DSA5202 Lec 4 → Bayesian Approach

Recap: → Uncertainty Quantification

↓

Machine Learning Problem
- ① Dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$
- ② Model $\mathcal{H}$ → hypothesis space
- ③ Traning Algo → approximate $\hat{\theta}$

$\theta^* \approx \hat{\theta}$ → Random Variable w.r.t dataset $\mathcal{D}$

θ^*: Ground-truth

$f_{\hat{\theta}}(x)$: model prediction on new data

Frequentist Perspective → has the intrinsic Randomness due to the randomness of dataset $\mathcal{D}$

Core

Example: ① $\mathbb{E}[X] = \theta^*$

$\hat{\theta} = \frac{1}{n}\sum_{i=1}^n X_i$

asymptotic

⇒ from CLT, we have $\hat{\theta} \sim \text{Gaussian}(\theta^*, \frac{1}{n}\text{Var}[X])$

→ uncertainty quantification

$\Rightarrow \mathbb{P}\left(\theta^* \in \left[\text{std}\left(\hat{\theta} - q_{1-\frac{\alpha}{2}}\right), \text{std}\left(\hat{\theta} - q_{\frac{\alpha}{2}}\right)\right]\right) = 1-\alpha$

→ Here, randomness comes from $\mathcal{D} = \{X_i\}_{i=1}^n$

Example: ② $Y = X\beta^* + \varepsilon$ $\quad f_{\hat{\beta}}(x) = x^T\hat{\beta}$ (Linear Reg)

Here, $\hat{\beta} = (X^TX)^{-1}X^TY$ → randomness comes from Y (ε)

$\hat{\beta} \sim \text{Gaussian}(\beta^*, \sigma^2(X^TX)^{-1})$

$\Rightarrow f_{\hat{\beta}}(x) = x^T\hat{\beta} \sim \text{Gaussian}(x^T\beta^*, \sigma^2 x^T(X^TX)^{-1}x)$

$\Rightarrow \mathbb{P}(f^*(x) = x^T\beta^* \in [\cdots, \cdots]) = 1-\alpha$

→ uncertainty quantification

Example : ③ → Bootstrap

we have $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ + Hypothesis Space $\mathcal{H}$

optimal hyper-parameter $\hat{\theta} \approx \theta_n$

model prediction $f_{\hat{\theta}}(x)$

Basically, $f_{\hat{\theta}}(x)$ can be viewed as the random variable of $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$

use Bootstrap to estimate { variance → confidence interval

draw with replacement

$\{(X_{i1}, Y_{i1}), \ldots, (X_{in}, Y_{in})\}_{i=1}^B$

↳ $\{f_{\hat{\theta}_i}(x)\}_{i=1}^B$

☆ this can also construct CI under normal assumption

⇒ ① variance $\mathrm{Var}[f_{\hat{\theta}}(x)] \approx \sum_{i=1}^B \left(f_{\hat{\theta}_i}(x) - \overline{f_{\hat{\theta}}(x)}\right)^2$

$\overline{f_{\hat{\theta}}(x)} = \sum_{i=1}^B f_{\hat{\theta}_i}(x)$

⇒ ② confidence interval → $\{f_{\hat{\theta}_i}(x) - f_{\hat{\theta}}(x)\}_{i=1}^B$

⇓

$q_{\frac{\alpha}{2}} = f_{\hat{\theta}_{\frac{\alpha}{2}}} - f_{\hat{\theta}}$ quantile

$q_{1-\frac{\alpha}{2}} = f_{\hat{\theta}_{1-\frac{\alpha}{2}}} - f_{\hat{\theta}}$

$\mathbb{P}(f_{\hat{\theta}}(x) - f_{\theta^*}(x) \le t) \approx \frac{1}{B}\sum_{i=1}^B \mathbb{1}\{f_{\hat{\theta}_i}(x) - f_{\hat{\theta}}(x) \le t\}$

$\Rightarrow \mathbb{P}\left(f_{\theta^*}(x) \in \left[2f_{\hat{\theta}}(x) - f_{\hat{\theta}_{1-\frac{\alpha}{2}}}(x),\ 2f_{\hat{\theta}}(x) - f_{\hat{\theta}_{\frac{\alpha}{2}}}(x)\right]\right) = 1 - \alpha$

Bootstrap quantile

Bootstrap quantile

To conclude: from Frequentist Perspective:

① all randomness comes from the data $\mathcal{D}$ itself!

② Ground-truth parameters are fixed!

Today's lecture: Bayesian Approach → β is random!

