1장. 기계학습 소개

실전코딩

본 강의자료는 한빛아카데미에서 제공하는 강의자료를 바탕으로 작성되었음

각 절에서 다루는 내용

- 1.1절: 기계 학습의 정의와 개념, 인공지능을 구현하는 도구로서의 역할을 설명한다.
- 1.2절: 특징 공간과 공간 변환을 소개한다.
- 1.3절: 데이터의 중요성과 희소성을 강조한다.
- 1.4절: 선형 회귀를 이용하여 기계 학습을 직관적으로 설명한다.
- 1.5절: 모델 선택의 중요성과 방법, 과소적합과 과잉적합을 설명한다.
- 1.6절: 현대 기계 학습에서 매우 중요한 규제 기법으로 데이터 확대와 가중치 감쇠를 간략히 기술한다.
- 1.7절: 기계 학습의 유형으로 지도 학습, 비지도 학습, 강화 학습, 준지도 학습을 소개한다.
- 1.8절: 기계 학습의 간략한 역사와 인공지능의 사회적 의미를 살펴본다.

1.5 모델 선택

- 1.5.1 과소적합과 과잉적합
- 1.5.2 바이어스와 분산
- 1.5.3 검증집합과 교차검증을 이용한 모델 선택 알고리즘
- 1.5.4 모델 선택의 한계와 현실적인 해결책

1.5.1 과소적합과 과잉적합

- [그림 1.13]의 1차 모델은 과소적합 (under-fitting)
  - 모델의 '용량이 작아' 오차가 클 수밖에 없는 현상
- 비선형 모델을 사용하는 대안
  - [그림 1-13]의 2차, 3차, 4차, 12차는 다항식 곡선을 선택한 예
  - 1차(선형)에 비해 오차가 크게 감소함

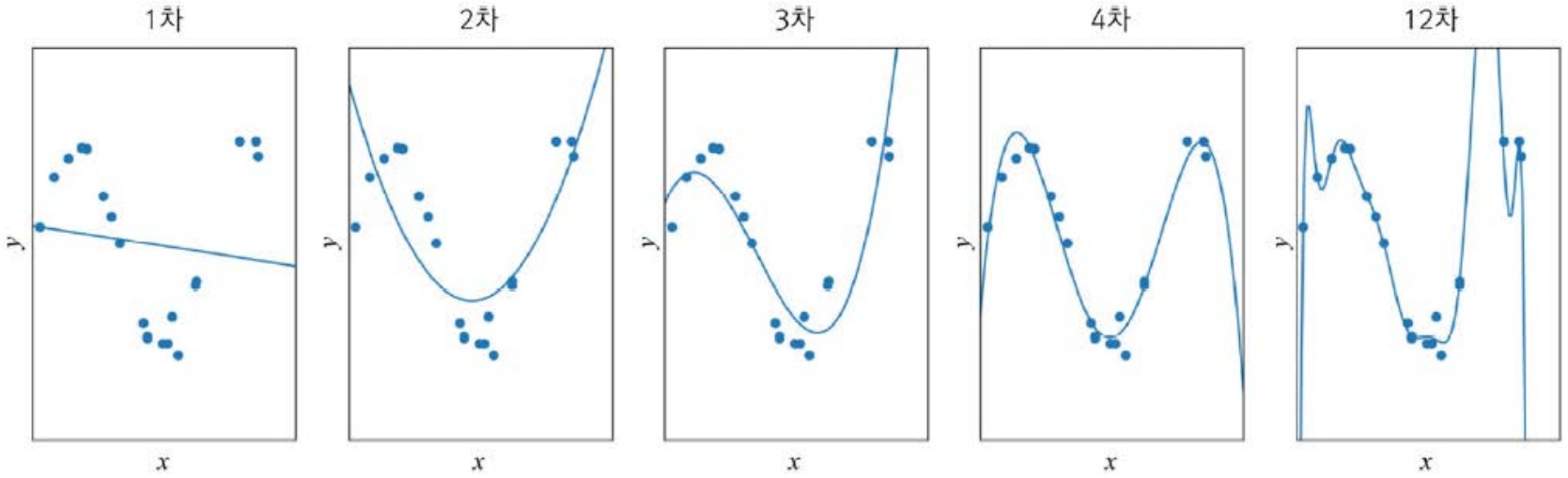


그림 1-13 과소적합과 과잉적합 현상

1.5.1 과소적합과 과잉적합

과잉적합 (over-fitting)

