5.5 Maximum Likelihood Estimation

- "추정량(estimator)들은 어디서 유래하는가?"
- 각각의 모델들에 대해 어떤 좋은 (분포)함수를 유도하는 일반적인 (추정) 원리가 있는가?
- (알 수는 없지만) 어떤 $P_{\text{data}}(x)$ 확률분포를 따르는 데이터 공간에서 독립적으로 추출한 m 개의 샘플 데이터가 있다고 치자.

$$\mathbb{X} = \{\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(m)}\}$$

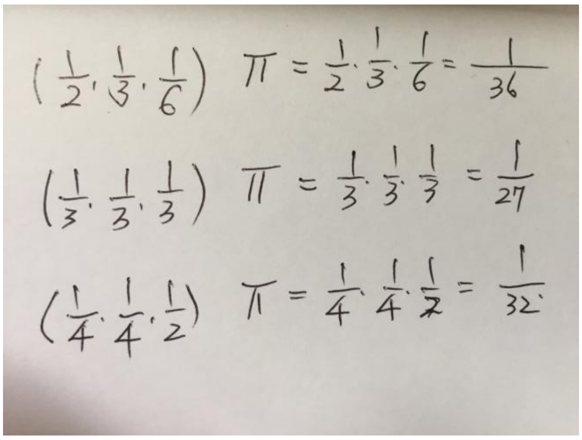
- 위의 X 분포를 설명하는 여러 확률 분포를 가정.

$$p_{\text{model}}(\mathbf{x}; \boldsymbol{\theta})$$

- MLE 는 아래를 만족시키는, θ로 구성된 확률분포를 말함.

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} p_{\text{model}}(\mathbb{X}; \boldsymbol{\theta}) \tag{5.56}$$

$$= \arg\max_{\boldsymbol{\theta}} \prod_{i=1}^{m} p_{\text{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \tag{5.57}$$



- (5.57)은 수치해석적으로 underflow 가능성이 있으므로 아래와 같이 바꿔서 표현하자.

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log p_{\text{model}}(\boldsymbol{x}^{(i)}; \boldsymbol{\theta}). \tag{5.58}$$

- 기대값으로 바꿔 표현

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \mathbb{E}_{\mathbf{x} \sim \hat{p}_{\text{data}}} \log p_{\text{model}}(\boldsymbol{x}; \boldsymbol{\theta}). \tag{5.59}$$

- KL-divergence 관점에서 생각하면, 관측된 분포인 $\hat{p}_{\text{data}}$는 상수이므로, (5.59)를 최대화한다는 것은 아래 (5.60), (5.61)을 최소화시킨다는 뜻이 된다. 즉, MLE는 D_{KL}을 최소화시키면서 모델분포가 관측 분포와 최대한 근접하게 만드는 것이라는 의미가 된다.

$$D_{\text{KL}}\left(\hat{p}_{\text{data}} \| p_{\text{model}}\right) = \mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}}\left[\log \hat{p}_{\text{data}}(\boldsymbol{x}) - \log p_{\text{model}}(\boldsymbol{x})\right]. \tag{5.60}$$

$$-\mathbb{E}_{\mathbf{x}\sim\hat{p}_{\text{data}}}\left[\log p_{\text{model}}(\boldsymbol{x})\right] \tag{5.61}$$

- (5.61)을 negative log likelihood (NLL)이라고도 한다.
- MLE는 NLL 또는 관측된 분포와 모델 분포간의 cross-entropy를 최소화하는 것...

5.5.1 Conditional Log-Likelihood and Mean Square Error

- MLE를 supervised learning에 이용할 수 있음. (X: 입력, Y: 출력)이라 할 때,

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} P(\boldsymbol{Y} \mid \boldsymbol{X}; \boldsymbol{\theta}). \tag{5.62}$$

- 훈련데이터가 i.i.d(independent and identically distributed)라면, 아래와 같이 쓸 수 있음.

$$\boldsymbol{\theta}_{\text{ML}} = \arg\max_{\boldsymbol{\theta}} \sum_{i=1}^{m} \log P(\boldsymbol{y}^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}). \tag{5.63}$$

- 선형회귀와 MLE
  - 중심극한정리(central limit theorem)를 이용해서 아래와 같이 정의($\hat{y}(\boldsymbol{x};\boldsymbol{w})$:평균 예측모델):

$$p(y \mid \boldsymbol{x}) = \mathcal{N}(y; \hat{y}(\boldsymbol{x};\boldsymbol{w}), \sigma^2)$$

  - (5.63)을 위의 $p(y|x)$ 정의에 맞춰 전개하면 아래와 같다.(3p. 전개 참고)

$$\sum_{i=1}^{m} \log p(y^{(i)} \mid \boldsymbol{x}^{(i)}; \boldsymbol{\theta}) \tag{5.64}$$

$$= -m \log \sigma - \frac{m}{2} \log(2\pi) - \sum_{i=1}^{m} \frac{\left\|\hat{y}^{(i)} - y^{(i)}\right\|^2}{2\sigma^2}, \tag{5.65}$$

  - 선형회귀의 $\text{MSE}_{\text{train}}$는 아래와 같다.

$$\text{MSE}_{\text{train}} = \frac{1}{m} \sum_{i=1}^{m} ||\hat{y}^{(i)} - y^{(i)}||^2, \tag{5.66}$$

5.5.2 MLE 의 (좋은) 성질들

- 이론적으로 샘플 크기(m)이 무한대로 커질수록 MLE가 점근적으로 가장 최선의 추정량임. 즉, consistency 성질(5.4.5)을 갖는데, 아래 조건들이 만족되어야 함.
  - 실제 분포인 p_{data}가 모델 $p(\bullet ;\theta)$ 내에 들어 있어야 함.
  - p_{data} 가 정확히 하나의 θ에 대응이 되어야 함.
- MLE 외에도 consistency 성질을 만족시키는 다른 추정 원리들이 있으나, 통계적인 효율성(statistic efficiency)면에서 차이들이 있음. (ex. 낮은 일반화 오류를 얻기 위한 m의 수가 차이 나는 것 등).
  - "parametric case" 연구 분야에서는, 어떤 consistent estimator들도 MLE를 능가하지 못함.(m이 충분히 크면 MLE가 제일 작은 MSE를 갖게 하는 추정원리임).

$$N(x;\mu,\sigma^2) \Rightarrow p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$N(y;\hat{y},\sigma^2) \Rightarrow p(\hat{y}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|\hat{y}-y\|^2}{2\sigma^2}\right)$$

$$L = \prod_i^m p(\hat{y}^{(i)}) = \prod_i^m \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{\|\hat{y}^{(i)}-y^{(i)}\|^2}{2\sigma^2}\right)$$

$$= \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^m \cdot \exp\left(\sum_i^m -\frac{\|\hat{y}^{(i)}-y^{(i)}\|^2}{2\sigma^2}\right)$$

$$= (2\pi\sigma^2)^{-\frac{m}{2}} \cdot \exp\left(\quad '' \quad\right)$$