심층 학습

윤예준

5.9 확률적 경사 하강법(stochastic gradient descent, SGD) - 1

- 음의 로그가능도(negative log-likelihood)

$$J(\boldsymbol{\theta}) = \mathbb{E}_{\mathbf{x},y\sim \hat{p}_{\text{자료}}} L(\boldsymbol{x}, y, \boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}). \qquad (5.96)$$

- 샘플별 손실 함수: $L(x, y, \theta) = -\log p(y|x;\theta)$

$$\nabla_{\boldsymbol{\theta}} J(\boldsymbol{\theta}) = \frac{1}{m}\sum_{i=1}^{m} \nabla_{\boldsymbol{\theta}} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}). \qquad (5.97)$$

- 미니배치(minibatch)
  - $\mathbb{B} = \{x^{(1)}, \ldots, x^{(m')}\}$

5.9 확률적 경사 하강법(stochastic gradient descent, SGD) - 2

- 미니배치(minibatch)
  - $\mathbb{B} = \{x^{(1)}, \ldots, x^{(m')}\}$
- 미니배치 기울기 추정값 공식

$$\boldsymbol{g} = \frac{1}{m'} \nabla_{\boldsymbol{\theta}} \sum_{i=1}^{m'} L(\boldsymbol{x}^{(i)}, y^{(i)}, \boldsymbol{\theta}). \tag{5.98}$$

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} - \epsilon \boldsymbol{g}. \tag{5.99}$$

5.10 기계 학습 알고리즘 만들기

- 심층학습 알고리즘
  - 비교적 간단한 '조리법 ' 을 필요에 따라 구체적으로 적용한 사례에 해당
  - 비교적 간단한 '조리법': 자료 집합의 명세와 비용함수, 최적화 절차, 모형 결합

- 예시: 선형 회귀 알고리즘의 비용함수와 모형 명세

$$J(\boldsymbol{w}, b) = -\mathbb{E}_{\mathbf{x}, y \sim \hat{p}_{\text{자료}}} \log p_{\text{모형}}(y \mid \boldsymbol{x}) \qquad (5.100)$$

$$p_{\text{모형}}(y|x) = \mathcal{N}(y; x^T w + b, 1)$$

$$J(\boldsymbol{w}, b) = \lambda \|\boldsymbol{w}\|_2^2 - \mathbb{E}_{\mathbf{x}, y \sim \hat{p}_{\text{자료}}} \log p_{\text{모형}}(y \mid \boldsymbol{x}). \qquad (5.101)$$

$$J(\boldsymbol{w}) = \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{자료}}} \|\boldsymbol{x} - r(\boldsymbol{x}; \boldsymbol{w})\|_2^2 \qquad (5.102)$$