- 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
- 적절한 용량의 모델을 선택하는 모델 선택 작업이 필요함

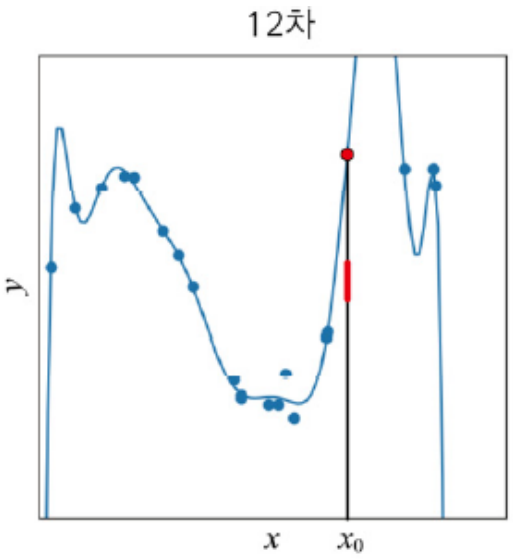


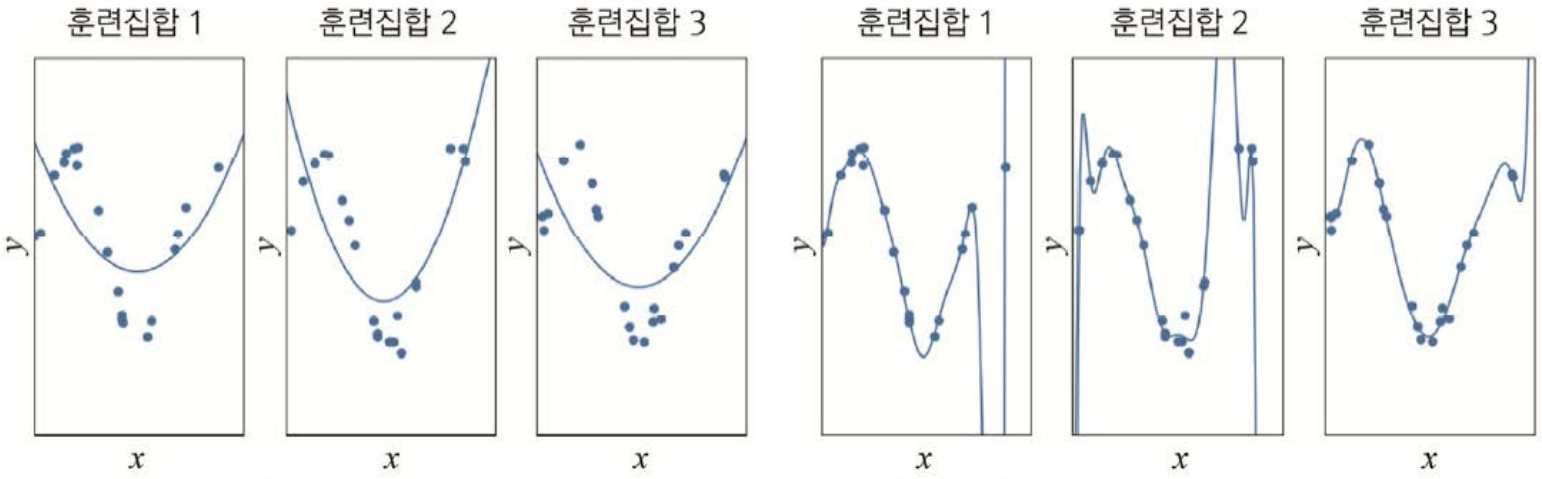
그림 1-14 과잉적합되었을 때 부정확한 예측 현상

1.5.2 바이어스와 분산

- 1차~12차 다항식 모델의 비교 관찰
  - 1~2차는 훈련집합과 테스트집합 모두 낮은 성능
  - 12차는 훈련집합에 높은 성능을 보이나 테스트집합에서는 낮은 성능 → 낮은 일반화 능력
  - 3~4차는 훈련집합에 대해 12차보다 낮겠지만 테스트집합에는 높은 성능 → 높은 일반화 능력

