Lecture 8: Uncertainty Quantification (I)

Soufiane Hayou

Model inference

Assume that our data comes from a model f_{θ^*} (the oracle function).

- We build a model $y \sim f_{\hat{\theta}}(x)$
- Given $\hat{\theta}$, what will likely be the range/distribution of $g(\hat{\theta})$?
- Given x, what will likely be the range/distribution of $g(\hat{\theta}) = f_{\hat{\theta}(x)}$?

Model inference

Assume that our data comes from a model f_{θ^*} (the oracle function).

- We build a model $y \sim f_{\hat{\theta}}(x)$
- Given $\hat{\theta}$, what will likely be the range/distribution of $g(\hat{\theta})$?
- Given x, what will likely be the range/distribution of $g(\hat{\theta}) = f_{\hat{\theta}(x)}$?
- Frequentist approach: confidence interval for $g(\hat{\theta})$.

Model inference

Assume that our data comes from a model f_{θ^*} (the oracle function).

- We build a model $y \sim f_{\hat{\theta}}(x)$
- Given $\hat{\theta}$, what will likely be the range/distribution of $g(\hat{\theta})$?
- Given x, what will likely be the range/distribution of $g(\hat{\theta}) = f_{\hat{\theta}(x)}$?
- Frequentist approach: confidence interval for $g(\hat{\theta})$.
- Bayesian approach: conditional distribution for $g(\hat{\theta})$.

Model inference

Assume that our data comes from a model f_{θ^*} (the oracle function).

- We build a model $y \sim f_{\hat{\theta}}(x)$
- Given $\hat{\theta}$, what will likely be the range/distribution of $g(\hat{\theta})$?
- Given x, what will likely be the range/distribution of $g(\hat{\theta}) = f_{\hat{\theta}(x)}$?
- Frequentist approach: confidence interval for $g(\hat{\theta})$.
- Bayesian approach: conditional distribution for $g(\hat{\theta})$.
- Also known as uncertainty quantification
- Source of uncertainty: estimation error of $\hat{\theta}$
- One popular way: find the variance of $g(\hat{\theta})$.

Confidence interval

- Build a region C around $g(\hat{\theta})$

Confidence interval

- Build a region C around $g(\hat{\theta})$
- Confidence interval: $\mathbb{P}(g(\theta) \in C) > 1 - \delta$
- The probability average out all possible data outcomes

Confidence interval

- Build a region C around $g(\hat{\theta})$
- Confidence interval: $\mathbb{P}(g(\theta) \in C) > 1 - \delta$
- The probability average out all possible data outcomes
- Suppose $g(\hat{\theta})$ is an unbiased estimator of $g(\theta)$
- Suppose $g(\hat{\theta})$ is Gaussian distributed

Confidence interval

- Build a region C around $g(\hat{\theta})$
- Confidence interval: $\mathbb{P}(g(\theta) \in C) > 1 - \delta$
- The probability average out all possible data outcomes
- Suppose $g(\hat{\theta})$ is an unbiased estimator of $g(\theta)$
- Suppose $g(\hat{\theta})$ is Gaussian distributed
- Find estimation variance $\sigma^2 = \mathbb{E}|g(\theta) - g(\hat{\theta})|^2$
- 95% confidence interval $[g(\hat{\theta}) - 1.96\sigma, g(\hat{\theta}) + 1.96\sigma]$

Confidence interval

- Build a region C around $g(\hat{\theta})$
- Confidence interval: $\mathbb{P}(g(\theta) \in C) > 1 - \delta$
- The probability average out all possible data outcomes
- Suppose $g(\hat{\theta})$ is an unbiased estimator of $g(\theta)$
- Suppose $g(\hat{\theta})$ is Gaussian distributed
- Find estimation variance $\sigma^2 = \mathbb{E}|g(\theta) - g(\hat{\theta})|^2$
- 95% confidence interval $[g(\hat{\theta}) - 1.96\sigma, g(\hat{\theta}) + 1.96\sigma]$
- Question: how do we find σ?
- We will focus on the case $g(\theta) = f_\theta(x)$.