5.11 심층 학습의 개발 동기가 된 기준 문제점들

- 차원의 저주
- 국소 일치성과 평활성 정칙화
- 다양체 학습

5.11.1 차원의 저주

- 자료의 차원이 높을 때(즉, 차원들이 많을 때) 특히 풀기 어려워지는 문제

- 통계적 난제

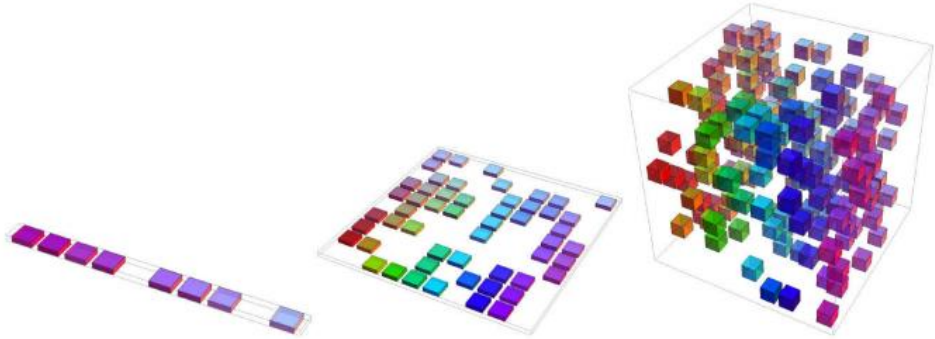

그림 5.9: 자료의 관련 차원의 수가 늘어남에 따라(왼쪽에서 오른쪽), 우리가 관심이 있는 가능한 구성의 수가 지수적으로 증가한다. (왼쪽) 이 1차원의 예에는 변수가 하나뿐이고, 구별해야 할 변수의 서로 다른 값은 단 10가지이다. 각 값은 각각의 영역(그림의 정사각형 격자 칸)에 대응된다. 모든 영역에 견본이 배정되기에 충분할 정도로 견본들이 많으면 학습 알고리즘은 수월하게 잘 일반화된다. 한 가지 간단한 일반화 방법은 각 영역 안에서 대상 함수의 값을 추정하는 것이다(더 나아가서, 이웃 영역들의 값과 보간할 수도 있다). (가운데) 2차원 자료 공간에서는 각 변수의 서로 다른 10가지 값을 구별하기가 좀 더 어렵다. 총 $10\times10=100$개의 영역을 관리해야 하며, 그러한 모든 영역에 견본이 배정될 정도로 견본의 수가 많아야 한다. (오른쪽) 3차원의 경우에는 영역의 수가 $10^3=1{,}000$개이고 견본 개수 역시 그 정도의 규모로 많아야 한다. d차원의 경우 축마다 구별되는 값이 각각 v가지이면 총 $O(v^d)$개의 영역과 견본이 필요하다. 그림들은 니콜라스 차파도스Nicolas Chapados가 친절하게 제공했다.

5.11.2 국소 일치성과 평활성 정칙화

- 기계 학습 알고리즘이 잘 일반화되려면, 알고리즘이 배워야할 함수의 종류에 관한 **사전 믿음(prior belief)**들을 알고리즘에 제공할 필요가 있음.

- 사전 믿음은 함수 자체에 직접 영향을 미치고 매개변수들에는 간접적으로 영향을 미침

- 특정 부류의 함수들을 선호하도록 알고리즘을 편향되게 구현함으로써 사전 믿음들을 암묵적으로 지정할 수 있음.

- 이러한 편향을 여러 함수의 관한 믿음의 정도를 확률분포의 형태로 표현하곤 하지만, 항상 그런 것은 아님. 심지어, 확률분포의 형태로 표현 불가능한 경우도 있음.

5.11.2 국소 일치성과 평활성 정칙화

- **평활성(매끈함) 사전분포(smoothness prior)**
  - 국소 불변성 사전분포(local constancy prior)라고도 불림
  - **함수가 작은 영역 안에서 아주 크게 변해서는 안 된다는 제약을 나타냄**

- 간단한 축에 속하는 여러 알고리즘은 좋은 일반화를 보장하기 위한 수단이 이 사전분포밖에 없음.

- 사전 믿음을 표현하는 방법은 여러가지임. 이런 여러 표현 방법은 모두, 학습 과정이 특정한 조건을 만족하는 함수f^*를 배우도록 격려하기 위해 고안된 것.

$$f^*(\boldsymbol{x}) \approx f^*(\boldsymbol{x} + \epsilon) \qquad (5.103)$$

5.11.2 국소 일치성과 평활성 정칙화

- 국소 불변성 접근 방식의 극단적인 예
  - k-최근접 이웃 알고리즘에 속하는 알고리즘
    - 이런 알고리즘의 예측값은, k개의 최근접 이웃이 동일한 점들.
    - k-최근접 이웃 알고리즘은 근처에 있는 훈련 샘플들의 출력을 그대로 돌려줌.
    - 그러나 대부분의 핵 기계 방법들은 이웃 훈련 샘플들에 연관된 훈련 집합 출력들을 보간해서 돌려줌.
    - 이러한 핵들의 중요한 부류로 **국소 핵(local kernel)**들이 있음.
    - 국소 핵에서는 $u = v$일 때 $k(u, v)$가 크고, u와 v가 멀어질수록 $k(u, v)$가 감소
- 평활성에만 의존하는 학습의 한계들이 적용되는 예
  - 결정트리
    - 입력 공간을 잎 노드 개수만큼의 영역들로 분할하고 각 영역에 대해 개별적인 하나의 매개변수를 사용하기 때문.
    - 대상 함수를 정확히 표현하는 데 필요한 결정 트리의 잎 노드가 적어도 n개라고 할 때, 학습 모델을 그 결정 트리에 잘 적합시키려면 적어도 n개의 훈련 샘플이 필요.
    - 예측 결과의 통계적 신뢰도를 일정 수준 이상으로 확보하려면 훈련 샘플의 개수가 n의 몇 배이어야함.

