Introduction to the Bootstrap

Prof. Sharon McNicholas

STATS 780/CSE 780

Introduction

- Before proceeding to boosting and bagging, we take a look at an extremely powerful technique in (modern) statistics.

- The following two texts are very useful for further reading:

  - Efron, B. and Tibshirani, R. J. (1993). *An Introduction to the Bootstrap*. Boca Raton: Chapman & Hall/CRC.

  - Davison, A. C. and Hinkley, D. V. (1997). *Bootstrap Methods and their Application*. New York: Cambridge University Press.

Resampling

- We have a sample (or ensemble) $\mathcal{Q} = \{x_1, x_2, \ldots, x_n\}$ and an estimator $\hat{\theta}$ based upon that sample.
- Suppose that we wish to estimate the bias and standard error of this estimator $\hat{\theta}$.
- We can use a **resampling technique**.
- Resampling simply means that we draw samples from an ensemble that is itself a sample.
- There are a variety of resampling techniques available; the most famous of these is called the bootstrap.
- Before we look at the bootstrap, we need to see the **plug-in principle**.

The Plug-In Principle

- Let $X_1, X_2, \ldots, X_n$ be iid random variables, then the cdf

$$G(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}_{X_i \leq x},$$

where $\mathbb{I}_k$ is an indicatior function, defines an **empirical distribution**.

- The **plug-in estimate** of the parameter of interest θ_F is given by

$$\hat{\theta} = \theta_{\hat{F}},$$

where $\hat{F}$ is an empirical distribution.

- For example, summary statistics are plug-in estimators.

The Bootstrap: The Idea

- Efron (2002)[a] gives the following description of the bootstrap.
  1. Suppose that the data are a random sample from some unknown probability distribution F.
  2. We are interested in the parameter θ.
  3. We want to know $\mathsf{SE}_F(\hat{\theta})$.
  4. We compute $\mathsf{SE}_{\hat{F}}(\hat{\theta})$, where $\hat{F}$ is the empirical distribution of F.

- The purpose of the bootstrap is (generally) to **assess variability**.

[a]Efron, B. (2002). The bootstrap in modern statistics *in* 'Statistics in the 21st Century', Raftery, A. E., Tanner, M. A. and Wells, M. T. (Eds.), Florida: Chapman & Hall / CRC, 326–332.

The Bootstrap: The Idea contd.

- Suppose again that we have an ensemble $\mathcal{Q} = \{x_1, x_2, \ldots, x_n\}$ and that we want to estimate the standard error of an estimator $\hat{\theta}$.

- We can do this by **sampling with replacement** from $\mathcal{Q}$, n times, to get a **bootstrap sample** $\mathcal{Q}^* = \{x_1^*, x_2^*, \ldots, \ldots, x_n^*\}$ and then computing $\hat{\theta}^*$ based on $\mathcal{Q}^*$.

- Repeating this process m times gives values $\hat{\theta}^*(1), \hat{\theta}^*(2), \ldots, \hat{\theta}^*(m)$ based on bootstrap samples $\mathcal{Q}^*(1), \mathcal{Q}^*(2), \ldots, \mathcal{Q}^*(m)$

- Then,
$$\widehat{\mathsf{SE}}_{\mathsf{boot}} = \sqrt{\frac{1}{m-1}\sum_{j=1}^{m}(\hat{\theta}^*(j) - \hat{\theta}^*(\cdot))^2},$$
where $\hat{\theta}^*(\cdot) = \sum_{i=1}^{m} \hat{\theta}^*(i)/m$.

The Bootstrap: Some Details

- As $m \to \infty$, $\widehat{\mathsf{SE}}_{\mathsf{boot}} \to \widehat{\mathsf{SE}}_{\hat{F}}$.
- $\widehat{\mathsf{SE}}_{\mathsf{boot}}$ and $\widehat{\mathsf{SE}}_{\hat{F}}$ are **non-parametric bootstrap estimates** since they are based on $\hat{F}$ rather than F.
- Clearly, we want m as large as possible but how large is large enough?
- There is no concrete answer but experience helps one get a feel for it.
- Exercise 1. Suppose we are interested in the median of these data: 10, 27, 31, 40, 46, 50, 52, 104, 146. Compute $\widehat{\mathsf{SE}}_{\mathsf{boot}}$.

STATS 780/CSE 780 Prof. Sharon McNicholas

The Bootstrap: Estimating Bias

- Suppose we want to estimate the bias of $\hat{\theta}$ given an ensemble $\mathcal{Q}$.
- The bootstrap estimate of the bias, using the familiar notation, is
$$\widehat{\mathsf{Bias}}_{\mathsf{boot}} = \hat{\theta}^*(\cdot) - \hat{\theta},$$
where $\hat{\theta}$ is computed based the empirical distribution $\hat{F}$.
- Exercise 2: Compute $\widehat{\mathsf{Bias}}_{\mathsf{boot}}$ for Exercise 1.

Bootstrap Percentile Intervals

- In addition to estimating the standard error and the bias, the bootstrap can be used to estimate confidence intervals.

- There are a number of ways to do this, and there is a good body of literature around this topic.

- The most straightforward method is to compute **bootstrap percentile intervals**.

- Let's look at an example taken from Efron and Tibshirani (1993).

Bootstrap PI Example

- Suppose we have the following data from a study set up to determine whether or not regular doses of a certain quantity of aspirin were effective at preventing stroke.

	Strokes	Subjects
Aspirin	119	11037
Placebo	98	11034

- Suppose that we are interested in comparing the treatments by looking at the ratio of the rates of strokes. Then:

$$\hat{\theta} = \frac{119/11037}{98/11034} = 1.21.$$

Bootstrap PI Example contd.

- To the untrained eye, this may seem to suggest that Aspirin is harmful.
- However, to see if this ratio is significantly different from 1, we need to estimate the variability of this estimatior.
- To do this we repeat the following many times:
  1. Form two populations: one consisting of 119 1's and 10918 0's and the the other having 98 1's and 10936 0's.
  2. Sample with replacement, 11037 items from the first population and 11034 from the second (this gives two bootstrap samples).
  3. Compute the ratio $\hat{\theta}^*$.

Bootstrap PI Example contd.

- Suppose that we repeat these three steps $m = 1000$ times.
- We can then write down a 95% bootstrap percentile interval by looking at the values at the 2.5th and 97.5th percentiles.
- I did this earlier and got the interval $(0.93, 1.60)$ for θ.
- Note that 1 is inside this interval.
- Note also that a 95% confidence interval for θ is given by $(0.93, 1.59)$.
- By the CLT, as $m \to \infty$ these two intervals will be more and more similar, as in this example.

Comments

- Again, I suggest the two aforementioned books for further reading, i.e., Efron and Tibshirani (1993) and Davison and Hinkley (1997).