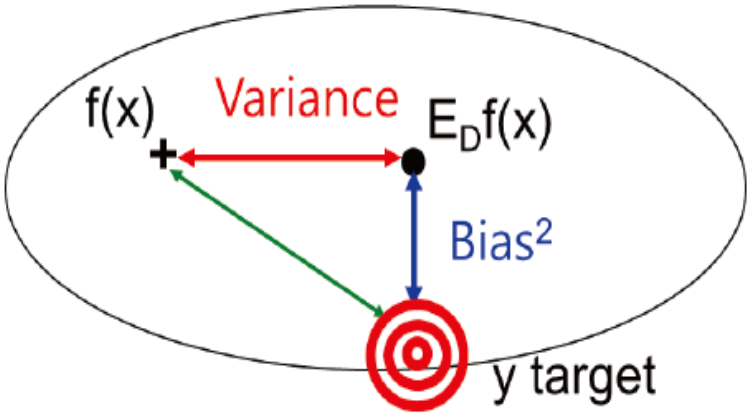

그림 3.12 편향 오차와 분산 오차의 분석

요약

- 모델 복잡도와 정규화 이론
  - 오캄의 면도날 원리
  - 모델 복잡도에 따른 모델 선택 문제와 정규화 방법
  - SRM
  - MDL
  - MAP
- 편향분산 분석
  - 단순한 모델: 편향 오차 ↑, 분산 오차 ↓
  - 복잡한 모델: 편향 오차 ↓, 분산 오차 ↑