Linear transformation

- Suppose $\mathrm{cov}(\hat{\theta}) = \Sigma_\theta$ (assume we have access to Σ_θ)
- What is variance of $g(\hat{\theta}) = a^T\hat{\theta}$?

- Suppose $\mathrm{cov}(\hat{\theta}) = \Sigma_\theta$ (assume we have access to Σ_θ)
- What is variance of $g(\hat{\theta}) = a^T\hat{\theta}$?
- it's given by $a^T\Sigma_\theta a$.

Linear transformation

- Suppose $\mathrm{cov}(\hat{\theta}) = \Sigma_\theta$ (assume we have access to Σ_θ)
- What is variance of $g(\hat{\theta}) = a^T\hat{\theta}$?
- it's given by $a^T\Sigma_\theta a$.
- In linear regression: $f(x) = x^T\hat{\beta}$

Linear transformation

- Suppose $\text{cov}(\hat{\theta}) = \Sigma_\theta$ (assume we have access to Σ_θ)
- What is variance of $g(\hat{\theta}) = a^T\hat{\theta}$?
- it's given by $a^T\Sigma_\theta a$.
- In linear regression: $f(x) = x^T\hat{\beta}$
- Nonlinear problem: need Monte Carlo.

Bootstrap method

Bootstrap approach

- Draw n_B samples from S randomly with replacement, denote as S^b.
- Repeat this for $b = 1, \ldots, B$.
- For each bootstrap dataset S^b, we apply ML procedure to find $\hat{f}_{S^b}$
- We now have $\hat{f}_{S^1}(x), \ldots, \hat{f}_{S^B}(x)$
- They "simulate" the scenario where data are randomly obtained.
- Their variance is approximately the same as the $\hat{f}_S(x)$

Example: Linear regression

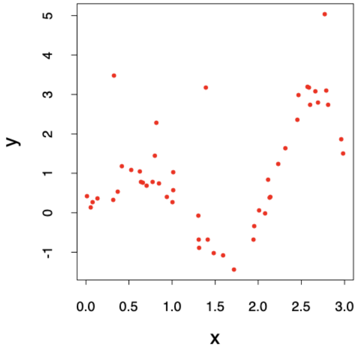
True variance: the case where the only source of randomness is y

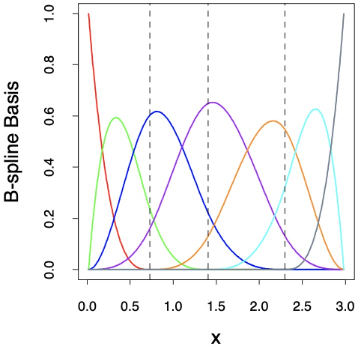
- Recall that $\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \beta^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon$
- Given $\mathbf{X}$, the estimated covariance matrix is $\sigma_\epsilon^2(\mathbf{X}^T\mathbf{X})^{-1}$
- At x_0, the variance of $\hat{f}(x)$ is $\sigma_\epsilon^2 x_0^T(\mathbf{X}^T\mathbf{X})^{-1}x_0$

