STAT201B Lecture 2 Non-Parametric Inference

1 Plug-in (Substitution) Principle: 一种 non-parametric estimation method

1.1 Plug-in (Substitution) method

问题描述:

- 令 $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$, F 可以为 parametric 或 non-parametric
- quantities of interest 以一种 nonparametric way 与 F 相关 (如 mean, median, variance, quantiles 等), 即无论 F 为 parametric 还是 nonparametric, quantities of interest 都可以写作一个关于 F 的函数 $\theta(F)$

方法:
使用 $\theta(\hat{F}_n)$ 来估计 $\theta(F)$, 其中 $\hat{F}_n$ 是 F 的 empirical distribution

1.2 Empirical CDF

令 $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$,
则 **empirical CDF $\hat{F}_n$** 对每个 datapoint 赋予 $1/n$ 的权重:

$$\hat{F}_n(x) = \frac{\sum_{i=1}^N I(X_i \le x)}{n} = \#\{X_i \le x\}/n$$

Remark

1. $P_{\hat{F}_n}(X \le x) = \hat{F}_n(x)$ 通常与 $P_F(X \le x) = F(x)$ 不同
   e.g. 令 $X_1, X_2, X_3, X_4 \overset{i.i.d.}{\sim} U(0,1)$, 观测值为 $X_1 = 0.3, X_2 = 0.5, X_3 = 0.1, X_4 = 0.7$, 则 $\hat{F}_n(0.5) = 0.75, F_n(0.5) = 0.5$
2. $Y_i = I(X_i \le x), i = 1, \dots, n$ 为 iid Bernoulli r.v. with $p = P(Y_i = 1) = P(X_i \le x) = F(x)$
3. $P_{\hat{F}_n}(X = t) = \frac{\sum_{i=1}^n I(X_i = t)}{n}$

1.3 Plug-in estimators 的例子

1. 若 $\theta(F) = E_F(X)$,
   则 plug-in estimate 为

$$\begin{aligned} \theta(\hat{F}_n) &= \mathbb{E}_{\hat{F}_n}(X) = \textstyle\sum_t t P_{\hat{F}_n}(X = t) \\ &= \textstyle\sum_t t \cdot \frac{\sum_{i=1}^n I(X_i = t)}{n} \\ &= \frac{\sum_{i=1}^n X_i}{n} \\ &= \bar{X}_n \end{aligned}$$

2. 若 $\theta(F) = Var_F(X)$,
   则 plug-in estimate 为

$$\begin{aligned} \theta(\hat{F}_n) = & \; var_{\hat{F}_n}(X) = \mathrm{E}_{\hat{F}_n}(X^2) - (\mathrm{E}_{\hat{F}_n}(X))^2 \\ = & \; \frac{\sum_{i=1}^n X_i^2}{n} - (\frac{\sum_{i=1}^n X_i}{n})^2 \\ = & \; \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n} \end{aligned}$$

3. 若 $\theta(F) = median(X) = inf_t\,\{t | F(t) \ge \frac{1}{2}\}$,
   则 plug-in estimate 为

$$\theta(\hat{F}_n) = inf_t \left\{ t | \hat{F}_n(t) \ge \frac{1}{2} \right\}$$

1.4 Empirical CDF 的性质

- 对任意 fixed x, 有
  - $E[\hat{F}_n(x)] = F(x)$
  - $V[\hat{F}_n(x)] = \frac{F(x)[1-F(x)]}{n}$ (利用 $I(X_i \le x)$ 独立同分布于伯努利分布易证)
  - $MSE[\hat{F}_n(x)] = V[\hat{F}_n(x)] \to 0$
  - $\hat{F}_n(x) \xrightarrow{p} F(x)$
- **Glivenko-Cantelli Theorem:** (提供更加强的 convergence)
  令 $X_1, \ldots, X_n \overset{i.i.d.}{\sim} F$,
  则

$$\sup_x |\hat{F}_n - F(x)| \xrightarrow{a.s.} 0$$

1.5 Plug-in estimator 的性质

令函数 $\theta(F)$ 为 continuous in the sup-norm, 即

$$\forall \epsilon > 0, \exists \delta > 0, \text{such that } ||G - F||_\infty < \delta \implies |\theta(G) - \theta(F)| < \epsilon$$