1.5.2 바이어스와 분산

- 훈련집합을 여러 번 수집하여 1차~12차에 적용하는 실험
  - 2차는 매번 큰 오차 → 바이어스가 큼*(정답에서 멀다)*. 하지만 비슷한 모델을 얻음 → 낮은 분산 *(예측값들의 분포가 비슷)*
  - 12차는 매번 작은 오차 → 바이어스가 작음*(정답에서 가깝다)*. 하지만 크게 다른 모델을 얻음 → 높은 분산 *(예측값들의 분포가 다르다)*
  - 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

1.5.2 바이어스와 분산

기계 학습의 목표

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

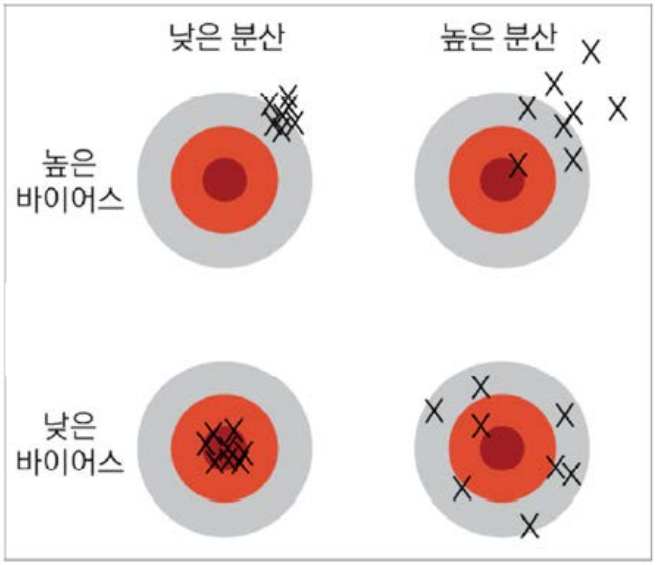


그림 1-16 바이어스와 분산

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

1.5.3 검증집합과 교차검증을 이용한 모델 선택 알고리즘

■ 검증집합을 이용한 모델 선택

- 훈련집합(training set: 70%)과 테스트집합(test set: 15%)과 다른 별도의 검증집합(validation set: 15%)을 가진 상황

알고리즘 1-2 검증집합을 이용한 모델 선택

입력 : 모델집합 Ω, 훈련집합, 검증집합, 테스트집합
출력 : 최적 모델과 성능

```
1  for (Ω에 있는 각각의 모델)
2      모델을 훈련집합으로 학습시킨다.
3      검증집합으로 학습된 모델의 성능을 측정한다.  // 검증 성능 측정
4  가장 높은 성능을 보인 모델을 선택한다.
5  테스트집합으로 선택된 모델의 성능을 측정한다.
```

1.5.3 검증집합과 교차검증을 이용한 모델 선택 알고리즘

교차검증cross validation

- **비용 문제로 별도의 검증집합이 없는 상황에 유용한 모델 선택 기법**
- 훈련집합을 등분하여, 학습과 평가 과정을 여러 번 반복한 후 평균 사용

알고리즘 1-3 교차검증에 의한 모델 선택

입력 : 모델집합 Ω, 훈련집합, 테스트집합, 그룹 개수 k
출력 : 최적 모델과 성능

```
1  훈련집합을 k개의 그룹으로 등분한다.
2  for (Ω에 있는 각각의 모델)
3        for (i=1 to k)
4              i번째 그룹을 제외한 k-1개 그룹으로 모델을 학습시킨다.
5              학습된 모델의 성능을 i번째 그룹으로 측정한다.
6        k개 성능을 평균하여 해당 모델의 성능으로 취한다.
7  가장 높은 성능을 보인 모델을 선택한다.
8  테스트집합으로 선택된 모델의 성능을 측정한다.
```

