36-705: Intermediate Statistics **Fall 2017**

Lecture 33: November 29

Lecturer: Siva Balakrishnan

Today we will continue discussing the bootstrap, and then try to understand why it works in a simple case. In the last lecture we discussed estimating the variance of an estimator. Today, we will discuss constructing confidence intervals using the bootstrap, and then start discussing model selection.

33.1 Bootstrap Confidence Intervals

The bootstrap can also be used to obtain confidence intervals. If your estimator has a normal limit then you could just use a Wald interval with the bootstrap variance estimate, i.e. $C_n = [\widehat{\theta}_n - \widehat{s}z_{\alpha/2}, \widehat{\theta}_n + \widehat{s}z_{\alpha/2}]$.

It is often more accurate to use the distribution of the bootstrap estimates itself to construct the bootstrap confidence interval.

33.1.1 Hypothetical confidence interval

Suppose we knew the distribution of our estimator, in particular suppose we knew the distribution of $\sqrt{n}(\widehat{\theta}_n - \theta)$. Let us denote the distribution by G and denote its $\alpha/2$ and $1-\alpha/2$ quantiles by $g_{\alpha/2}$ and $g_{1-\alpha/2}$.

Then a $1-\alpha$ confidence interval would be:

$$C_n = \left[\widehat{\theta}_n - \frac{g_{1-\alpha/2}}{\sqrt{n}}, \widehat{\theta}_n - \frac{g_{\alpha/2}}{\sqrt{n}}\right].$$

This might seem a little strange, but this is probably because you are used to confidence intervals based on the normal distribution which has symmetric quantiles. To verify this,

$$\begin{aligned}\mathbb{P}(\theta \in C_n) &= \mathbb{P}\left(g_{\alpha/2} \leq \sqrt{n}(\widehat{\theta}_n - \theta) \leq g_{1-\alpha/2}\right) \\ &= 1 - \alpha/2 - \alpha/2 = 1 - \alpha.\end{aligned}$$

Again the point is that we do not know the distribution G above so we try to approximate this using the bootstrap.

33.1.2 Bootstrap confidence interval algorithm

Bootstrap Confidence Interval

1. Draw a bootstrap sample $X_1^*, \ldots, X_n^* \sim P_n$. Compute $\widehat{\theta}_n^* = g(X_1^*, \ldots, X_n^*)$.
2. Repeat the previous step, B times, yielding estimators $\widehat{\theta}_{n,1}^*, \ldots, \widehat{\theta}_{n,B}^*$.
3. Let
$$\widehat{G}(t) = \frac{1}{B}\sum_{j=1}^{B} I\Big(\sqrt{n}(\widehat{\theta}_{n,j}^* - \widehat{\theta}_n\Big) \leq t).$$
4. Let
$$C_n = \left[\widehat{\theta}_n - \frac{g_{1-\alpha/2}}{\sqrt{n}},\ \widehat{\theta}_n - \frac{g_{\alpha/2}}{\sqrt{n}}\right]$$
where $g_{\alpha/2} = \widehat{G}^{-1}(\alpha/2)$ and $g_{1-\alpha/2} = \widehat{G}^{-1}(1-\alpha/2)$.
5. Output C_n.

33.2 Variants

There are many many many papers that have been written about the bootstrap. Particularly, there are lots of variants – the "studentized" bootstrap where you throw in some estimates of the standard deviation in constructing confidence intervals, the block bootstrap for time-series, the residual bootstrap or the wild bootstrap for regression, the parametric bootstrap for parametric models, the smooth bootstrap and ideas related to sub-sampling to avoid certain regularity conditions, the less computationally intensive but less general Jackknife and so on.

33.3 Justifying the Bootstrap

This part is going to be a little bit technical. Before we get into it, we should try to figure out what it means to "justify the bootstrap". Roughly, we want that the quantiles of the bootstrap distribution of our statistic should be close to the quantiles its actual distribution, i.e. suppose we define:

$$\widehat{F}_n(t) = \mathbb{P}_n(\sqrt{n}(\widehat{\theta}_n^* - \widehat{\theta}_n) \geq t | X_1, \ldots, X_n),$$

to be the CDF of the bootstrap distribution, and

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \geq t),$$

to be the CDF of the true sampling distribution of our statistic, then the bootstrap works if for instance:

$$\sup_t |\widehat{F}_n(t) - F_n(t)| \to 0.$$

This turns out to be true in quite a bit of generality, only requiring mild conditions (Hadamard differentiability, see Bootstrap chapter in van der Vaart), but we will prove it in the simplest case: when $\widehat{\theta}_n$ is a sample mean. In this case there are much simpler ways to construct confidence intervals (using Normal approximations) but that is not really the point.