1. Bayes Theorem

likelihood

prior distribution

$$P(\beta \mid \mathcal{D}) = \frac{P(\mathcal{D} \mid \beta) \cdot P(\beta)}{P(\mathcal{D})}$$

$P(\mathcal{D})$ → evidence for this model

posterior distribution

$$= \frac{P(\mathcal{D} \mid \beta)\, P(\beta)}{\int_{\beta} P(\mathcal{D} \mid \beta) \cdot P(\beta)\, d\beta}$$

$$\propto P(\mathcal{D} \mid \beta) \cdot P(\beta)$$

↳ our interest to determine the posterior distribution $P(\beta \mid \mathcal{D})$

Advantage: ← encode knowledge to distribution

① easy to combine prior knowledge

② robust against over-fitting ← prior distribution

③ give the whole distribution of β ↕ regularization

$p(\beta|\mathcal{D})$ $\Rightarrow$ easy to construct CI

Disadvantages:

① Expensive to achieve closed-form sol^n

↳ need approximation via MCMC

② sensitive to prior choices

2. Example: Bayesian Linear Regression

Setting: $\begin{cases} \beta \sim N(\mu_0, \Sigma_0) \\ y|X,\beta \sim N(X\beta, \Sigma_\varepsilon) \end{cases}$

① likelihood: $p(\mathcal{D}|\beta) = p(y|X,\beta)$

$\propto \exp\{-(y-X\beta)^T \Sigma_\varepsilon^{-1} (y-X\beta)\}$

② prior: $p(\beta) = \exp\{-(\beta-\mu_0)^T \Sigma_0^{-1} (\beta-\mu_0)\}$

$\Rightarrow$ ③ posterior: $p(\beta|\mathcal{D}) \propto p(\mathcal{D}|\beta) \cdot p(\beta)$

$= \exp\{-(y-X\beta)^T \Sigma_\varepsilon^{-1} (y-X\beta)\} \times$

$\exp\{-(\beta-\mu_0)^T \Sigma_0^{-1} (\beta-\mu_0)\}$

Simplification: $\exp\{-(y-X\beta)^T\Sigma_\varepsilon^{-1}(y-X\beta)-(\beta-\mu_0)^T\Sigma_0^{-1}(\beta-\mu_0)\}$

$$\propto \exp\{-\beta^T X^T\Sigma_\varepsilon^{-1}X\beta-\beta^T\Sigma_0^{-1}\beta$$

$$+2y^T\Sigma_\varepsilon^{-1}X\beta+2\mu_0^T\Sigma_0^{-1}\beta\}$$

$$=\exp\{-\beta^T(X^T\Sigma_\varepsilon^{-1}X+\Sigma_0^{-1})\beta+2(y^T\Sigma_\varepsilon^{-1}X+\mu_0^T\Sigma_0^{-1})\beta\}$$

if $X\sim N(\mu,\Sigma)$, $p(x)\propto\exp\{-x^T\Sigma^{-1}x+2\mu^T\Sigma^{-1}x\}$

$$\Rightarrow\begin{cases}\Sigma^{-1}=(X^T\Sigma_\varepsilon^{-1}X+\Sigma_0^{-1})\\ \mu^T\Sigma^{-1}=y^T\Sigma_\varepsilon^{-1}X+\mu_0^T\Sigma_0^{-1}\end{cases}$$

$$\Rightarrow\begin{cases}\Sigma=(X^T\Sigma_\varepsilon^{-1}X+\Sigma_0^{-1})^{-1}\\ \mu=(X^T\Sigma_\varepsilon^{-1}X+\Sigma_0^{-1})^{-1}(X^T\Sigma_\varepsilon^{-1}y+\Sigma_0^{-1}\mu_0)\end{cases}$$

Therefore, we can conclude that:

for Bayesian LR, under the setting that:

$$\begin{cases}\beta\sim N(\mu_0,\Sigma_0)\\ y|X,\beta\sim N(X\beta,\Sigma_\varepsilon)\end{cases}$$ (special case)

posterior dist. is exact Gaussian for Bayesian LR

we have: $\beta|\mathcal{D}\sim N(\mu_p,\Sigma_p)$

Here, $\begin{cases}\mu_p=(X^T\Sigma_\varepsilon^{-1}X+\Sigma_0^{-1})^{-1}(X^T\Sigma_\varepsilon^{-1}y+\Sigma_0^{-1}\mu_0)\\ \Sigma_p=(X^T\Sigma_\varepsilon^{-1}X+\Sigma_0^{-1})^{-1}\end{cases}$

Issue: ① Generally speaking, we cannot attain the exact solution for posterior distribution → disadvantage 1

① If our choice our β (prior distribution) is far from reality, then the achieved $\beta|\mathcal{D}$ (posterior distribution) will be twisted.

