STAT 383C: Statistical Modeling I **Fall 2015**

Lecture 23 — November 22

Lecturer: Purnamrita Sarkar *Scribe: Mariia Shovkun*

Disclaimer: These scribe notes have been slightly proofread and may have typos etc.

Note: The latex template was borrowed from EECS, U.C. Berkeley. Also these set of notes closely followed some of Larry Wasserman's class notes from CMU.

23.1 Non-Parametric bootstrap - new

Let us assume that we have n data points. $X_1...X_n$ are i.i.d. $X_1...X_n \sim F$, where F is a set of cell distributions. F is a set of all cdf's.

T: $F \to$ R

θ =T(F) which is a sophistical function

$$F \text{ is } \begin{cases} \text{discrete}, pmf, p(x) \\ \text{continuous}, pdf, f(x) \end{cases}$$

We also define:

$$\int g(x)dF(x) = \begin{cases} \sum g(x_j)p(x_j), \text{discrete} \\ \int g(x)f(x), \text{continuous} \end{cases} \tag{23.1}$$

Example: Mean, variance, and median are all functionals of the population cdf F:

$$\mu := \int xdF(x) \tag{23.2}$$

$$\sigma^2 := \int (x-\mu)^2 dF(x) = \int x^2 dF(x) - \left(\int xdF(x)\right)^2 \tag{23.3}$$

$$\text{Population median} := F^{-1}(\frac{1}{2}) \tag{23.4}$$

More complicated examples would include the largest eigenvalue of the covariance matrix.

Now, why are we doing all this? As it turns out, most estimators can be written as substituting the F by $\hat{F}_n$, where $\hat{F}_n$ is empirical C.D.F; it puts $\frac{1}{n}$ mass on each data point.

$$F(t) = P(x \le t) \qquad \hat{F}_n(t) = \frac{\sum_{i=1}^n 1(x_i \le t)}{n} \tag{23.5}$$

If $\theta = T(F)$, then the estimator is:

$$\hat{\theta}_n = T(\hat{F}_n)$$

We will now put down more groundwork by introducing a linear statistical functional. $T(F)$ is a linear functional, if

$$T(F) = \int a(x)dF(x).$$

The "plug-in" estimator is given by:

$$T(\hat{F}_n) = \int a(x)dF_n(x) = \frac{\sum_{i=1}^n a(x_i)}{n}$$

More examples of "plug-in" estimators: $\mu = \int xdF(x)$

$$\hat{\mu} = \frac{\sum_{i=1}^n x_i}{n} \tag{23.6}$$

$$\hat{\sigma}^2 = \frac{\sum x_i^2}{n} - \left(\frac{\sum x_i}{n}\right)^2 \tag{23.7}$$

$$\text{Estimator for the median} = \hat{F}_n^{-1}(1/2) \tag{23.8}$$

The plugin estimator of skewness $\kappa = \frac{E(x-\mu)^3}{\sigma^3}$, is given by:

$$\hat{\kappa} = \frac{\sum_i (x_i - \bar{x})^3/n}{\hat{\sigma}^3}$$

Well, bootstrap mainly comes into play when you want to calculate the variance of an estimator, or a confidence interval. Lets start this by first calculating a non-parametric CI for the median.

Non-parametric confidence interval for the median of F Let $Y_1 ... Y_n \sim F$. We want t_1, t_2 such that,

$$p(t_1 \leq \theta \leq t-2) \geq 1-\alpha$$

where θ is the median.

Define:

$$Z_i = \frac{\text{sign}(Y_i - \theta) + 1}{2} = \begin{cases} 1, \text{if } Y_i > \theta \\ 0, \text{if } Y_i < \theta \end{cases}$$

Since $Z_i \sim$ Bernoulli(1/2), then

$$T = \sum_{i=1}^n Z_i \sim \text{Binomial}(n, \frac{1}{2})$$

Lets find k_1,, k_2, such that

$$P(k_1 \leq T \leq k_2) \geq 1-\alpha$$

$$\{T \geq k_1\} = \{\#\{i : Y_i > \theta\} \geq k_1\} \iff \theta < Y_{(n-k_1+1)}$$

Using the same approach

$$\{T \leq k_2\} = \{\#\{i : Y_i < \theta\} \leq k_2\} \iff \theta > Y_{(n-k_2)}$$

Here $Y_{(i)}$ is the i^{th} order statistic.