1.5.3 검증집합과 교차검증을 이용한 모델 선택 알고리즘

부트스트랩boot strap

난수를 이용한 샘플링 반복

알고리즘 1-4 부트스트랩을 이용한 모델 선택

입력 : 모델집합 Ω, 훈련집합, 테스트집합, 샘플링 비율 $p(0<p\leq 1)$, 반복횟수 T
출력 : 최적 모델과 성능

```
1  for (Ω에 있는 각각의 모델)
2      for (i=1 to T)
3          훈련집합 𝕏에서 pn개 샘플을 뽑아 새로운 훈련집합 𝕏′를 구성한다. 이때 대치를 허용한다.
4          𝕏′로 모델을 학습시킨다.
5          𝕏 − 𝕏′를 이용하여 학습된 모델의 성능을 측정한다.
6      T개 성능을 평균하여 해당 모델의 성능으로 취한다.
7  가장 높은 성능을 보인 모델을 선택한다.
8  테스트집합으로 선택된 모델의 성능을 측정한다.
```

1.5.4 모델 선택의 한계와 현실적인 해결책

- [알고리즘 1-2, 1-3, 1-4]에서 모델 집합 Ω (오메가)
  - [그림 1-13]에서는 서로 다른 차수의 다항식이 Ω인 셈
  - 현실에서는 아주 다양
    - 신경망(3,4,8장), 강화 학습(9장), 확률 그래피컬 모델(10장), SVM(11장), 트리 분류기 (12장) 등이 선택 대상
    - 신경망을 채택하더라도 MLP(3장), 깊은 MLP(4장), CNN(4장) 등 아주 많음

- **현실에서는 경험으로 큰 틀 선택한 후**
  - **모델 선택 알고리즘으로 세부 모델 선택하는 전략 사용**
  - **예) CNN을 사용하기로 정한 후, 은닉층 개수, 활성함수, 모멘툼 계수 등을 정하는데 모델 선택 알고리즘을 적용함**

1.5.4 모델 선택의 한계와 현실적인 해결책

- 이런 경험적인 접근방법에 대한『Deep Learning』책의 비유

"To some extent, we are always trying to fit a square peg(the data generating process) into a round hole(our model family). 어느 정도 우리가 하는 일은 항상 둥근 홈(우리가 선택한 모델)에 네모 막대기(데이터 생성 과정)를 끼워 넣는 것이라고 말할 수 있다[Goodfellow2016(222쪽)]."

- **현대 기계 학습의 전략**
  - **용량이 충분히 큰 모델을 선택 한 후, 선택한 모델이 정상을 벗어나지 않도록 여러 가지 규제regularization 기법을 적용함**
  - 예) [그림 1-13]의 경우 12차 다항식을 선택한 후 적절히 규제를 적용