↘ disadvantage 2 sensitive to prior choices

3. Monte Carlo Simulation Outline

Recap: In Bayesian LR, based on $\begin{cases} \beta \sim N(\mu_0, \Sigma_0) \\ y|X,\beta \sim N(X\beta, \Sigma_L) \end{cases}$

we can achieve $\beta|\mathcal{D} \sim N(\mu_p, \Sigma_p)$

↓

exact posterior distribution

⟶ ① we can attain exact $\begin{cases} \text{expectation} & \mu_p \\ \text{variance} & \Sigma_p \end{cases}$

(if our interest is $\begin{cases} \mathbb{E}[\beta] \\ Var[\beta] \end{cases}$ (posterior))

⟶ ② we can sample from it easily

★ sometimes even if we have $g(\beta)$, it is expensive to do integral

(if our interest is $\begin{cases} \mathbb{E}[g(\beta)] \\ Var[g(\beta)] \end{cases}$ (posterior), now we have no access to the posterior distribution for $g(\beta)$ explicitly)

$\Longrightarrow$ $\beta_1, \dots, \beta_n \sim \beta|\mathcal{D}$ (since $\beta|\mathcal{D}$ is good enough) → exact Gaussian

⇓

approximate $\begin{cases} \mathbb{E}[g(\beta)] \\ Var[g(\beta)] \end{cases}$ via Monte Carlo

→ ③ Sometimes it is difficult to sample from $\beta | \mathcal{D}$ directly ⟹ use some Markov Chain Monte Carlo

⇓

sample from $\beta | \mathcal{D}$ approximately

[E.g.] Metropolis-Hastings Algo [M-H Algo]

Gibbs Sampling

4. Monte Carlo Approach

↓

Our interest is like $\begin{cases} \mathbb{E}[f(\beta) | \mathcal{D}] \\ \text{Var}[f(\beta) | \mathcal{D}] \end{cases}$

Assume: ① it is intractable to compute $\int_\beta f(\beta) \cdot p(\beta | \mathcal{D}) d\beta$ directly

② We can sample from $\beta | \mathcal{D}$ directly

Solution: $\mathbb{E}[f(\beta) | \mathcal{D}] = \int_\beta f(\beta) p(\beta | \mathcal{D}) d\beta$

e.g. $f(\beta) = h_\beta(x) \Rightarrow h_\beta(x)$: neural network prediction

$\mathbb{E}_{\beta|\mathcal{D}}[f(\beta)] = \mathbb{E}_{\beta|\mathcal{D}}[h_\beta(x)]$

no closed-form solution

⇒ approximate through $\mathbb{E}_\beta[f(\beta) | \mathcal{D}]$

$\approx \hat{f}_m = \frac{1}{m} \sum_{i=1}^{m} f(\beta^i)$ $\quad \beta^i \sim p(\beta | \mathcal{D})$

① From LLN, we have $\hat{f}_m \xrightarrow[a.s.]{P} \mathbb{E}_\beta[f(\beta)|\mathcal{D}]$ $m \to \infty$

② if $\beta^i \sim p(\beta|\mathcal{D})$ i.i.d. then $\text{var}[\hat{f}_m] = \frac{1}{m}\text{var}_\beta[f(\beta)|\mathcal{D}]$

From CLT, we have:

$$\hat{f}_m - \mathbb{E}_\beta[f(\beta)|\mathcal{D}] \approx \mathcal{N}\left(0, \frac{1}{m}\text{var}_\beta[f(\beta)|\mathcal{D}]\right)$$

$\frac{1}{\sqrt{m}}$ convergence rate

Remark: for MCMC method, generally we do not have this property, which means the variance will be slightly bigger.

From autocorrelation to ESS

Idea: use Autocorrelation to calculate this term!

$$\text{Var}\left\{\frac{1}{T}\sum_{t=1}^{T} f(\theta^{(t)})\right\} = \frac{1}{T^2}\sum_{t=1}^{T}\left(\text{Var}\{f(\theta^{(t)})\} + \sum_{s\neq t}\text{Cov}\{f(\theta^{(t)}), f(\theta^{(s)})\}\right)$$

if independent, then ESS = T!

$$= \frac{\text{Var}\{f(\theta^{(0)})\}}{T^2}\sum_{t=1}^{T}\left(1 + \sum_{s\neq t}\text{Corr}\{f(\theta^{(t)}), f(\theta^{(s)})\}\right)$$

$$= \frac{\text{Var}\{f(\theta^{(0)})\}}{T^2}\sum_{t=1}^{T}\left(1 + \sum_{s\neq t}\rho_{s-t}\right)$$

$$= \frac{\text{Var}\{f(\theta^{(0)})\}}{T^2}\sum_{t=1}^{T}\left(1 + \sum_{s\neq t}\rho_{|s-t|}\right)$$

$$\approx \frac{\text{Var}\{f(\theta^{(0)})\}}{T^2}\sum_{t=1}^{T}\left(1 + 2\sum_{L=1}^{\infty}\rho_L\right)$$

$$= \frac{\text{Var}\{f(\theta^{(0)})\}}{T(1+2\sum_{L=1}^{\infty}\rho_L)^{-1}} \quad \text{ESS}$$

can estimate with plug-in estimators

Remark: when dimensionality is high, we prefer Frequentist Approach

↳ $\beta \in \mathbb{R}^p$ and p is large

since we need: $\mathbb{E}_\beta[f(\beta)|\mathcal{D}] \approx \frac{1}{m}\sum_{i=1}^{m} f(\beta^i)$ for Bayesian

