DSA5202 Lec 8 → Uncertainty Qualification

→ Confidence for model prediction ?

important in sensitive domain { medical field, finance }

Today's Lecture:

1. Model inference (Problem formulation)

→ We assume that our groud-truth (x, y) comes from:

$$\underline{y = f_{\theta^*}(x)} \qquad f_{\theta^*}(\cdot) \rightarrow \text{oracle function}$$

$$\Rightarrow \mathcal{D} = \{(x_i, y_i)\}_{i=1}^N = \{(x_i, f_{\theta^*}(x_i))\}_{i=1}^N$$

→ our estimator of θ^* is $\hat{\theta}$. comes from:

ERM ←
$$\boxed{\hat{\theta} = \underset{\theta \in \Theta}{\arg\min} \frac{1}{N} \sum_{i=1}^{N} \ell(f_\theta(x_i), \underbrace{f_{\theta^*}(x_i)}_{y_i})}$$

Hypothesis Space $\mathcal{H} = \{ f : f(x) = f_\theta(x), \theta \in \Theta \}$

→ $\boxed{\text{parametric model}}$ like NN

→ We solve the ERM approximately via { GD, SGD, mGD }

⇓ optimization perspective

⇓ achieve $\hat{\theta} \approx \theta_n$

Question: when we achieve θ_n $(\hat{\theta})$, we might be interested:

① the distribution of $\hat{\theta}$ / $g(\hat{\theta})$

② the distribution of our prediction $f_{\hat{\theta}}(x)$

☆ Note: Randomness comes from: Dataset $(x_i, y_i) \sim \mathcal{D}$ → z_i

→ $\hat{\theta} = g(z_1, \dots, z_n)$

Today's content

θ^* is constant

Solution: { Frequentist approach: confidence interval for $\boxed{g(\theta^*)}$

Bayesian approach: → $\boxed{\text{conditional distribution}}$

our interest is: $g(\theta) = f_\theta(x)$

2. Definition of Quantile q_α for R.V. z

$\Longrightarrow$ $q_\alpha \longrightarrow \underbrace{P(z \le q_\alpha)}_{\boxed{\text{cdf}}} = \alpha$

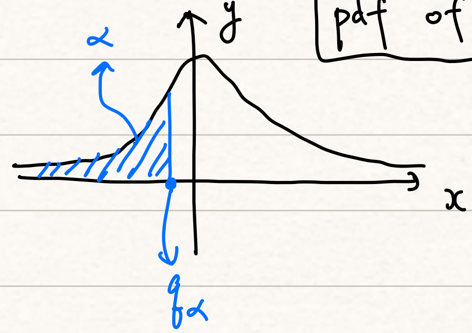
(plug-in estimator) estimator of q_α.

$$\frac{\sum_{i=1}^{n} \mathbb{1}\{z_i \le \hat{q}_\alpha\}}{n} = \alpha$$

$\boxed{\text{ECDF}}$

[e.g.]. $z \sim \mathcal{N}(0,1)$

$\boxed{\text{pdf of } z}$



→ plug-in estimator: use ECDF to replace CDF

CDF: $q_\alpha :\Rightarrow F(q_\alpha) = \alpha \longrightarrow$ ECDF: $\hat{q}_\alpha :\Rightarrow \hat{F}(\hat{q}_\alpha) = \alpha$

$$F(x) = P(X \le x) \longrightarrow \hat{F}_n(x) = \frac{\sum_{i=1}^{n} \mathbb{1}\{X_i \le x\}}{n}$$

$\boxed{X_i \xrightarrow{\text{i.i.d}} F}$

Definition of Confidence Interval (CI) C_α

$\Longrightarrow$ $\theta \to$ population parameter for $X \sim F(\cdot)$

$\boxed{X_i \sim F(\cdot)}$ i.i.d sample from population

$$C_\alpha \longrightarrow \mathbb{P}(\theta \in C_\alpha(X_1, \dots, X_n)) = \alpha$$

$$\Leftrightarrow \mathbb{P}(L_\alpha(X_1, \dots, X_n) \le \theta \le U_\alpha(X_1, \dots, X_n)) = \alpha$$

☆ **Note:** In statistical inference, the CI measures that:

$\Rightarrow$ if we repeat the estimation for B times, that is:

