심층학습

윤예준

저자: 이안 굿펠로, 요슈아 벤지오, 에런 쿠빌

목차

- 5장 기계 학습의 기초
  - 학습 알고리즘
  - 수용력, 과대적합, 과소적합

5.1 학습 알고리즘

- 학습(learning)이란?
  - “한 컴퓨터 프로그램이 어떤 과제류(class of tasks) T에 속하는 과제들을 수행하며 그 수행의 성과를 측정한 측도가 P라고 할 때, 만일 어떤 경험 E 때문에 T의 어떤 과제에 대한 성과 측도 P가 개선되었다면, 그 컴퓨터 프로그램은 경험 E로부터 학습한다고 말할 수 있다.” [Mitchell, 1997]

5.1.1 과제 T

- 기계 학습의 과제는 일반적으로 기계 학습 시스템이 견본(example)을 처리하는 방식을 서술하는 형태로 정의된다.
  - 견본이란 기계 학습 시스템의 처리 대상인 어떤 물체나 사건으로부터 정량적으로 측정한 특징(feature)들의 집합이다.

- 분류(classification)
  - 주어진 입력이 k개의 범주 중 어떤 범주에 속하는지 판단하는 과제
  - 이 과제를 풀기 위한 학습 알고리즘은 함수 $f: R^n \rightarrow \{1, \ldots, k\}$를 산출해야한다.

- 결측 입력이 있는 자료의 분류(classification with missing input)
  - 해결 방법 중 하나
    - 모든 관련 변수에 관한 확률분포를 학습하고 결측값들을 주변화하기

5.1.1 과제 T

- 회귀(regression)
  - 주어진 입력에 기초해서 하나의 수치를 예측
  - 학습 알고리즘이 배워야 할 것은 하나의 함수 $f: \mathbb{R}^n \rightarrow \mathbb{R}$
  - 분류 과제와 비슷하나, 출력의 형식이 다름
  - Ex) 보험 가입자가 받을 수 있는 보험금 예측
- 전사(transcription)
  - 비교적 구조적이지 않은 형태로 표현된 어떤 자료를 입력받아서 그 자료에 있는 정보를 이산적인 텍스트 형식으로 출력
  - Ex) OCR(광 문자 인식) 프로그램: 텍스트 이미지가 담긴 사진을 관측해서 그 텍스트에 해당하는 문자열 출력
- 기계 번역(machine translation)
  - 입력은 어떤 언어의 기호들로 이루어진 문자열일 때, 이러한 입력을 다른 언어의 기호들로 이루어진 문자열로 변환

5.1.1 과제 T

- 구조적 출력(structured output)
  - 출력이 하나의 벡터이고 벡터의 성분들 사이에 중요한 관계가 존재하는 형태의 모든 과제
  - Ex) 구문 분석(parsing): 주어진 자연어 문장의 구문 구조를 서술하는 트리를 형성하는 과정을 말한다. 이때 그 트리의 각 노드에 해당 문장 요소의 품사를 뜻하는 꼬리표를 부여한다.
- 비정상 검출(anomaly detection)
  - 일단의 사건들 또는 물체들을 살펴보고 그중 특이하거나 비정상적인 것들을 골라낸다.
  - Ex) 신용카드 사기 검출: 신용카드 회사는 사용자의 카드 결제 습관을 모형화해서, 사용자가 카드를 잘못 사용했을 가능성이 큰 거래를 찾아낸다.
- 합성과 표본추출(synthesis and sampling)
  - 훈련 자료에 있는 견본들과 비슷한 새 견본들을 생성
  - 매체 관련 응용에서 대량의 콘텐츠를 사람이 직접 생성하려면 비용이 너무 크거나, 너무 지루하거나, 시간이 너무 많은 경우 이러한 기계 학습을 이용한 합성과 표본추출이 유용
  - Ex) 비디오 게임에서 큰 물체나 지형에 입힐 텍스터를 사람이 일일이 픽셀을 찍어서 그리는 대신 컴퓨터를 이용해서 자동으로 생성