5.11.2 국소 일치성과 평활성 정칙화

- 일반화
  - 입력 공간을 O(k)개의 서로 다른 영역들로 구분하기 위해서는 O(k)개의 견본이 필요.
  - 일반적으로 O(k)개의 영역 각각에 O(1)개의 매개변수가 배정되므로, 매개변수 개수 역시 O(k)일 때가 많음.

- 훈련 샘플보다 많은 수의 영역을 구분할 수 있는 복잡한 함수를 표현하는 방법은 바탕 함수가 매끄럽다는 가정만으로는 학습 과정이 그런 함수를 표현할 수 없음.
- 평활성 가정과 관련 비매개변수적 알고리즘들은, 학습할 실제 바탕 함수의 대부분의 봉우리(peak)들에서는 큰(높은) 점들을 관측하고 대부분의 계곡들에서는 값이 작은(낮은) 점들을 관측하기만 한다면 대단히 잘 작동
  - 그러나 학습할 함수가 충분히 매끄럽고 그리 많지 않은 차원들에서만 변화할 때만 충족
  - 차원이 높을 때는 아주 매끄러운 함수도 문제가 될 수 있음.
  - 함수가 각 차원에서 서로 다른 방식으로 변할 수 있기 때문.

5.11.2 국소 일치성과 평활성 정칙화

- 복잡한 함수
  - 함수 효율적으로 표현하는 것 가능
  - 새 입력들에 잘 일반화 되는 것 모두 가능
  - 핵심
    - O(k)개의 표본들로 그보다 훨씬 많은 영역을 정의 할 수 있음
    - 단, 바탕 자료 생성 분포에 관한 추가적인 가정들을 도입해서 영역들 사이에 일정한 의존성을 부여해야함.

- 심층 학습의 핵심 착안
  - 자료가 **인자들의 조합**을 통해서 생성된다는 것

5.11.3 다양체 학습

- 다양체
  - 연결된 영역
  - 수학적으로, 다양체는 각 점 주변의 이웃과 연관된 점들의 집합

- 기계학습에서의 다양체
  - 그냥 더 높은 차원의 공간에 내장된, 그러나 그보다 낮은 차원 또는 자유도(degree of freedom)로도 잘 근사할 수 있는 일단의 연결된 점들로 느슨하게 정의

- 다양체 학습(manifold learning) 알고리즘
  - $\mathbb{R}^n$공간의 대부분이 유효하지 않은 입력으로 구성되고 흥미로운 입력들은 일부 점들로 이루어진 몇몇 다양체들에만 존재
  - 학습 대상 함수의 출력에서 흥미로운 변동들은 그 다양체에 놓인 방향에서만 발생하거나 한 다양체에서 다른 다양체로 이동할 때만 발생한다고 가정함으로써 어려움 극복
  - 다양체 학습은 연속값 자료를 사용하는 비지도 학습을 위해 도입되었지만, 다양체 학습의 확률 집중 착안을 이산 자료를 사용하는 지도 학습으로도 일반화 할 수 있음 => 핵심 가정은 확률질량이 고도로 집중되어 있다는 것

5.11.3 다양체 학습

- 다양체 가설 지지하는 논점 2가지
  - 논점1. 실세계에서 볼 수 있는 이미지나 텍스트 문자열, 음향에 관한 확률분포가 실제로 집중되어 있다는 것
    - 확률분포가 집중되어 있다는 것이 자료가 적당히 적은 수의 다양체들에 놓여있음을 말해주는 충분조건은 아님
    - 다양체상에서 변환을 적용해서 도달할 수 있는 이웃 견본들이 존재하는 점을 입증 해야함
  - 논점2. 이웃둘과 변환들을 적어도 비공식적으로는 상상할 수 있다는 것
    - 이미지 처리의 경우에는 이미지 공간 안의 한 다양체 안에서 이동하는 데 필요한 다양한 변환을 생각해 낼 수 있음
    - 대부분의 응용에서는 하나가 아니라 여러 개의 다양체가 관여할 가능성이 큼
- 자료가 저차원 다양체에 놓여있는 경우
  - 기계 학습 알고리즘이 이런 자료를 $\mathbb{R}^n$의 좌표가 아니라 그 다양체를 기준으로 한 좌표로 표현하는 것이 훨씬 자연스러움.