

# Deep Learning Book

## 5.5 Maximum Likelihood Estimation

October 9, 2016

## 5.5 Maximum Likelihood Estimation

- ▶ “추정함수(estimator)”들은 어떻게 구해야 하는가? (때려맞추는 방법 말고...)
- ▶ 각각의 다른 모델들에 대한 좋은 (분포 추정)함수를 유도할 수 있는 일반적인 원리가 있는가?
- ▶ 어떤  $p_{\text{data}}(\mathbf{x})$  확률분포를 따르는 공간에서 독립적으로 추출한  $m$  개의 데이터 샘플 가정:

$$\mathbb{X} = \{\mathbf{x}^{(1)}, \dots, \mathbf{x}^{(m)}\} \quad (1)$$

- ▶ MLE(Maximum Likelihood Estimation)는 위의 확률분포를 설명하는 여러  $p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})$  중에서 다음 조건을 만족시키는 것을 찾는 것이다:

$$\boldsymbol{\theta}_{ML} = \arg \max_{\boldsymbol{\theta}} p_{\text{model}}(\mathbb{X}; \boldsymbol{\theta}) \quad (2)$$

$$= \arg \max_{\boldsymbol{\theta}} \prod_{i=1}^m p_{\text{model}}(\mathbf{x}^{(i)}; \boldsymbol{\theta}) \quad (3)$$

## 5.5 Maximum Likelihood Estimation (Cont.)

- ▶ (3) 수식은 underflow 가능성이 있으므로, 아래와 같이 바꿔 표현한다(최적화 관점에서 동일의미):

$$\theta_{ML} = \arg \max_{\theta} \sum_{i=1}^m \log p_{\text{model}}(\mathbf{x}^{(i)}; \theta) \quad (4)$$

- ▶ 훈련데이터로부터 정의되는 관측된(empirical) 분포인  $\hat{p}_{\text{data}}$ 에 대한 기대값 표현으로 다시 바꾸면:

$$\theta_{ML} = \arg \max_{\theta} \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(\mathbf{x}^{(i)}; \theta) \quad (5)$$

- ▶ KL-divergence 관점에서 보자. 관측분포인  $\hat{p}_{\text{data}}$ 는 (모델들 관점에서는) 일종의 상수항이므로, 아래 (6),(7) 수식을 최소화시킨다는 것은 위의 (5) 수식을 최대화한다는 의미이다. 즉, MLE는 모델분포를 관측분포와 최대한 근사하도록 만든다.

$$D_{KL}(\hat{p}_{\text{data}} \| p_{\text{model}}) = \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{data}}} [\log \hat{p}_{\text{data}}(\mathbf{x}) - \log p_{\text{model}}(\mathbf{x})] \quad (6)$$

$$\mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{data}}} [-\log p_{\text{model}}(\mathbf{x})] \Rightarrow (\text{Negative Log-Likelihood}) \quad (7)$$

## 5.5.1 Conditional Log-Likelihood and MSE

- ▶ MLE를 지도학습 관점에서 살펴보자.
- ▶  $\mathbf{X}, \mathbf{Y}$ 를 각각 관측된 데이터의 모든 입력, 출력이라고 할 때, 조건부  $MLE$ 는 다음과 같이 정의할 수 있다:

$$\boldsymbol{\theta}_{ML} = \arg \max_{\boldsymbol{\theta}} P(\mathbf{Y}|\mathbf{X}; \boldsymbol{\theta}) \quad (8)$$

- ▶ 훈련데이터가 i.i.d(independent and identically distributed) 라면, 아래와 같이 분해 표현 가능:

$$\boldsymbol{\theta}_{ML} = \arg \max_{\boldsymbol{\theta}} \sum_{i=1}^m \log P(\mathbf{y}^{(i)}|\mathbf{x}^{(i)}; \boldsymbol{\theta}) \quad (9)$$

## 5.5.1 Conditional Log-Likelihood and MSE (Cont.)

- ▶ 선형회귀를 앞에서 정의한 조건부 MLE((9) 수식)과 연관지어 생각해 보자.
- ▶ 선형회귀는 입력값  $\mathbf{x}$ 에 대해서 출력값  $\hat{y}$ 를 - MSE를 최소화시키면서 - 추정하는 알고리즘이다.
- ▶  $\hat{y}$ 라는 단 하나의 값을 출력하는 대신, 조건부 확률 분포  $p(y|\mathbf{x})$ 를 생성하는 모델을 생각해 보자:

$$p(y|\mathbf{x}) = \mathcal{N}(y; \hat{y}(\mathbf{x}; \mathbf{w}), \sigma^2) \quad (10)$$

$\hat{y}(\mathbf{x}; \mathbf{w}) \Rightarrow \text{predicts the mean}$

- ▶  $p(y|\mathbf{x}; \boldsymbol{\theta})$ 는 정규분포 정의에 따라 아래와 같음:

$$p(y|\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|y - \hat{y}\|^2}{2\sigma^2}\right) \quad (11)$$

## 5.5.1 Conditional Log-Likelihood and MSE (Cont.)

- ▶ (9),(11) 수식에 따라 조건부 log-likelihood를 전개하면,

$$\sum_{i=1}^m \log p(y^{(i)} | \mathbf{x}^{(i)}; \boldsymbol{\theta}) \quad (12)$$

$$= \sum_{i=1}^m \log \left( \frac{1}{\sqrt{2\pi}\sigma^2} \exp\left(-\frac{\|y^{(i)} - \hat{y}^{(i)}\|^2}{2\sigma^2}\right) \right) \quad (13)$$

$$= \sum_{i=1}^m \log \left( (2\pi\sigma^2)^{-\frac{1}{2}} \right) + \sum_{i=1}^m \log \left( \exp\left(-\frac{\|y^{(i)} - \hat{y}^{(i)}\|^2}{2\sigma^2}\right) \right) \quad (14)$$

$$= -m \log \sigma - \frac{m}{2} \log 2\pi - \sum_{i=1}^m \frac{\|y^{(i)} - \hat{y}^{(i)}\|^2}{2\sigma^2} \quad (15)$$

- ▶ (15) 수식과 아래 Mean Squared Error를 비교해보라:

$$\text{MSE}_{\text{train}} = \frac{1}{m} \sum_{i=1}^m \|\hat{y}^{(i)} - y^{(i)}\|^2 \quad (16)$$

## 5.5.2 Properties of Maximum Likelihood

- ▶ 이론적으로 샘플 크기( $m$ )이 무한대로 커질수록 MLE가 점근적으로 가장 최적의 추정량임. 즉, 아래 조건들이 만족될 때 5.4.5절의 “consistency” 성질을 갖게 됨.
  - ▶ 실제 분포인  $p_{\text{data}}$ 가 모델  $p(\cdot; \theta)$  내에 들어 있어야 함.
  - ▶  $p_{\text{data}}$ 가 정확히 하나의  $\theta$ 에 대응이 되어야 함.
- ▶ MLE외에도 consistency 성질을 만족시키는 다른 추정 원리들이 있으나(ex. Maximum spacing estimator), 통계적인 효율성 (statistical efficiency)면에서 차이가 있음(ex. 낮은 일반화 오류를 얻는 수준에 이르기까지 필요한  $m$ 의 크기 등).
  - ▶ “parametric case” 연구 분야에서는 어떤 consistent estimator들도 MLE를 능가하지 못한다(= $m$ 이 충분히 크면 MLE가 제일 작은 MSE를 갖게 하는 추정원리다)는 연구 결과가 있음.