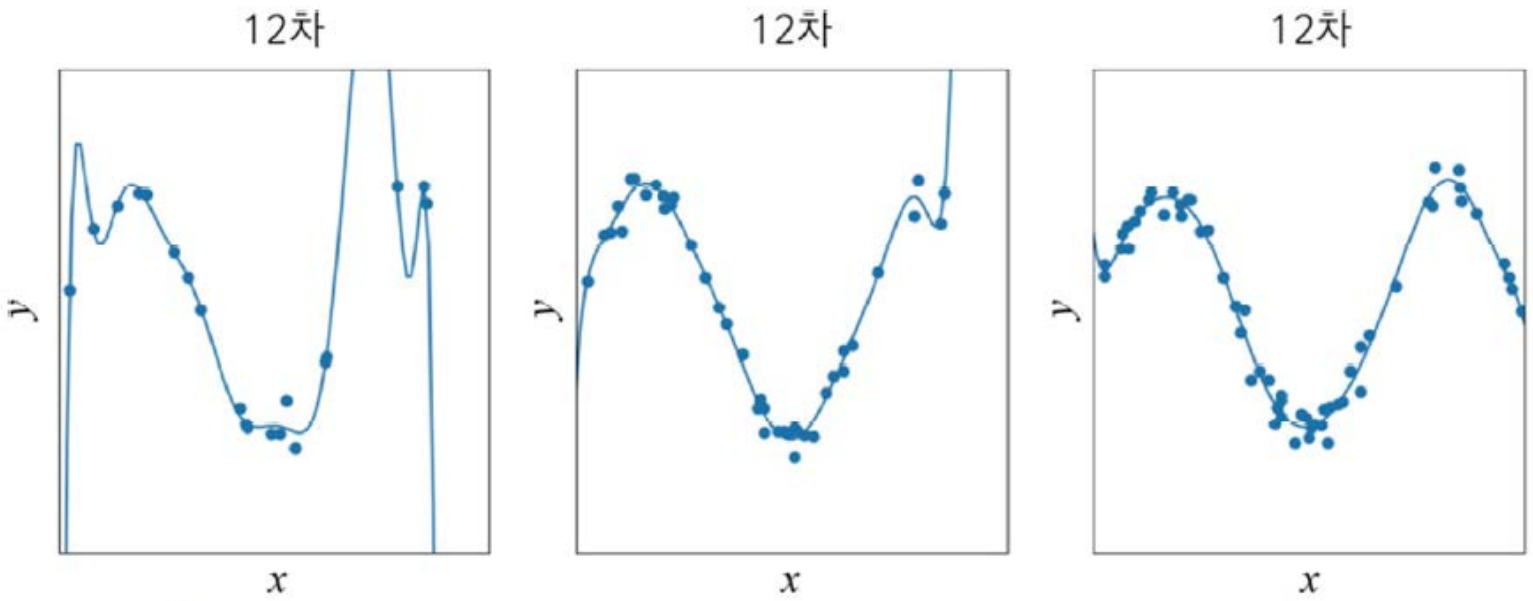
1.6 규제

- 1.6.1 데이터 확대
- 1.6.2 가중치 감쇠

- 규제를 중요하게 다룬 책 [Goodfellow2016(7장)] [Haykin2009(7장)]
- 이 책은 5.3~5.4절에서 자세히 다룸
  - 가중치 벌칙, 조기 멈춤, 데이터 확대, 드롭아웃, 앙상블 등

1.6.1 데이터 확대

- 데이터를 더 많이 수집하면 일반화 능력이 향상됨 (하지만 현실에선 어려움)



(a) 훈련집합의 크기 = 20 (b) 훈련집합의 크기 = 40 (c) 훈련집합의 크기 = 60

그림 1-17 데이터를 확대하여 일반화 능력을 향상함

1.6.1 데이터 확대

- 데이터 수집은 많은 비용이 듦
  - 그라운드 트루스를 사람이 일일이 레이블링해야 함
- 인위적으로 데이터 확대
  - 훈련집합에 있는 샘플을 변형함
  - 약간 회전 또는 와핑 (부류 소속이 변하지 않게 주의)

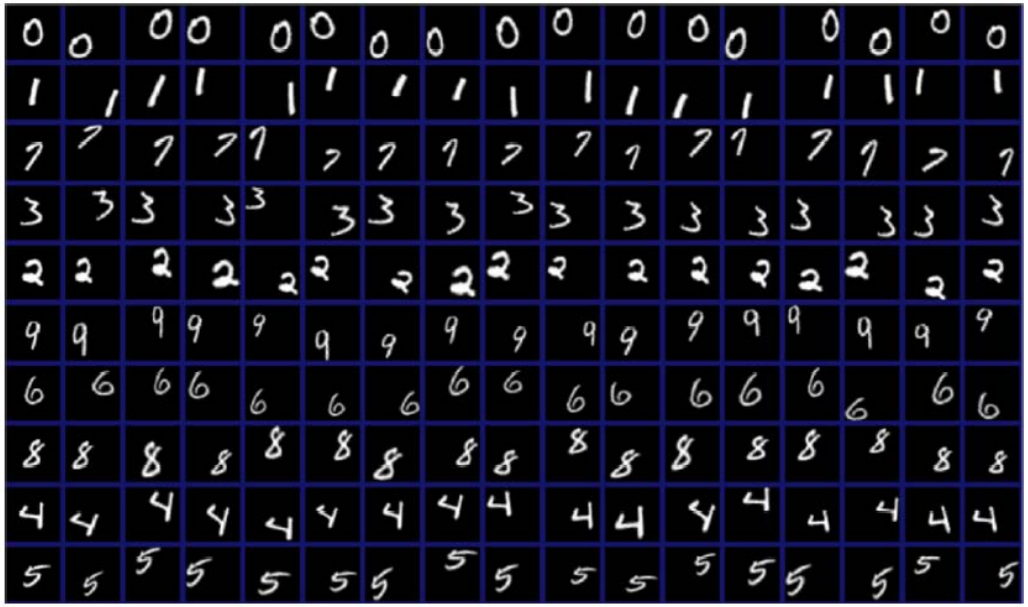


그림 5-24 필기 숫자 데이터의 다양한 변형[8]