$X_{i1}, \dots, X_{in} \sim F(\cdot) \quad i = 1, 2, 3, \dots, B$

$\downarrow$

Confidence Interval $\boxed{CI_\alpha^i}$ $[L_\alpha(X_{i1}, \dots, X_{in}), U_\alpha(X_{i1}, \dots, X_{in})] \quad i \in [B]$

$$\Rightarrow \boxed{\frac{\sum_{i=1}^{B} \mathbb{1}\{\theta \in CI_\alpha^i\}}{B} \xrightarrow[a.s.]{P} \mathbb{P}(\theta \in CI_\alpha) = \alpha}$$

3. Frequentist Approach (Confidence Interval)

$\longrightarrow$ $\boxed{\text{idea}}$: for our model $f_{\hat{\theta}}(\cdot)$ and x_{new}, we want an interval prediction CI such that $\mathbb{P}(f_{\theta^*}(x_{new}) \in CI) \geq 1 - \alpha$

$\downarrow$

$\boxed{\alpha \to \text{confidence level}}$

Assmption: 1. $g(\hat{\theta}) \xrightarrow{\text{unbiased estimator}} g(\theta^*) \Leftrightarrow \mathbb{E}[g(\hat{\theta})] = g(\theta^*)$

2. $g(\hat{\theta}) \Longrightarrow$ Gaussian Distributed R.V.

$\Vert$

$\Downarrow$

$$g(\hat{\theta}) \sim \text{Gaussian}\left(\mathbb{E}[g(\hat{\theta})],\ Var[g(\hat{\theta})]\right)$$
$$= \text{Gaussian}\left(g(\theta^*),\ \mathbb{E}[\{g(\hat{\theta}) - g(\theta^*)\}^2]\right)$$

Therefore, we just need to figure out $\sigma^2 = \mathbb{E}[\{g(\hat{\theta}) - g(\theta^*)\}^2]$

$\downarrow$

Then, 95% CI of $\boxed{g(\theta^*)}$ is :

$$[\, g(\hat{\theta}) - \underbrace{\boxed{1.96}}_{z_{\frac{\alpha}{2}}}\sigma,\ g(\hat{\theta}) + \underbrace{\boxed{1.96}}_{z_{\frac{\alpha}{2}}}\sigma \,]$$

Remark : we have $\mathbb{P}(\underbrace{g(\theta^*)}_{\text{true value}} \in [g(\hat{\theta}) - 1.96\sigma,\ g(\hat{\theta}) + 1.96\sigma]) = 0.95$

Issue : How to achieve σ ? $\boxed{\sigma^2 = Var[g(\hat{\theta})]}$

$\downarrow$

$\boxed{\text{standard error of estimator } g(\hat{\theta})}$ $\rightarrow$ $\boxed{\text{Sampling distribution}}$

Solution :
1. exact method for sampling distribution $g(\hat{\theta})$
2. Bootstrap for sampling distribution $g(\hat{\theta})$
3. exact method for standard error $Var[g(\hat{\theta})]$
4. Bootstrap for standard error $Var[g(\hat{\theta})]$

using asymptotic distribution of $g(\hat{\theta})$ $\rightarrow$ $\boxed{g(\hat{\theta}) \sim N(\mathbb{E}[g(\hat{\theta})],\ Var[g(\hat{\theta})])}$

4. Standard Error $Var[g(\hat{\theta})]$ [Exact Method]

$\longrightarrow$ Here, we focus on $g(\hat{\theta}) = f_{\hat{\theta}}(x_{new})$

① Linear (Regression) case

consider $g(\hat{\theta}) = a^T \hat{\theta}$

$\longrightarrow \text{var}[g(\hat{\theta})] = \text{var}[a^T \hat{\theta}]$

Denote $\Sigma_\theta = \text{cov}[\hat{\theta}]$

$= a^T \text{cov}[\hat{\theta}] \cdot a$

$= a^T \Sigma_{\hat{\theta}} \cdot a$

✲

$\boxed{\text{Assumption}}$ in Linear Regression :

$y_i = f^*(x_i) + \boxed{\varepsilon_i} \longrightarrow$ all the randomness comes from this term

x_i is fixed & non-random

$\longrightarrow$ in Linear Regression : $f^*(x) = x^T \beta^*$

$y_i = f^*(x_i) + \varepsilon_i$

$f_{\hat{\beta}}(x) = x^T \hat{\beta} \longrightarrow$ if we have $\text{cov}[\hat{\beta}]$, then we are done!