5.1.1 과제 T

- 결측값 대체(imputation of missing values)
  - 새 견본 $x \in \mathbb{R}^n$을 입력받는데, 그 x에는 일부 성분 x_i들이 빠져 있다. 알고리즘은 이러한 결측 성분들의 값을 예측하는 것
- 잡음 제거(denoising)
  - 깨끗한 견본(clean example) $x \in \mathbb{R}^n$이 알려지지 않은 어떤 손상 과정을 거친 결과로 만들어진 손상된 견본(corrupted example) $\tilde{x} \in \mathbb{R}^n$을 입력받는다. 학습 알고리즘은 송산된 견본 $\tilde{x}$로부터 깨끗한 견본 x를 예측해야한다. 좀 더 일반적으로, 학습알고리즘은 조건부 확률분포 $\mathrm{p}(x|\tilde{x})$를 예측해야 한다.
- 밀도추정 또는 확률질량함수 추정
  (density estimation or probability mass function estimation)
  - 함수 $p_{\text{모형}}: \mathbb{R}^n \rightarrow \mathbb{R}$을 배워야한다. 여기서 $p_{\text{모형}}(x)$는 견본들이 있던 공간에 따라서는 확률밀도함수로 해석할 수도 있고(x가 연속 확률변수일 때) 확률질량함수로 해석할 수도 있다(x가 이산 확률변수일 때).

5.1.2 성과 측도 P

- P는 시스템이 수행하는 과제 T에 따라 다름

- 분류, 결측 입력이 있는 자료의 분류, 전사
  - 정확도: 주어진 견본 전체 중 모형이 정확한 결과를 출력한 견본들의 비율
  - 오류율: 모형이 틀린 결과를 출력한 견본들의 비율

- 밀도 추정 같은 과제
  - 평균 로그 확률 (가장 흔히 사용)

- 입력한 적이 없는 자료에 대한 기계 학습 알고리즘의 성과 측도를 얻기 위해, 기계 학습 시스템을 훈련하는데 사용한 것과 다른 **test data set**으로 학습 시스템의 성과 측정

5.1.3 경험 E

- 비지도 학습
  - 다수의 특징을 담은 자료 집합(dataset)을 경험하고, 구조가 가진 유용한 성질들을 배움
  - 주어진 자료 집합을 만들어 낸 생성원에 해당하는 전체 확률분포를 명시적으로 또는 암묵적으로 학습하는 것.
- 지도 학습
  - 다수의 특징을 담은 dataset을 경험하나, 이 경우는 자료집합의 각 견본에 label 또는 targe과 연관되어 있음

- 비지도 학습과 지도학습 모두 수행할 수 있는 예
  - 벡터 $x \in \mathbb{R}^n$에 대한 결합확률분포

$$p(\mathbf{x}) = \prod_{i=1}^{n} p(\mathrm{x}_i \mid \mathrm{x}_1, \ldots, \mathrm{x}_{i-1}). \qquad (5.1)$$

5.1.3 경험 E

- $p(y|x)$를 학습하는 지도 학습 문제를 전통적인 비지도 학습 기술들을 이용해서 푸는 경우
  - 결합분포 $p(x, y)$를 배운 후 식 5.2를 추론한다.

$$p(y \mid \mathbf{x}) = \frac{p(\mathbf{x}, y)}{\sum_{y'} p(\mathbf{x}, y')}. \qquad (5.2)$$

- 준지도학습(semi-supervised learning)
  - 학습 지도를 위한 표지(목표)가 붙은 견본들과 그렇지 않은 견본들로 구성된 자료집합을 사용
  - 다중 인스턴스 학습에서는 자료 집합에 특정 부류의 견본이 들어 있는지의 여부를 학습 알고리즘에 알려주지만, 자료 집합의 개별 견본에는 아무런 표지도 붙어 있지 않음

5.1.3 경험 E

- 강화학습
  - 경험하는 자료 집합이 고정되지 않는 기계 학습 알고리즘 예
  - 하나의 환경과 상호작용한다. 즉, 학습 시스템과 그 경험들 사이에는 하나의 되먹임 루프(feedback loop)가 존재한다.

