심층학습 5장. 기계 학습의 기초 (5.3장~)

송선영

5.3 초매개변수와 검증 집합

- **초매개변수 (hyperparameter)**
  - 알고리즘의 행동을 제어하는데 사용
  - 초매개변수들의 값을 학습 알고리즘 자체가 적응시키지는 않음
    - 단, 한 학습 알고리즘이 다른 학습 알고리즘의 최상의 초매개변수 값들을 배우는 방식은 가능

  - 학습 알고리즘이 학습하지 않는 설정 중 최적화하기 어려운 설정을 초매개변수로 두는 경우도 있음
  - 더 자주 쓰이는 접근 방식은, 훈련 집합으로 학습하기에 적합하지 않은 설정을 초매개변수로 두는 것임

  - **만일 초매개변수를 훈련 집합을 통해 학습한다면 과대적합이 일어남**
    - 이를 해결하기 위해 훈련 알고리즘이 관측하지 않은 견본들로 이루어진 검증 집합이 필요
    - 검증 집합은 초매개변수를 선택할때 쓰임
    - 일반적으로 훈련 자료의 약 80%를 훈련에, 나머지 20%를 검증에 사용함

- **교차검증**
  - 자료 집합을 하나의 고정된 훈련집합과 하나의 고정된 시험집합으로 나눌 경우, 시험집합이 너무 작으면 문제가 될 수 있음 → 시험 집합이 작다는 것은 추정된 평균 시험 오차에 대한 통계적 불확실성이 존재한다는 의미
  - 그렇기 때문에 교차검증이 필요함

  - ***k* 중 교차 검증 (*k*-fold cross-validation)**
    - 자료 집합을 k 개의 서로 겹치지 않는 부분집합들로 나눔
    - *k*회의 시행으로 얻은 수치들로 평균 시험 오차를 구함
    - 시행 *i*에서는 자료의 *i*번째 부분집합을 시험 집합으로 사용하고, 나머지는 훈련 집합을 사용함

5.4 추정량, 편향, 분산

- **점 추정**
  - 어떤 수량에 대한 최고의 예측 하나를 제공하는 것을 목표로 함

  - 집합 $\{x^{(1)}, \dots, x^{(m)}\}$이 독립적이고 동일하게 분포된 m 개의 자료점으로 이루어진 집합일 때, 이 자료집합의 점 추정량 또는 통계량을 아래 식으로 정의할 수 있음

$$\hat{\boldsymbol{\theta}}_m = g(\boldsymbol{x}^{(1)}, \dots, \boldsymbol{x}^{(m)}). \tag{5.19}$$

    - g가 반드시 θ의 참값에 가까운 하나의 값을 돌려주는 함수이어야 하는 것은 아님
    - g의 치역이 θ가 가질 수 있는 값들의 집합과 반드시 같아야 하는 것도 아님

    - 좋은 추정량은 훈련 자료를 생성한 바탕 θ의 참값에 가까운 값을 산출

- 빈도주의자의 관점에서
    - θ의 참값이 고정되어 있지만 알려지지는 않음
    - 점 추정량 $\hat{\theta}$는 하나의 확률변수임

    - 입력과 목표 변수의 관계를 추정하는 것도 점 추정 또는 함수 추정이라고 부름
        - 함수 추정: 주어진 입력 벡터x에 기초해서 변수y의 값을 예측하는 것

5.4 추정량, 편향, 분산

- **편향**
  - 추정량의 편향(bias)는 식 5.20과 같이 정의됨

$$\mathrm{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\theta}, \tag{5.20}$$

  - 기댓값 $\mathbb{E}$: 자료 전체에 대한 평균
  - θ: 자료 생성 분포를 정의하는데 쓰인 θ의 참 값

  - 불편추정량
    - 편향이 전혀 없을 때의 추정량
    - 즉, $bias(\hat{\theta}_m) = 0$ ($\mathbb{E}(\hat{\theta}_m) = \mathbb{E}(\hat{\theta}_m) = 0$)

  - 점근 불편추정량
    - $\log_{m\to\infty} bias(\hat{\theta}_m) = 0$ ($\log_{m\to\infty} \mathbb{E}(\hat{\theta}_m) = 0$) 인 추정량

5.4 추정량, 편향, 분산

- 예제: 베르누이 분포

$$\mathrm{bias}(\hat{\boldsymbol{\theta}}_m) = \mathbb{E}(\hat{\boldsymbol{\theta}}_m) - \boldsymbol{\theta}, \tag{5.20}$$

$$P(x^{(i)};\theta) = \theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}. \tag{5.21}$$

$$\hat{\theta}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}. \tag{5.22}$$