Suppose that $X_1, \ldots, X_n \sim P$ where X_i has mean μ and variance σ^2. Suppose we want to construct a confidence interval for μ.

Let $\widehat{\mu}_n = \frac{1}{n}\sum_{i=1}^n X_i$ and define

$$F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\mu}_n - \mu) \leq t). \tag{33.1}$$

We want to show that

$$\widehat{F}_n(t) = \mathbb{P}\Big(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n) \leq t \;\Big|\; X_1, \ldots, X_n\Big)$$

is close to F_n.

Theorem 33.1 (Bootstrap Theorem) *Suppose that $\mu_3 = \mathbb{E}|X_i|^3 < \infty$. Then,*

$$\sup_t |\widehat{F}_n(t) - F_n(t)| = O_P\left(\frac{1}{\sqrt{n}}\right).$$

To prove this result, let us recall that Berry-Esseen Theorem.

Theorem 33.2 (Berry-Esseen Theorem) *Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2. Let $\mu_3 = \mathbb{E}[|X_i - \mu|^3] < \infty$. Let $\overline{X}_n = n^{-1}\sum_{i=1}^n X_i$ be the sample mean and let Φ be the cdf of a $N(0,1)$ random variable. Let $Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma}$. Then*

$$\sup_z \Big|\mathbb{P}(Z_n \leq z) - \Phi(z)\Big| \leq \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}. \tag{33.2}$$

Proof of the Bootstrap Theorem. Let $\Phi_\sigma(t)$ denote the cdf of a Normal with mean 0 and variance σ^2. Let $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \widehat{\mu}_n)^2$. Thus, $\widehat{\sigma}^2 = \mathrm{Var}(\sqrt{n}(\widehat{\mu}_n^* - \widehat{\mu}_n)|X_1, \ldots, X_n)$. Now, by the triangle inequality,

$$\begin{aligned}\sup_t |\widehat{F}_n(t) - F_n(t)| &\leq \sup_t |F_n(t) - \Phi_\sigma(t)| + \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| + \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \\ &= \mathrm{I} + \mathrm{II} + \mathrm{III}.\end{aligned}$$

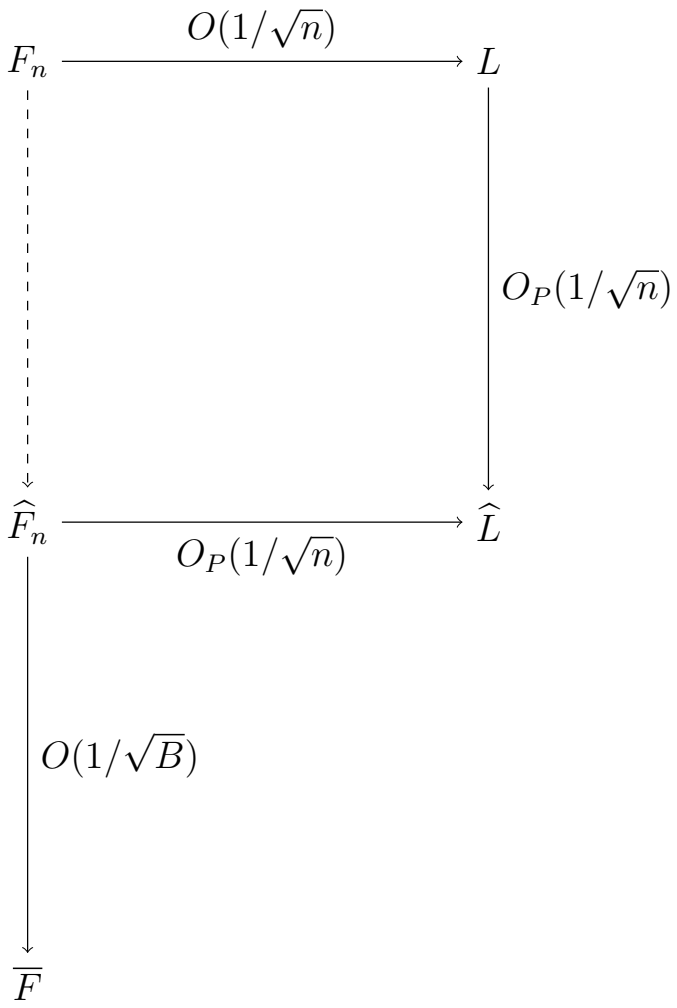

Figure 33.1: *The distribution* $F_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}_n - \theta) \leq t)$ *is close to some limit distribution* L*. Similarly, the bootstrap distribution* $\widehat{F}_n(t) = \mathbb{P}(\sqrt{n}(\widehat{\theta}^*_n - \widehat{\theta}_n) \leq t | X_1, \ldots, X_n)$ *is close to some limit distribution* $\widehat{L}$*. Since* $\widehat{L}$ *and* L *are close, it follows that* F_n *and* $\widehat{F}_n$ *are close. In practice, we approximate* $\widehat{F}_n$ *with its Monte Carlo version* $\overline{F}$ *which we can make as close to* $\widehat{F}_n$ *as we like by taking* B *large.*