- 설계 행렬
  - 한 행이 하나의 견본이고, 각 열은 해당 견본의 각 특징인 행렬이다.
  - Ex) Iris 자료 집합은 각각 네 개의 특징으로 이루어진 견본 150개로 구성된다. 이러한 자료 집합은 성분 $X_{i,1}$이 i번 식물의 꽃받침 길이, $X_{i,2}$가 i번 식물의 꽃받침 너비, 등등인 행렬 $X \in \mathbb{R}^{150\times 4}$으로 서술할 수 있음

- 하나의 자료 집합을 설계 행렬로 서술하기 위해서는 각 견본을 하나의 벡터로 서술 할 수 있어야함.
  - 서술하기 어려운 경우: 사진 => 픽셀 수가 다름
  - 이런 경우 자료 집합을 m개의 원소(성분)로 이루어진 하나의 집합 $\{x^{(1)}, x^{(2)}, \ldots, x^{(m)}\}$으로 서술한다.

5.1.4 예제: 선형회귀

- 목표: 벡터 $x \in \mathbb{R}^n$(입력)으로부터 스칼라 $y \in \mathbb{R}$의 값(출력)을 예측하는 것

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x}. \tag{5.3}$$

- 매개변수는 시스템의 행동을 제어하는 값
  - 위 예시 모형은 각 특징 x_i에 매개변수 w_i를 곱한 것들을 모두 더한다.
  - 이러한 w를, 각 특징이 예측에 미치는 영향을 결정하는 **가중치(weight)**들의 집합이라고 생각하면 됨

- 선형회귀의 과제 T
  - 식 5.3을 출력함으로써 x로부터 y를 예측하는 것

5.1.4 예제: 선형회귀

- 성과 측도 P
  - 평균제곱오차(mean squared error, MSE)

$$\mathrm{MSE}_{\text{시험}} = \frac{1}{m}\sum_i (\hat{\boldsymbol{y}}^{(\text{시험})} - \boldsymbol{y}^{(\text{시험})})_i^2. \tag{5.4}$$

  - $\hat{y}^{(\text{시험})} = y^{(\text{시험})}$일 때 오차 측도 0

$$\mathrm{MSE}_{\text{시험}} = \frac{1}{m}\| \hat{\boldsymbol{y}}^{(\text{시험})} - \boldsymbol{y}^{(\text{시험})} \|_2^2. \tag{5.5}$$

5.1.4 예제: 선형회귀

- $\text{MSE}_{\text{훈련}}$을 최소화하는 방법은 기울기 벡터가 0인 점을 구하는 것

$$\nabla_{\boldsymbol{w}} \text{MSE}_{\text{훈련}} = 0 \tag{5.6}$$

$$\Rightarrow \nabla_{\boldsymbol{w}} \frac{1}{m} \| \hat{\boldsymbol{y}}^{(\text{훈련})} - \boldsymbol{y}^{(\text{훈련})} \|_2^2 = 0 \tag{5.7}$$

$$\Rightarrow \frac{1}{m} \nabla_{\boldsymbol{w}} \| \boldsymbol{X}^{(\text{훈련})} \boldsymbol{w} - \boldsymbol{y}^{(\text{훈련})} \|_2^2 = 0 \tag{5.8}$$

$$\Rightarrow \nabla_{\boldsymbol{w}} (\boldsymbol{X}^{(\text{훈련})} \boldsymbol{w} - \boldsymbol{y}^{(\text{훈련})})^\top (\boldsymbol{X}^{(\text{훈련})} \boldsymbol{w} - \boldsymbol{y}^{(\text{훈련})}) = 0 \tag{5.9}$$

$$\Rightarrow \nabla_{\boldsymbol{w}} (\boldsymbol{w}^\top \boldsymbol{X}^{(\text{훈련})\top} \boldsymbol{X}^{(\text{훈련})} \boldsymbol{w} - 2\boldsymbol{w}^\top \boldsymbol{X}^{(\text{훈련})\top} \boldsymbol{y}^{(\text{훈련})} + \boldsymbol{y}^{(\text{훈련})\top} \boldsymbol{y}^{(\text{훈련})}) = 0 \tag{5.10}$$