5.4 추정량, 편향, 분산

- **예제: 베르누이 분포**

$$\mathrm{bias}(\hat{\theta}_m) = \mathbb{E}[\hat{\theta}_m] - \theta \tag{5.23}$$

$$= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \theta \tag{5.24}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right] - \theta \tag{5.25}$$

$$= \frac{1}{m}\sum_{i=1}^{m} \sum_{x^{(i)}=0}^{1} \left(x^{(i)}\theta^{x^{(i)}}(1-\theta)^{(1-x^{(i)})}\right) - \theta \tag{5.26}$$

$$= \frac{1}{m}\sum_{i=1}^{m}(\theta) - \theta \tag{5.27}$$

$$= \theta - \theta = 0. \tag{5.28}$$

- $bias(\theta^\wedge) = 0$ 이므로 추정량 $\theta^\wedge$는 편향되지 않았음

5.4 추정량, 편향, 분산

- **예제: 가우스 분포의 평균의 추정량**

$$p(x^{(i)};\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2}\frac{(x^{(i)}-\mu)^2}{\sigma^2}\right). \tag{5.29}$$

$$\hat{\mu}_m = \frac{1}{m}\sum_{i=1}^{m} x^{(i)}. \tag{5.30}$$

$$\mathrm{bias}(\hat{\mu}_m) = \mathbb{E}[\hat{\mu}_m] - \mu \tag{5.31}$$

$$= \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right] - \mu \tag{5.32}$$

$$= \left(\frac{1}{m}\sum_{i=1}^{m} \mathbb{E}\left[x^{(i)}\right]\right) - \mu \tag{5.33}$$

$$= \left(\frac{1}{m}\sum_{i=1}^{m} \mu\right) - \mu \tag{5.34}$$

$$= \mu - \mu = 0. \tag{5.35}$$

5.4 추정량, 편향, 분산

- **예제: 가우스 분포의 분산의 추정량**

$$\hat{\sigma}_m^2 = \frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2. \tag{5.36}$$

$$\text{bias}(\hat{\sigma}_m^2) = \mathbb{E}[\hat{\sigma}_m^2] - \sigma^2. \tag{5.37}$$

$$\mathbb{E}[\hat{\sigma}_m^2] = \mathbb{E}\left[\frac{1}{m}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right] \tag{5.38}$$

$$= \frac{m-1}{m}\sigma^2. \tag{5.39}$$

$$\tilde{\sigma}_m^2 = \frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2. \tag{5.40}$$

$$\mathbb{E}[\tilde{\sigma}_m^2] = \mathbb{E}\left[\frac{1}{m-1}\sum_{i=1}^{m}\left(x^{(i)} - \hat{\mu}_m\right)^2\right] \tag{5.41}$$

$$= \frac{m}{m-1}\mathbb{E}[\hat{\sigma}_m^2] \tag{5.42}$$

$$= \frac{m}{m-1}\left(\frac{m-1}{m}\sigma^2\right) \tag{5.43}$$

$$= \sigma^2. \tag{5.44}$$

5.4 추정량, 편향, 분산

- **분산과 표준오차**
  - 추정량의 분산
    - 훈련 집합이 확률변수일 때의 분산

$$\mathrm{Var}(\hat{\theta}) \tag{5.45}$$

  - 추정량의 표준오차
    - 분산의 제곱근 $SE(\theta^{\wedge})$
    - 평균의 표준오차

$$\mathrm{SE}(\hat{\mu}_m) = \sqrt{\mathrm{Var}\left[\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right]} = \frac{\sigma}{\sqrt{m}}. \tag{5.46}$$

  - 표본이 충분히 클 때, 평균의 분포가 정규분포와 거의 같아진다는 중심극한정리에 근거하여, 기댓값의 참값이 임의의 선택된 구간에 속할 확률을 표준오차를 이용해 계산할 수 있음
    - Ex. 평균이 $\mu_m^{\wedge}$이고 ,분산이 $SE(\mu_m^{\wedge})^2$ 인 정규분포에서 평균을 포함하는 95% 신뢰구간은 식 5.47처럼 나타낼 수 있음

$$(\hat{\mu}_m - 1.96\mathrm{SE}(\hat{\mu}_m), \hat{\mu}_m + 1.96\mathrm{SE}(\hat{\mu}_m)). \tag{5.47}$$