Let $Z \sim N(0,1)$. Then, $\sigma Z \sim N(0,\sigma^2)$ and from the Berry-Esseen theorem,

$$\begin{aligned} \mathrm{I} &= \sup_t |F_n(t) - \Phi_\sigma(t)| = \sup_t \left|\mathbb{P}\left(\sqrt{n}(\widehat{\mu}_n - \mu) \le t\right) - \mathbb{P}\left(\sigma Z \le t\right)\right| \\ &= \sup_t \left|\mathbb{P}\left(\frac{\sqrt{n}(\widehat{\mu}_n - \mu)}{\sigma} \le \frac{t}{\sigma}\right) - \mathbb{P}\left(Z \le \frac{t}{\sigma}\right)\right| \le \frac{33}{4}\frac{\mu_3}{\sigma^3\sqrt{n}}. \end{aligned}$$

Using the same argument on the third term, we have that

$$\mathrm{III} = \sup_t |\widehat{F}_n(t) - \Phi_{\widehat{\sigma}}(t)| \le \frac{33}{4}\frac{\widehat{\mu}_3}{\widehat{\sigma}^3\sqrt{n}}$$

where $\widehat{\mu}_3 = \frac{1}{n}\sum_{i=1} |X_i - \widehat{\mu}_n|^3$ is the empirical third moment. By the strong law of large numbers, $\widehat{\mu}_3$ converges almost surely to μ_3 and $\widehat{\sigma}$ converges almost surely to σ. So, almost surely, for all large n, $\widehat{\mu}_3 \le 2\mu_3$ and $\widehat{\sigma} \ge (1/2)\sigma$ and $\mathrm{III} \le \frac{33}{4}\frac{4\mu_3}{\sqrt{n}}$. From the fact that $\widehat{\sigma} - \sigma = O_P(\sqrt{1/n})$ it may be shown that $\mathrm{II} = \sup_t |\Phi_\sigma(t) - \Phi_{\widehat{\sigma}}(t)| = O_P(\sqrt{1/n})$. (This may be seen by Taylor expanding $\Phi_{\widehat{\sigma}}(t)$ around σ.) This completes the proof. □

We have shown that $\sup_t |\widehat{F}_n(t) - F_n(t)| = O_P\left(\frac{1}{\sqrt{n}}\right)$. From this, it may be shown that, for each $0 < \beta < 1$, $t_\beta - z_\beta = O_P\left(\frac{1}{\sqrt{n}}\right)$.

So far we have focused on the mean. Similar theorems may be proved for more general parameters. The details are complex so we will not discuss them here.

33.4 Model Selection

In non-parametric regression we had an unknown bandwidth parameter. In practice, this tuning parameter is chosen using cross-validation.

Before we discuss cross-validation lets understand a train-test split, i.e. suppose we split the data into two parts we can estimate our regression function for a grid of bandwidths $\{h_1, \ldots, h_M\}$ on one part of the data. Now, we want to pick one of these bandwidths.

In this case, we could simply check how well we can predict on the test set, i.e.,

$$\widehat{R}(\widehat{f}_{h_1}, f) = \frac{1}{n_{\text{test}}}\sum_{i=1}^{n_{\text{test}}}(Y_i - \widehat{f}_{h_1}(X_i))^2,$$

and repeat this for each of the bandwidths and then pick the bandwidth that minimizes this. Why does this work? Essentially, train-test splits allow us to *estimate* the risk, and then we are picking the value of the tuning parameter that minimizes our estimated risk.

We should be a little bit careful about what risk we are estimating:

$$\mathbb{E}(\widehat{R}(\widehat{f}_{h_1}, f)) = \mathbb{E}(f(X) + \epsilon - \widehat{f}_{h_1}(X))^2 = \mathbb{E}(f(X) - \widehat{f}_{h_1}(X))^2 + \sigma^2.$$

We can of course ignore the σ^2, but one should notice that the risk we are estimating:

$$\mathbb{E}(f(X) - \widehat{f}_{h_1}(X))^2 = \int (f(x) - \widehat{f}_{h_1}(x))^2 p(x) dx,$$

where p is the density of the covariates. This is sometimes called the $L_2(\mathbb{P})$-risk, as opposed to the L_2-risk which we defined earlier:

$$R = \int (f(x) - \widehat{f}_{h_1}(x))^2 dx.$$

Most people would consider the $L_2(\mathbb{P})$-risk to be more natural, since it puts less weight in places where you have less data.