1.6.2 가중치 감쇠

가중치를 작게 조절하는 기법

- [그림 1-18(a)]의 12차 곡선은 가중치가 매우 큼

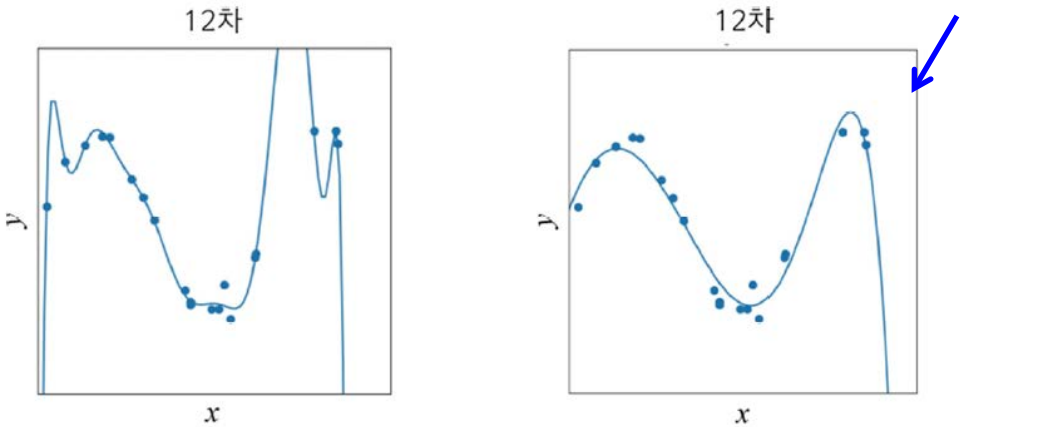
$$y = 1005.7x^{12} - 27774.4x^{11} + \cdots - 22852612.5x^{1} - 12.8$$

- **가중치 감쇠는 개선된 목적함수를 이용하여 가중치를 작게 조절하는 규제 기법**
  - 식 (1.11)의 두 번째 항은 규제 항(regularization term)으로서 가중치 크기를 작게 유지해 줌

$$J(\Theta) = \frac{1}{n}\sum_{i=1}^{n}(f_\Theta(\mathbf{x}_i) - y_i)^2 + \lambda\|\Theta\|_2^2 \qquad (1.11)$$

$$\begin{aligned}\boldsymbol{\Theta} &= \boldsymbol{\Theta} - \rho\nabla J_{regularized}(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y}) \\ &= \boldsymbol{\Theta} - \rho(\nabla J(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y}) + 2\lambda\boldsymbol{\Theta}) \\ &= (1 - 2\rho\lambda)\boldsymbol{\Theta} - \rho\nabla J(\boldsymbol{\Theta}; \mathbb{X}, \mathbb{Y})\end{aligned}$$

$$y = 10.779x^{12} - 42.732x^{11} + \cdots - 2.379x^{1} + 0.119$$



(a) 가중치 감쇠 적용 안 함[식 (1.8)의 목적함수] (b) 가중치 감쇠 적용함[식 (1.11)의 목적함수]

그림 1-18 가중치 감쇠에 의한 규제 효과

$$L_p = \left(\sum_{i}^{n} |x_i|^p\right)^{\frac{1}{p}}$$

Norm: 벡터의 길이 혹은 크기를 측정하는 방법
→ 각 요소별로 절대값을 p번 곱한 값의 합을 p제곱근한 값

L2-norm: $\|w\|_2^2 = (w_0{}^2 + w_1{}^2 + \cdots w_M{}^2)^{1/2}$
→ Weight decay 기법임

1.7 기계 학습 유형

- 1.7.1 지도 방식에 따른 유형
- 1.7.2 다양한 기준에 따른 유형

1.7.1 지도 방식에 따른 유형

지도 학습

- 특징 벡터 $\mathbb{X}$와 목표값 $\mathbb{Y}$가 모두 주어진 상황
- 회귀와 분류 문제로 구분

비지도 학습

- 특징 벡터 $\mathbb{X}$는 주어지는데 목푯값 $\mathbb{Y}$가 주어지지 않는 상황
- 군집화 과업 (고객 성향에 따른 맞춤 홍보 응용 등)
- 밀도 추정, 특징 공간 변환 과업
- 6장의 주제