    - 알고리즘 A의 오차에 대한 95% 신뢰구간의 상계가 알고리즘 B의 오차에 대한 95% 신뢰구간의 한계보다 작다면 알고리즘 A가 알고리즘 B보다 낫다는 분석을 얻을 수 있음

5.4 추정량, 편향, 분산

- 예제: 베르누이 분포

$$\mathrm{Var}\left(\hat{\theta}_m\right) = \mathrm{Var}\left(\frac{1}{m}\sum_{i=1}^{m} x^{(i)}\right) \tag{5.48}$$

$$= \frac{1}{m^2}\sum_{i=1}^{m} \mathrm{Var}\left(x^{(i)}\right) \tag{5.49}$$

$$= \frac{1}{m^2}\sum_{i=1}^{m} \theta(1-\theta) \tag{5.50}$$

$$= \frac{1}{m^2} m\theta(1-\theta) \tag{5.51}$$

$$= \frac{1}{m}\theta(1-\theta). \tag{5.52}$$

5.4 추정량, 편향, 분산

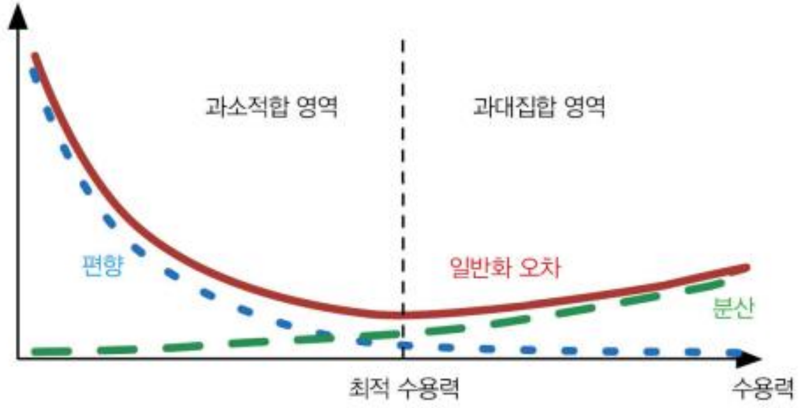
- **평균제곱오차의 최소화를 위한 편향과 분산의 절충**
  - 편향: 함수나 매개변수의 기댓값과 참값의 편차를 측정
  - 분산: 자료의 임의의 표본에서 기대할 수 있는 기대 추정값과 특정 표본에 대한 추정값의 편차를 측정

  - 편향과 분산을 저울질 할 때 (편향이 큰 추정량과 분산이 큰 추정량이 있을 때 등) 흔히 쓰이는 방법
    - 교차검증 적용
    - 평균제곱오차(MSE)를 비교해서 선택
      - MSE: 매개변수 θ의 추정량과 참값 사이의 전반적인 기대 편차
      - MSE가 더 작은 추정량을 선택한다는 것은 편향과 분산 둘 다 모두 작은 값을 유지되는 추정량을 선택한다는 것

      - 일반화 오차를 MSE로 측정하는 경우

$$\mathrm{MSE} = \mathbb{E}[(\hat{\theta}_m - \theta)^2] \quad (5.53)$$

$$= \mathrm{Bias}(\hat{\theta}_m)^2 + \mathrm{Var}(\hat{\theta}_m). \quad (5.54)$$



- 수용력 증가: 분산 증가, 편향 감소

5.4 추정량, 편향, 분산

- **일치성**
  - 자료 집합에 있는 자료점들의 개수 m이 증가함에 따라 점 추정값이 해당 매개변수들의 참값에 수렴하는 성질을 공식으로 표현하면 식 5.55

$$\operatorname{plim}_{m\to\infty}\hat{\theta}_m = \theta. \tag{5.55}$$

  - 식 5.55의 조건을 일치성 조건이라 부름
    - 식 5.55의 조건: 임의의 $\epsilon > 0$ 에 대해, $m \to \infty$에 따라 $P(|\hat{\theta}_m - \theta| > \epsilon) \to 0$
    - 식 5.55의 조건을 약한 일치성
    - $\theta^{\wedge}$가 θ에 거의 확실하게 수렴하는 것을 강한 일치성
  - 일치성 조건을 만족하는 추정량의 경우, 자료 견본들의 수가 늘어남에 따라 추정량에 의한 편향은 반드시 줄어듦
    - 단, 그 역은 성립하지 않음

Thank You

감사합니다.