Practitioners might be concerned that this the train-test split is wasteful of the data: you might need a pretty large test set to get a good estimate and this is data that you might have instead used to estimate the model. Also, train-test splits have a "lottery" effect: you might get unlucky in the way you split the data and this could affect results.

K-fold cross-validation tries to get around this by splitting the data into K pieces (think of K as a small number like 5). Now, we repeat the train-test split K times, each time we use $K-1$ pieces for training and the K-th piece for testing. In this way we get, K estimates of the error for each value of the bandwidth. We average these K numbers to get our risk estimate. Finally, we choose the value of the bandwidth that minimizes the risk. The extreme case of K-fold cross-validation is called leave-one-out or n-fold cross-validation. Here we leave out one observation, and try to predict it and then cycle through the observations.

The basic question is then: is there a sense in which cross-validation is "doing the right thing"?

33.5 A simple analysis of the train-test split

Lets try to understand cross-validation in a simple scenario. We will do this in the context of point estimation, but one could use *exactly* the same argument for bandwidth selection.

Say we have models $\mathcal{M}_1, \ldots, \mathcal{M}_M$. These are different models that we think might be reasonable fits to the data. Now, we observe our data $(X_1, \ldots, X_{2n})$ and randomly split it into train and test sets of size n each. We really should refer to the test set as a validation set but we will ignore this for today.

On the train set, we fit our models (say using the MLE), and compute point estimates $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$. Now, suppose that we want to select the model/estimate that fits the data well. We will use the negative log-likelihood as our measure, i.e., we want an estimate that has low negative log-likelihood. This is the same as using the KL divergence as our loss function.

We can use the test set to estimate the negative log-likelihood:

$$R_i = \frac{-1}{n}\sum_{i=1}^{n} \log f_{\widehat{\theta}_i}(X_{n+i}).$$

Note that:

$$\mathbb{E}(R_i) = -\mathbb{E}_{f_{\theta^*}} \log f_{\widehat{\theta}_i}(X) = KL(f_{\theta^*}||f_{\widehat{\theta}_i}) - \mathbb{E}_{f_{\theta^*}} \log f_{\theta^*}(X),$$

so we are estimating the KL divergence upto some term that does not depend on $\widehat{\theta}_i$. So minimizing $\mathbb{E}(R_i)$ is equivalent to minimizing the KL divergence.

We can now use the LLN to argue that if the test-set size goes to ∞ then our risk estimates converge to their expectations, and then we will find the model/estimate with the lowest KL to the true model.

Suppose however we wanted to be more precise, and try to understand the role of the test set size and the number of models M?

We could use Hoeffding's inequality. This will need an assumption that $|\log f_\theta(X)| \leq B$ for every θ and X that we care about (this can be relaxed using more complex techniques). Now, notice that the following is an important but straightforward consequence of Hoeffding's inequality:

$$\mathbb{P}(\max |R_i - \mathbb{E}(R_i)| \geq \epsilon) \leq 2M \exp(-2n\epsilon^2/(4B^2)).$$

This is true since for each i we know that

$$\mathbb{P}(|R_i - \mathbb{E}(R_i)| \geq \epsilon) \leq 2 \exp(-2n\epsilon^2/(4B^2)).$$

so we can obtain the desired inequality via a union bound (if the max exceeds ϵ at least one of the terms must exceed ϵ).

Define,

$$\epsilon_n = \sqrt{\frac{4B^2 \log(2M/\alpha)}{n}},$$

then we know that

$$\mathbb{P}(\max |R_i - \mathbb{E}(R_i)| \geq \epsilon_n) \leq \alpha.$$

Suppose we select the model $\widehat{i} = \arg\min_i R_i$, and let $i^* = \arg\min_i \mathbb{E}(R_i)$, then we have that with probability at least $1 - \alpha$:

$$\mathbb{E}(R_{\widehat{i}}) \leq R_{\widehat{i}} + \epsilon_n \leq R_{i^*} + \epsilon_n \leq \mathbb{E}(R_{i^*}) + 2\epsilon_n.$$

So the model we select will be sub-optimal by at most $2\epsilon_n$. In regression, we would use exactly the same reasoning, but just replace the risk with the squared loss.

Reasoning about K-fold cross-validation turns out to be much more challenging, because the data re-use breaks independence assumptions.

The analysis above should remind you of the analysis we did before of Empirical Risk Minimization. The goals are slightly different, as is the final guarantee. It is worth thinking about what exactly the data splitting buys you. In particular, we do not require uniform convergence of the empirical to the true risk over all the model classes $\mathcal{M}_1, \ldots, \mathcal{M}_M$, rather we only require a good estimate of the risk for the *fixed* models indexed by $\widehat{\theta}_1, \ldots, \widehat{\theta}_M$.