1.7.1 지도 방식에 따른 유형

강화 학습

- 목표값이 주어지는데, 지도 학습과 다른 형태임
- 예) 바둑
  - 수를 두는 행위가 샘플인데, 게임이 끝나면 목푯값 하나가 부여됨
    - 이기면 1, 패하면 -1을 부여
  - 게임을 구성한 샘플들 각각에 목푯값을 나누어 주어야 함
- 9장의 주제

준지도 학습

- 일부는 $\mathbb{X}$와 $\mathbb{Y}$를 모두 가지지만, 나머지는 $\mathbb{X}$만 가진 상황
- 인터넷 덕분으로 $\mathbb{X}$의 수집은 쉽지만, $\mathbb{Y}$는 수작업이 필요하여 최근 중요성 부각
- 7장의 주제

1.7.2 다양한 기준에 따른 유형

- 오프라인 학습과 온라인 학습
  - 이 책은 오프라인 학습을 다룸
  - 온라인 학습은 인터넷 등에서 추가로 발생하는 샘플을 가지고 점증적 학습

- 결정론적 학습과 스토캐스틱 학습
  - 결정론적에서는 같은 데이터를 가지고 다시 학습하면 같은 예측기가 만들어짐
  - 스토캐스틱 학습은 학습 과정에서 난수를 사용하므로 같은 데이터로 다시 학습하면 다른 예측기가 만들어짐. 보통 예측 과정도 난수 사용
  - 10.4절의 RBM과 DBN이 스토캐스틱 학습

- **분별 모델(discriminative model)과 생성 모델(generative model) → 다음주에**
  - 분별 모델은 부류 예측에만 관심. 즉 $P(y|\mathbf{x})$의 추정에 관심
  - 생성 모델은 $P(\mathbf{x})$ 또는 $P(\mathbf{x}|y)$를 추정함
    - 따라서 새로운 샘플을 '생성'할 수 있음
    - 4.5절의 GAN, 10.4절의 RBM은 생성 모델
    - 8.5절의 순환신경망(RNN)을 생성 모델로 활용하는 응용 예제

1.8 기계 학습의 과거와 현재, 미래

- 1.8.1 인공지능과 기계 학습의 간략한 역사
- 1.8.2 기술 추세
- 1.8.3 사회적 전망

1.8.1 인공지능과 기계 학습의 간략한 역사

- 베비지의 제자인 에이더 여사의 통찰력 (19세기 중반)
  - "... 해석엔진은 숫자 이외의 것도 처리할 수 있을 것이다. ... 예를 들어 화음과 음조를 해석엔진의 표기에 맞출 수 있다면, 해석엔진은 꽤 복잡한 곡을 작곡할 수도 있다." [Ada1843]
  - 200여 년이 지난 지금,
    - 흘려 쓴 필기 숫자를 0.23% 오류로 인식
    - 알파고는 이세돌을 이김
    - 자연영상에 대해 다섯 단어 가량의 문장으로 묘사함

1.8.1 인공지능과 기계 학습의 간략한 역사

1843	에이더 "… 해석엔진은 꽤 복잡한 곡을 작곡할 수도 있다."라는 논문 발표[Ada1843]
1950	인공지능 여부를 판별하는 튜링 테스트[Turing1950]
1956	최초의 인공지능 학술대회인 다트머스 콘퍼런스 개최. '인공지능' 용어 탄생[McCarthy1955]
1958	로젠블렛이 퍼셉트론 제안[Rosenblatt1958]
	인공지능 언어 Lisp 탄생
1959	사무엘이 기계 학습을 이용한 체커 게임 프로그램 개발[Samuel1959]
1969	민스키가 퍼셉트론의 과대포장 지적. 신경망 내리막길 시작[Minsky1969]
	제1회 IJCAI International Joint Conference on Artificial Intelligence 개최
1972	인공지능 언어 Prolog 탄생
1973	Lighthill 보고서로 인해 인공지능 내리막길, 인공지능 겨울AI winter 시작
1974	웨어보스가 오류 역전파 알고리즘을 기계 학습에 도입[Werbos1974]
1975경	의료진단 전문가 시스템 Mycin – 인공지능에 대한 관심 부활
1979	「IEEE Transactions on Pattern Analysis and Machine Intelligence」 저널 발간
1980	제1회 ICML International Conference on Machine Learning 개최
	후쿠시마가 NeoCognitron 제안[Fukushima1980]
1986	「Machine Learning」 저널 발간
	『Parallel Distributed Processing』 출간
	다층 퍼셉트론으로 신경망 부활