$$\Rightarrow 2\boldsymbol{X}^{(\text{훈련})\top} \boldsymbol{X}^{(\text{훈련})} \boldsymbol{w} - 2\boldsymbol{X}^{(\text{훈련})\top} \boldsymbol{y}^{(\text{훈련})} = 0 \tag{5.11}$$

$$\Rightarrow \boldsymbol{w} = (\boldsymbol{X}^{(\text{훈련})\top} \boldsymbol{X}^{(\text{훈련})})^{-1} \boldsymbol{X}^{(\text{훈련})\top} \boldsymbol{y}^{(\text{훈련})} \tag{5.12}$$

5.1.4 예제: 선형회귀

- 선형회귀 문제의 예

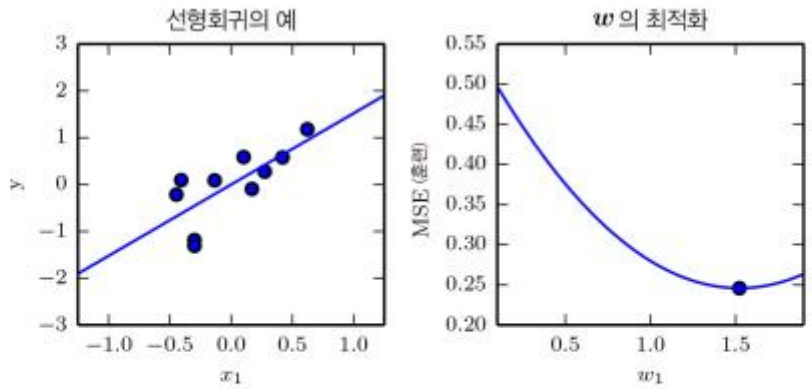

그림 5.1: 선형회귀 문제의 예. 훈련 집합은 열 개의 자료점으로 이루어지며, 각 자료점은 하나의 특징을 담는다. 견본의 특징이 하나뿐이므로 모형이 학습해야 할 가중치 벡터 $\boldsymbol{w}$의 성분도 w_1 하나뿐이다. (왼쪽) 선형회귀 모형이, $y = w_1 x$가 훈련 집합의 모든 점을 최대한 가까이 통과하는 직선이 되는 w_1 값을 배웠음을 알 수 있다. (오른쪽) 평균제곱오차 그래프. 표시된 점은 표준방정식으로 구한 w_1의 값을 나타낸다. 점의 위치에서 보듯이, 그 점에서 훈련 집합에 대한 평균제곱오차가 최소화된다.

5.1.4 예제: 선형회귀

- 절편 항 b가 추가된 선형회귀 모형

$$\hat{y} = \boldsymbol{w}^\top \boldsymbol{x} + b. \tag{5.13}$$

- 사상은 여전히 일차함수이지만, 특징에서 예측으로의 사상은 어파인 함수이다.

- 절편 항 b를 추가하는 대신, 그냥 가중치들만 있는 모형을 사용하되 항상 1로 설정되는 성분 하나를 x에 추가하는 것도 가능
  - 이러한 추가 1 성분의 가중치는 절편 항과 같은 역할을 한다.

5.2 수용력, 과대적합, 과소적합

- 기계학습의 주된 어려움
  - 모형을 훈련하는 데 사용한 입력뿐만 아니라 새로운, 이전에 본 적이 없는 입력에 대해서도 알고리즘이 잘 작동하게 만드는 것

- 일반화(generalization)
  - 이전에 관측한 적이 없는 입력들에 대해 잘 작동하는 능력

- 훈련 오차(training error)
  - 모형이 예측한 값과 훈련 집합에 있는 참값 사이의 오차

- 일반화 오차(generalization error) or 시험 오차(test error)
  - 새 입력에 대한 오차의 기댓값
    - 실제 실행 시 학습 시스템이 마주할 입력들의 분포에서 추출한 모든 가능한 서로 다른 입력에 대한 오차들의 평균