则

$$\theta(\hat{F}_n) \xrightarrow{p} \theta(F)$$

> **Remark**
>
> 表示当 quantity of interest 满足特定连续条件时, plug-in estimator 是 consistent estimator

2 Linear Functional 的 Plug-in Estimator

2.1 Definition: Statistical functional

Statistical function $T(F)$ (或 $\theta(F)$) 为 **any function of F**

> **Example**
>
> - mean: $\int x dF(x)$
> - variance: $\int x^2 dF(x) - \left(\int x dF(x)\right)^2$
> - p^{th} quantile: $F^{-1}(p) = inf\{x : F(x) \ge p\}$

2.2 Definition: Linear functional

Linear functional 可以被写作 $T(F) = \int r(x) dF(x) = \mathbb{E}_F[r(x)]$, 其中 $r(x)$ 为已知函数

> **Example**
>
> - mean 为一个 linear functional
> - variance 和 quantile function 不是 linear functional

2.3 Linear functional 的 plug-in estimator

若 T 为 linear functional, 则

$$T(\hat{F}_n) = \int r(x) d\hat{F}_n(x) = \frac{1}{n} \sum_{i=1}^{n} r(X_i)$$

Example

令 $X_1, \dots, X_n \overset{i.i.d.}{\sim} F$, 则以下 quantities of interest 的 plug-in estimator 为:

- X_1 的 expected value

$$\theta(F) = \mathbb{E}_F[X_1] = \int x dF(x)$$
$$\hat{\theta}_{\text{plug-in}}(F) = \theta(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$$

- $exp(X_1)$ 的 expected value

$$\theta(F) = \mathbb{E}_F[exp(X_1)] = \int exp(x) dF(x)$$
$$\hat{\theta}_{\text{plug-in}}(F) = \theta(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} exp(X_i)$$

- X_1 的 variance

$$\begin{aligned}\theta(F) &= \mathbb{E}_F[X_1^2] - (\mathbb{E}[X_1])^2 \\ &= \int x^2 dF(x) - \left(\int x dF(x)\right)^2\end{aligned}$$

$$\hat{\theta}_{\text{plug-in}}(F) = \theta(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^2$$

- F 的 median

3 Empirical CDF 的 Confidence Interval

3.1 Dvoretzky-Kiefer-Wolfowitz Inequality

令 $X_1, X_2, \dots, X_n \overset{i.i.d.}{\sim} F$,
则对任意 $\epsilon > 0$, 有

$$P(\sup_x |F(x) - \hat{F}_n(x)| > \epsilon) \le 2e^{-2n\epsilon^2}$$

换言之, 对 $\forall x$, 有

$$P(|F(x) - \hat{F}_n(x)| \le \epsilon) \ge 1 - 2e^{-2n\epsilon^2}$$

Remark

类似于 Chebyshev inequality, 该不等式为 empirical distribution function 与其 population distribution function 之间的 worst-case distance 提供了一个界限

3.2 Empirical CDF 的 confidence interval

以下 functions 构成了 F 的 global $1-\alpha$ confidence band:

$$L(x) = max\{\hat{F}_n(x) - \epsilon_n, 0\}$$
$$U(x) = min\{\hat{F}_n(x) + \epsilon_n, 1\}$$
$$\epsilon_n = \sqrt{log(2/\alpha)/(2n)}$$

Remark

令 $2e^{-2n\epsilon^2} = \alpha$ 即可求得 ϵ_n

即

$$P(L(x) \le F(x) \le U(x) \text{ for all } x) \ge 1 - \alpha$$

3.3 Linear functional 的 plug-in estimator 的 confidence interval

若 T 为 linear functional,

则 plug-in estimator 通常满足 $T(\hat{F}_n) \sim \mathcal{N}(T(F), \hat{se}^2)$, 因此可以将 $T(F)$ 的 $1-\alpha$ confidence interval 构建为

$$T(\hat{F}_n) \pm z_{\alpha/2} \cdot \hat{se}$$

Remark

由于 linear functional 的 plug-in estimator 的形式为 summation of random variables, 由 CLT 通常可以得到 asymptotic normality

Example

Example: Verify that the R expression

```
mean(x) + c(-2, 2) * sd(x)/sqrt(length(x))
```

produces an approximate 95% confidence interval for the mean waiting time for Old Faithful Geyser Data (built-in data in R).