$\boxed{\text{cov}[\hat{\beta}]} = \text{cov}[(X^TX)^{-1}X^T y]$ $\quad \begin{cases} y = X\beta^* + \varepsilon \\ E[y] = X\beta^* \end{cases}$

measure the uncertainty with respect to current dataset

$= \text{cov}[(X^TX)^{-1}X^T \varepsilon]$

$= \sigma^2 (X^TX)^{-1}$

$\Longrightarrow \text{Var}[f_{\hat{\beta}}(x)] = \sigma^2 x^T (X^TX)^{-1} x$ $\longrightarrow \hat{\sigma}^2 = \frac{SSE}{n-p+1}$

$\hookrightarrow$ exact variance for $f_{\hat{\beta}}(x)$

5. Standard Error $Var[g(\hat{\theta})]$ [Bootstrap]

↳ for general function $g(\cdot)$, not just linear case

issue : ① do not know $\hat{\theta} \sim ?$ → no access to $\Sigma_{\hat{\theta}}$

② even if we know $\Sigma_{\hat{\theta}}$, we cannot infer

$$Var[g(\hat{\theta})] \longleftrightarrow \Sigma_{\hat{\theta}}$$

due to the non-linearity of $g(\cdot)$ w.r.t $\hat{\theta}$

☆☆☆

Surrogate : approximate $Var[g(\hat{\theta})]$

Pipeline : $S = \{(x_i, y_i)\}_{i=1}^{N}$

draw with replacement S_b

1. draw n_b samples from S, $\boxed{b = 1, 2, \dots, B}$
2. Train model to achieve $\hat{\theta}$ from S_b

$\Downarrow$

get model $\hat{f}_{S^b}(x)$ $\boxed{b = 1, 2, \dots, B}$

3. $Var[g(\hat{\theta})] \approx \frac{1}{B} \sum_{b=1}^{B} \left\{ \hat{f}_{S^b}(x) - \overline{\hat{f}(x)} \right\}^2$

$g(\theta) = f_{\hat{\theta}}(x) := f_S(x)$

since $\hat{\theta}$ is attained via dataset S

bootstrap

$$\boxed{\overline{\hat{f}(x)} = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_{S^b}(x)}$$

6. Fisher Information for MLE asymptotic distribution

Problem Setting: $X_i \sim P_\theta \quad i = 1, 2, \dots, n.$

$$\Rightarrow \text{likelihood} \quad \mathcal{L}(\theta : \mathcal{D}) = \prod_{i=1}^{n} P(X_i \mid \theta)$$

$$\Rightarrow \text{log-likelihood} \quad \ell(\theta) = \sum_{i=1}^{n} \log P(X_i \mid \theta) := \sum_{i} \ell_i(\theta)$$

$$\Rightarrow \text{MLE estimator} \quad \hat{\theta}_{MLE} := \underset{\theta \in \Theta}{\operatorname{argmax}}\ \ell(\theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \sum_{i=1}^{n} \log P(X_i \mid \theta)$$

Notation

① Scoring Function:

$$S(\theta) = \nabla_\theta \ell(\theta) = \sum_{i=1}^{n} S_i(\theta) := \sum_{i=1}^{n} \nabla_\theta \ell_i(\theta)$$

② Fisher Information:

$$I(\theta) = \mathbb{E}[S(\theta) \cdot S(\theta)^T]$$

(independence between X_i)

$$= n\, \mathbb{E}[S_i(\theta) \cdot S_i(\theta)^T]$$

$$= n\, \mathbb{E}_X[\nabla_\theta \cdot \log P(X \mid \theta) \cdot \nabla_\theta \log P(X \mid \theta)^T]$$

$$\boxed{X \sim P(\cdot \mid \theta)}$$

Result 1: under some regularity condition,

$$I(\theta) = \mathbb{E}_{X\sim\theta}[\, S(\theta) \cdot S(\theta)^T \,]$$

$$= \mathbb{E}_{X\sim\theta}[\, \nabla \ell(\theta) \cdot \nabla \ell(\theta)^T \,]$$

$$= - \mathbb{E}_{X\sim\theta}[\, \nabla^2 \ell(\theta) \,]$$

Result 2: Asymptotic distribution for $\hat{\theta}_{MLE}$

$$\hat{\theta}_{MLE} \sim N(\theta^*, I(\theta^*)^{-1})$$

$\boxed{\text{asymptotic normal}}$

In practice, we use $\hat{\theta}_{MLE} - \theta^* \approx N(0, I(\hat{\theta}_{MLE})^{-1})$

In Summary:

$\boxed{\text{Uncertainty Qualification}}$ → confidence interval for ground-truth

$\boxed{\text{via model prediction}}$

From Frequentist Perspective: one general approach is:

$\boxed{\text{Bootstrap}}$ → simulate the randomness via { empirical cdf / model