5.2 수용력, 과대적합, 과소적합

- 자료 생성 과정
  - 자료 집합들에 관한 확률분포에 기초해서 훈련 자료와 시험 자료 생성
- 독립동일분포 가정(i.i.d)
  - 각 자료 집합의 견본들이 서로 독립이고, 훈련 집합과 자료 집합의 견본들이 같은 확률분포에 따라 동일하게 분포
- 자료 생성 분포
  - 공통의 확률 분포
- 무작위로 선택된 모형의 기대 훈련 오차가 그 모형의 기대 시험 오차와 같다.
- 기대 시험 오차 >= 기대 훈련 오차

5.2 수용력, 과대적합, 과소적합

- 기계 학습 알고리즘의 성과를 결정하는 능력
  - 훈련 오차를 작게 만드는 능력
  - 훈련 오차와 시험 오차의 차이를 작게 만드는 능력

- 주요 장애물
  - 과소적합: 모형이 훈련 집합의 오차 값을 충분히 작게 만들지 못할 때 발생
  - 과대적합: 훈련 오차와 시험 오차의 차이가 너무 클 때 발생

- 수용력
  - 모형이 다양한 함수들에 적합하는 능력

- 수용력(capacity)를 바꾸어서 과대적합 또는 과소적합 가능성 제어

5.2 수용력, 과대적합, 과소적합

- 수용력을 제어하는 한 가지 방법
  - 가설 공간을 적절히 선택하는 것
    - 가설 공간이란 학습 알고리즘이 하나의 해답으로 선택할 수 있는 함수들의 집합

$$\hat{y} = b + wx. \tag{5.15}$$

$$\hat{y} = b + w_1 x + w_2 x^2. \tag{5.16}$$

$$\hat{y} = b + \sum_{i=1}^{9} w_i x^i. \tag{5.17}$$

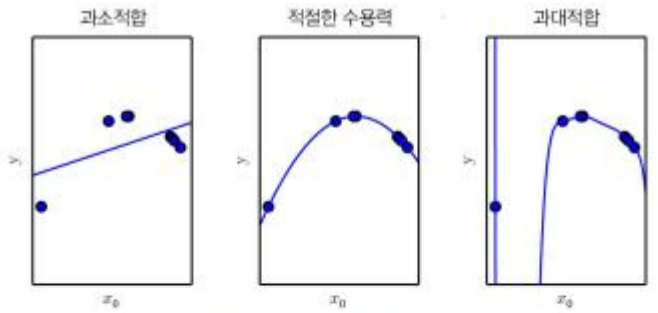

그림 5.2: 예제 훈련 집합에 세 개의 모형을 적합시킨 예. 예제 훈련 자료는 인공적으로, x 값들을 무작위로 추출하고 하나의 이차함수를 평가해서 y를 결정론적으로 선택하는 방식으로 생성한 것이다. (왼쪽) 일차함수는 자료에 과소적합한다. 일차함수는 자료가 나타내는 곡률을 포착하지 못한다. (가운데) 이차함수는 자료에 잘 적합하며, 새로운 자료점들에도 잘 일반화된다. 현저한 과소적합이나 과대적합은 발생하지 않는다. (오른쪽) 9차 다항식은 자료에 과대적합한다. 이 예는 무어-펜로즈 유사역행렬을 이용해서 과소결정(underdetermined; 미지수가 방정식보다 많은) 표준방정식을 푼다. 그 해는 주어진 모든 훈련 자료점을 정확히 지나가지만, 안타깝게도 원래의 바탕 함수의 진정한 구조를 모형이 포착하지는 못했다. 두 훈련 자료점 사이에 바탕 함수에는 없는 깊은 골짜기가 존재함을 주목하기 바란다. 또한, 자료의 오른쪽에서 함수가 급격기 증가하는데, 원래의 함수는 그 부분에서 감소한다.