단층 퍼셉트론
→ 선형적인 분리에 대한 학습만 가능 (XOR 문제 해결 못함)

[69년, Marvin Minsky (MIT AI Lab) XOR풀려면 MLP 밖에 없는 아무도 못풀거라고 했었음. 즉 학습시킬 수 없을 것이다] → AI의 암흑기

다층 퍼셉트론 사용하면
→ 비선형적인 분리도 가능하고, 또한 학습도 가능하다는 것이 Backpropagation 알고리즘을 통해 증명되었음(70년대). 그러나 관심이 없었다가 86년에 Hinton에 의해서 다시 발표됨

1.8.1 인공지능과 기계 학습의 간략한 역사

1987	Lisp 머신의 시장 붕괴로 제2의 인공지능 겨울
	UCI 리포지토리 서비스 시작
	NIPSNeural Information Processing Systems 콘퍼런스 시작
1989	「Neural Computation」 저널 발간
1993	R 언어 탄생
1997	IBM 딥블루가 세계 체스 챔피언인 카스파로프 이김
	LSTMLong short-term memory 개발됨
1998경	SVM이 MNIST 인식 성능에서 신경망 추월
1998	르쿤이 CNN의 실용적인 학습 알고리즘 제안[LeCun1998]
	『Neural Networks: Tricks of the Trade』 출간
1999	NVIDIA 사에서 GPU 공개
2000	「Journal of Machine Learning Research」 저널 발간
	OpenCV 최초 공개
2004	제1회 그랜드 챌린지(자율 주행)
2006	층별학습 탄생[Hinton2006a]
2007경	딥러닝이 MNIST 인식 성능에서 SVM 추월
2007	GPU 프로그래밍 라이브러리인 CUDA 공개

1.8.1 인공지능과 기계 학습의 간략한 역사

	어번 챌린지(도심 자율 주행)
	Scikit-learn 라이브러리 최초 공개
2009	Theano 서비스 시작
2010	ImageNet 탄생
	제1회 ILSVRC 대회
2011	IBM 왓슨이 제퍼디 우승자 꺾음
2012	MNIST에 대해 0.23% 오류율 달성
	AlexNet 발표 (3회 ILSVRC 우승)
2013	제1회 ICLR International Conference on Learning Representations 개최
2014	Caffe 서비스 시작
2015	TensorFlow 서비스 시작
	OpenAI 창립
2016	알파고와 이세돌의 바둑 대회에서 알파고 승리[Silver2016]
	『Deep Learning』 출간
2017	알파고 제로[Silver2017]

1.8.2 기술 추세

- 리뷰 논문
  - [LeCun2015, Jordan2015, Jones2014]
- 기계 학습은 인공지능 실현에 핵심 기술
- 기계 학습 알고리즘과 응용의 다양화
- 서로 다른 알고리즘과 응용의 융합
- **딥러닝이 기계 학습의 주류**
- **표현 학습이 중요해짐**

1.8.3 사회적 전망

미래의 직업 변화

- 시의적절하고 심사숙고 해야 할 객관적 담론
- 프레이는 702개 직업의 사라질 위기를 확률로 계산 [Frey2017]
- 텔레마케터 99% …… …… …… ….. ….. ….. 오락 치료사 0.28%

기계가 사람을 지배할지 모른다는 두려움

- 알파고 이후 매스컴을 통해 여과 없이 전파. 쓸데없는 과장에 불과
- 냇가에 다리 놓는 일과 목포-제주에 대교를 놓는 일은 규모만 다를 뿐 본질적으로 같은 일
- 오목 프로그램이나 바둑 프로그램은 규모만 다를 뿐 본질적으로 같은 일. 오목은 간단한 규칙으로 구현 가능하나 바둑은 미분을 사용한 복잡한 기계 학습 알고리즘 사용
- 현재 기계 학습은 온통 수학과 컴퓨터 알고리즘일 뿐