↳ works well in low-dimension case

similarly: $\text{var}_\beta[f(\beta)|\mathcal{D}] = \int_\beta (f(\beta) - \mathbb{E}_\beta[f(\beta)|\mathcal{D}])^2 p(\beta|\mathcal{D})\,d\beta$

$$\approx \frac{1}{m}\sum_{i=1}^{m}(f(\beta^i) - \hat{f}_m)^2$$

$$\begin{cases} \hat{f}_m = \frac{1}{m}\sum_{i=1}^{m} f(\beta^i) \approx \mathbb{E}_\beta[f(\beta)|\mathcal{D}] \\ \beta^i \sim p(\beta|\mathcal{D}) \end{cases}$$

5. Application of Monte Carlo method

Model: $\begin{cases} \beta \sim N(0, \tau I) \\ y|X,\beta \sim N(H\beta, \sigma I) \end{cases}$ $\quad H := \begin{pmatrix} h(x_1)^T \\ \vdots \\ h(x_n)^T \end{pmatrix}$

$\longrightarrow$ **General Linear Model**

$\Rightarrow \beta|\mathcal{D} \sim N(\mu_p, \Sigma_p)$

where $\begin{cases} \mu_p = (H^TH + \sigma\tau^{-1}I)^{-1}H^Ty \\ \Sigma_p = (\sigma^{-1}H^TH + \tau^{-1}I)^{-1} \end{cases}$ Ridge Regression estimator

$\longrightarrow$ if apply Maximize A Posterior (MAP) framework.

then $\hat{\beta}_{MAP} = (H^TH + \sigma\tau^{-1}I)^{-1}H^Ty$

$\hookrightarrow$ **Ridge Regularization**

$\Rightarrow$ Moreover, we have:

$$p(y|x_{new}, \mathcal{D})$$

$$= \int_w p(y, w|x_{new}, \mathcal{D})\, dw$$

$$= \int_w \underbrace{p(y|w, x_{new})}_{\sim N(w^Th(x_{new}), \beta^{-1})} \cdot \underbrace{p(w|\mathcal{D})}_{\sim N(\mu_p, \Sigma_p)}\, dw$$

$$\Rightarrow p(y|x_{new}, \mathcal{D}) \sim N(\mu_p^T h(x_{new}),\ \beta^{-1} + h(x_{new})^T \Sigma_p h(x_{new}))$$

$(\text{std})^2$

construct "confidence interval"

$$\left[\mu_p^T h(x_{new}) - \text{std},\ \mu_p^T h(x_{new}) + \text{std}\right]$$

Generally when our model is $y = f_w(x)$

and we have $w \mid \mathcal{D} \sim$ some distribution

Issue: ① it is difficult to determine $y \mid x_{new}, \mathcal{D}$ distribution

② cannot compute explicitly $\begin{cases} \mathbb{E}[y \mid x_{new}, \mathcal{D}] \\ \text{Var}[y \mid x_{new}, \mathcal{D}] \end{cases}$

$\Rightarrow$ Monte Carlo to approximate $\begin{cases} \mathbb{E}[y \mid x_{new}, \mathcal{D}] \to \hat{\mu} \\ \text{Var}[y \mid x_{new}, \mathcal{D}] \to \hat{\text{std}} \end{cases}$

then, the approximate "Confidence Interval" is:

$$\left[\hat{\mu} - \hat{\text{std}},\ \hat{\mu} + \hat{\text{std}}\right]$$

two case

① $w \mid \mathcal{D}$ → sample directly

$\Rightarrow$ like Bayesian LR case

$w \mid \mathcal{D} \sim N(\mu_p, \Sigma_p)$

② $w \mid \mathcal{D}$ → cannot sample directly

$\Rightarrow$ use MCMC to sample approximately

Bayesian inference vs frequentist inference

- Two fundamentally different ways of thinking about inference

Frequentist	Bayesian
• Parameter is fixed	• Parameter is random
• Data is random	• Data is fixed
• Randomness represents sampling	• Randomness represents uncertainty

Bayesian perspective allows us to reason about questions like "what is the probability of an earthquake?"