5.2 수용력, 과대적합, 과소적합

- 표현 수용력
  - 함수족 전체를 감당할 수 있는 수용력

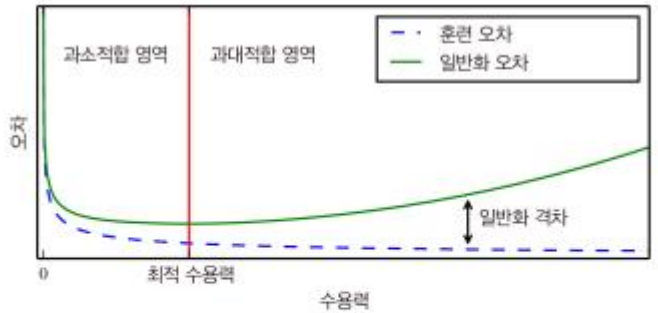

그림 5.3: 수용력과 오차의 전형적인 관계. 훈련 오차와 시험 오차가 다르게 행동한다. 그래프의 왼쪽 끝에서는 훈련 오차와 일반화 오차가 둘 다 높다. 이런 부분을 **과소적합 기간**(underfitting regime)이라고 부른다. 수용력이 증가하면서 훈련 오차가 감소하지만, 훈련 오차와 일반화 오차의 격차가 커진다. 그러다가 결국에는 그 격차가 훈련 오차의 감소량을 넘으면서 **과대적합 기간**(overfitting regime)이라고 부르는 영역에 들어선다. 이 영역에서는 수용력이 **최적 수용력**(optimal capacity)보다 크다.

- 오컴의 면도날
  - 알려진 관찰들을 동등하게 잘 설명하는 여러 가설이 있을 때, 가장 “단순한” 가설을 골라야한다는 것

- 바프니크-체르보넨키스 차원
  - 분류기가 임의로 label을 부여할 수 있는 서로 다른 m가지 자료점 x들로 이루어진 훈련 집합이 존재한다는 조건을 만족하는 m의 가장 큰 값이 그 분류기의 VC 차원이다.

5.2 수용력, 과대적합, 과소적합

- 비매개변수 모형
- 최근접 이웃 회귀
- 매개변수적 학습 알고리즘을 비매개변수 학습 알고리즘으로 만들 수 있음
- 베이즈 오차
  - 신탁 모형의 예측과 실제 분포 p(x,y) 사이의 오차

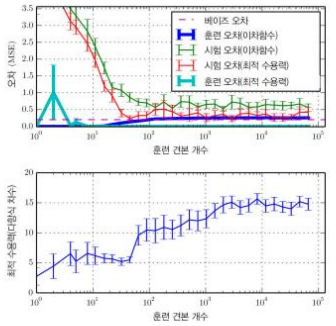

그림 5.4: 훈련 오차와 시험 오차에 대한, 그리고 모형의 최적 수용력에 대한 훈련 자료 집합 크기의 효과. 5차 다항식에 적당한 양의 잡음을 추가해서 하나의 시험 집합을 생성하고, 그런 다음 크기가 다른 여러 개의 훈련 집합을 생성해서 인위적인 회귀 문제를 만들었다. 그래프들은 크기별로 40개의 서로 다른 훈련 집합을 생성해서 측정한 오차들을 보여준다. 오차 막대는 95% 신뢰구간을 나타낸다. (위) 서로 다른 두 모형의 훈련 집합 MSE와 시험 집합 MSE. 한 모형은 이차이고 다른 한 모형은 시험 오차가 최소가 되는 차수를 선택한 것이다. 둘 다 닫힌 형식으로 적합된다. 이차 모형의 경우 훈련 집합의 크기가 증가함에 따라 훈련 오차가 증가한다. 이는 자료 집합이 클수록 적합이 어렵기 때문이다. 한편 시험 오차는 감소하는데, 이는 잘못된 가설이 적음수록 모형이 훈련 자료와 더 잘 부합하기 때문이다. 이차 모형은 그 수용력이 주어진 과제를 풀기에 부족하기 때문에 시험 오차가 높은 값으로 점근한다. 최적 수용력 모형의 경우 시험 오차는 베이즈 오차에 점근한다. 훈련 오차는 베이즈 오차에 훨씬 못 미칠 수 있는데, 이는 훈련 알고리즘이 훈련 집합의 특정 인스턴스들을 기억하는 능력 때문이다. 훈련 집합의 크기가 무한대에 가까워짐에 따라, 임의의 고정 수용력 모형(지금 예에서는 이차 모형)의 훈련 오차는 적어도 베이즈 오차까지는 증가하게 된다. (아래) 훈련 집합의 크기가 증가함에 따라 최적 수용력(그래프에는 최적의 다항식 회귀 함수의 차수로 표시되어 있다)이 증가한다. 주어진 과제를 풀기에 충분한 수준에 도달한 후에는 최적 수용력이 거의 변하지 않는다(평평한 대지를 형성).