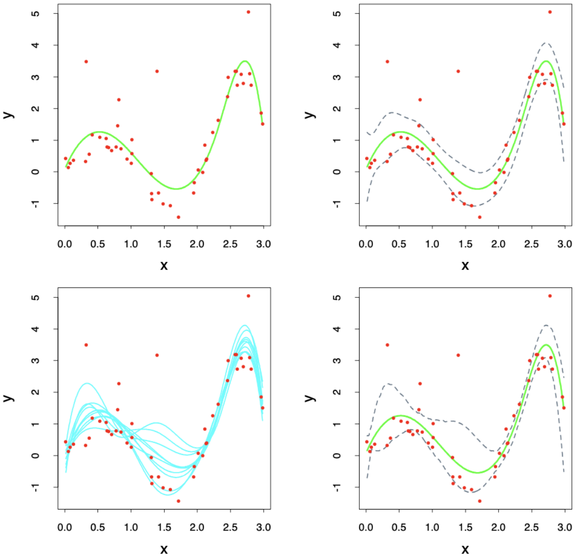
Bootstrap

- $\hat{\beta}_b = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}_b = \beta^* + (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\epsilon_b$
- Estimation for x_0: $x_0^T\hat{\beta}_b$
- The bootstrap covariance comes from randomness in ϵ_b
- The variance is $\sigma_\epsilon^2(\mathbf{X}^T\mathbf{X})^{-1}$.

- A standard way to deal with functional data
- We want to fit $\mu(x) = \beta_1 h_1(x) + \cdots \beta_7 h_7(x) + \epsilon$
- $h_j(x)$: spline functions
- Data fitting: $y_i = \mu(x_i) + \epsilon_i$
- Linear basis regression $y_i = \beta_1 h_1(x_i) + \cdots + \beta_7 h_7(x_i) + \epsilon_i$







y
x
0.0
0.5
1.0
1.5
2.0
2.5
3.0
-1
0
1
2
3
4
5

Using Fisher information

Maximum likelihood

In many applications we model the data from a parameteric family

$$z_i \sim p(\theta, z)$$

The overall likelihood is given by

$$p(\theta, S) = \prod_{i=1}^{n} p(\theta, z_i)$$

We can try to maximize its log $l(\theta, S) = \log(p(\theta, S))$

$$l(\theta, S) = \sum_{i=1}^{n} l(\theta, z_i) = \sum_{i=1}^{n} l(\theta, z_i)$$

The Maximum Likelihood Estimator (MLE) is obtained as

$$\hat{\theta} = \arg\max l(\theta, S).$$

Fisher information

The score function is defined as

$$\nabla_\theta l(\theta, S) = \sum_{i=1}^{n} \nabla_\theta l(\theta, z_i)$$

The Fisher information matrix is its expectation w.r.t z

$$I(\theta) = \mathbb{E}_S[\nabla l(\theta, S) \nabla l(\theta, S)^T] = n\mathbb{E}_z[\nabla l(\theta, z) \nabla l(\theta, z)^T]$$

The MLE estimator converges

$$\hat{\theta} \to \mathcal{N}(\theta^*, I(\theta^*)^{-1})$$

Using the approximation $\hat{\theta} \approx \theta^*$, we have

$$\hat{\theta} - \theta^* \approx \mathcal{N}(0, I(\hat{\theta})^{-1})$$

If we want to have a confidence interval of θ^*_j, we can use

$$\hat{\theta}_j - q^{(1-\alpha)}\sqrt{(I(\hat{\theta})^{-1})_{j,j}}, \quad \hat{\theta}_j - q^{(1-\alpha)}\sqrt{(I(\hat{\theta})^{-1})_{j,j}},$$

where $q^{1-\alpha}$ is the Gaussian quantile of order $1-\alpha$.

Linear regression

- Linear model $y_i = \beta_*^\top x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0, \sigma_\epsilon^2)$.
- For general β, the loglikelihood is given by

$$l(\beta) = -\frac{n}{2}\log(2\pi\sigma_\epsilon^2) - \frac{1}{2\sigma_\epsilon^2}\sum_{i=1}^{n}(y_i - x_i^T\beta)^2$$

- The Fisher information matrix is given by

$$I(\beta_*) = \frac{1}{\sigma_\epsilon^2} X^T X$$

- This agrees with earlier estimates.

Example

Suppose using linear regression, we have $X^T X = nI_d$, $\hat{\sigma^2} = 1$, $\hat{\beta} = [1, 0]$ What will be the confidence interval for $f([1, 1])$?

Answer: $[1 - 1.96 \times \frac{\sqrt{2}}{\sqrt{n}}, 1 + 1.96 \times \frac{\sqrt{2}}{\sqrt{n}}]$