$$P(Y_{(n-k_2)}) < \theta < T_{(n-k_1+1)} \geq 1 - \alpha$$

So $t_1 = Y_{(n-k_2)}$ and $T = Y_{(n-k_1+1)})$. How to define k_1 and k_2? For the large n $(T - n/2)/\sqrt{n/4} \to N(0,1)$. So we can read it from the normal table. Can get k_1, k_2 from Normal table because

$$\frac{T - n/2}{\sqrt{n/4}} \to N(0,1)$$

How about the variance of the median? Let $\hat{\theta}_n$ be the meadian of $X_1, \ldots, X_n \sim F$, which is a symmetric distribution with pdf f and location parameter θ and scale parameter $\sigma^2 = 1$.

$$\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow{d} N\left(0, \frac{1}{4f(\theta)^2}\right)$$

Note that in order to get the asymptotic variance, we will need to know the density f. This is where Bootstrap comes into play. Its a blackbox which estimates the variance of most estimators, as long as F is "well behaved". Lets start by writing the variance of the following estimator.

$$\hat{\theta_n} = g(X_1, ...X_n)$$

Assuming that $X_1...X_n$ are iid, $X_1...X_n \sim F$, and that F is known:

$$Var(\hat{\theta_n}) = \int \left(\hat{\theta_n} - E\hat{\theta_n}\right)^2 dF = \int \hat{\theta_n}^2 dF - \left(\int \theta_n dF\right)^2$$

Approximate the variance using sampling:
Pick the sample from $X_1...X_n \sim F$. Calculate

$$\hat{\theta_n}^{(i)} = g(X_1^{(i)}, X_2^{(i)}, ...X_n^{(i)})$$

Let us do B datasets. Then the

$$\hat{\text{Var}}(\hat{\theta_n}) = \frac{\sum \theta_n^{(i)2}}{B} - \left(\frac{\sum \theta_n^{(i)}}{B}\right)^2$$

At this step F is unknown. Plug in with the empirical distribution $X_1^{(1)}...X_n^{(1)} \sim \hat{F}_n$ Draw the samples from empirical distribution. This is essentially drawing n samples with replacement from the data. Let's draw B such sampled datasets.

$$\begin{array}{ll} X_1^{(1)} \ ... \ X_n^{(1)} & \theta_n^{*(1)} \\ ..... & .... \\ X_1^{(B)} \ ... \ X_n^{(B)} & \theta_n^{*(B)} \end{array}$$

$$S*^2 = \text{Var}\{\hat{\theta_n}^{*(1)}, \hat{\theta_n}^{*(B)}\}$$

Now there are two steps in approximation/error for non-parametric bootstrap.

We assume

$$\text{Var}_F(\hat{\theta_n}) \approx \text{Var}(\hat{\theta_n}|X_1, \ldots, X_n) \approx S*^2$$

The first approximation is the bottleneck here, since in some cases, $\hat{F}_n$ is not a good proxy for F. The second approximation can be made very small with large B. Larry Wasserman suggests take $B = 10,000$.

Remember Parametric bootstrap? F is known, parameter is unknown $X_1^{(1)}...X_n^{(1)} \sim F_{\hat{\theta}}$

Now when does Bootstrap not work? Typically when the test statistic (suitably scaled and centered) has a normal limit, Bootstrap works. Here is an example you have seen before.

Example.

$$X_1, ...X_n \approx U([0; \theta])$$

$$\hat{\theta} = \max(X_1, ... \ X_n)$$

It converges into exponential:

$$n\left(\theta - \hat{\theta_n}\right) \xrightarrow{d} \text{Exp}(\theta)$$

Next week we will talk about subsampling. The main idea is to take smaller size b samples **without** replacement.

$$\begin{array}{ll} \hat{\theta_b}^{\prime(1)} \leftarrow & \{X_1^{(1)} \ ... \ X_b^{(1)}\} \\ .... & ..... \\ \hat{\theta_b}^{\prime(B)} \leftarrow & \{X_1^{(B)} \ ... \ X_b^{(B)}\} \end{array}$$

The idea is to take $b \to \infty$ and $\frac{b}{n} \to 0$. Now we again calculate the variance of the $\hat{\theta_b}^{\prime(1)}, \ldots, \hat{\theta_b}^{\prime(B)}{}'s$. But since the sizes are different from n we will need to rescale. More later.