5.2.1 공짜 점심 없음 정리

- 모든 가능한 자료 생성 분포에 대해 평균을 구한다고 할 때, 이전에 관측한 적이 없는 자료점들을 분류하는 과제에서 모든 분류 알고리즘의 오차율은 서로 같다.

- 모든 가능한 자료 생성 분포에 대해 평균을 구할 때만 성립

- 확률분포들의 종류에 일정한 제약을 둔다면, 이런 분포들에 대해서는 잘 작동하는 학습 알고리즘을 설계하는 것이 가능

- 기계 학습의 연구 목표
  - 어떤 보편적인 학습 알고리즘이나 절대적으로 최고인 학습 알고리즘을 찾는 것 X
  - 인공지능 에이전트가 경험할 '현실 세계 ' 에서 어떤 종류의 분포들이 의미가 있는지, 그리고 이런 종류의 자료 생성 분포 들에서 뽑은 자료에 대해 잘 작동하는 기계 학습 알고리즘의 종류는 무엇인지 이해하는 것

5.2.2 정칙화

- 특정 해들에 대한 선호도를 명시적으로 또는 암묵적으로 표현하는 여러 접근 방식
  - 훈련 오차가 줄지는 않더라도 일반화 오차를 줄이려는 의도로 학습 알고리즘에 가하는 모든 종류의 수정이 정칙화에 해당

$$J(\boldsymbol{w}) = \mathrm{MSE}_{\text{훈련}} + \lambda \boldsymbol{w}^\top \boldsymbol{w}. \tag{5.18}$$

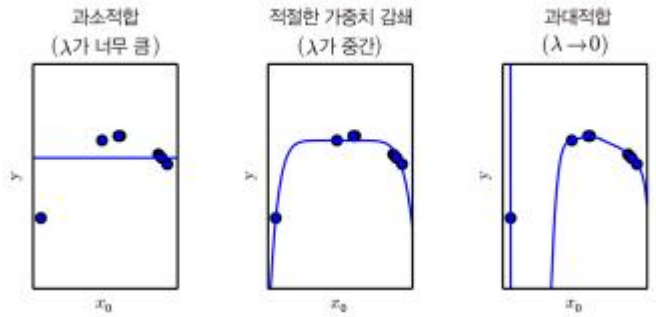

그림 5.5: 고차 다항식 회귀 모형을 그림 5.2의 예제 훈련 자료 집합에 적합시킨 예. 원래의 함수는 이차함수이지만, 여기서는 다항식의 차수가 9인 모형들만 사용했다. 그런 고차 모형들의 과대적합을 방지하기 위해 가중치 감쇄의 양을 여러 가지로 변화해서 적용했다. (왼쪽) λ가 아주 크면 모형은 기울기가 아예 없는 함수를 배우게 된다. 그러한 함수는 오직 한 가지 상수만 예측하므로 과소적합이 발생한다. (가운데) λ의 값이 중간 정도일 때 학습 알고리즘은 적절한 형태의 곡선을 복원한다. 모형 자체는 그보다 훨씬 복잡한 형태의 곡선들을 표현할 능력이 있지만, 가중치 감쇄 덕분에 모형은 더 작은 계수들로 서술되는 좀 더 단순한 함수를 선택한다. (오른쪽) 가중치 감쇄가 0에 접근할 때(즉, 무어-펜로즈 역함수를 이용해서 최소한의 정칙화로 미결 문제를 푸는 경우), 9차 다항식은 그림 5.2에서처럼 현저한 과대적합을 보인다.