Module 8: Resampling methods

Siyue Yang

06/07/2022

Outline

In this module, we will continue our discussion on resampling methods

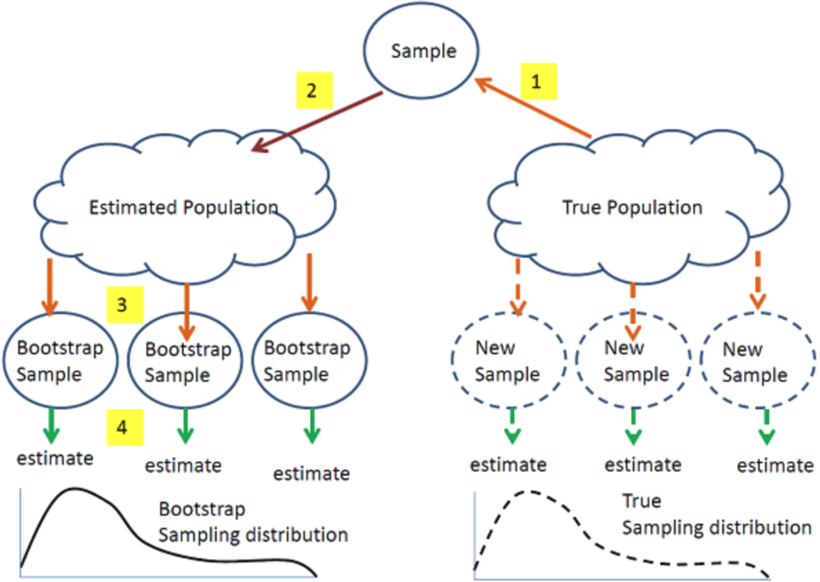
- Bootstrap
- Cross validation
- (Optional) Cluster computing

What is bootstrap?

A widely applicable, computer intensive resampling method used to compute standard errors, confidence intervals, and significance tests.

Why bootstrap?

- The exact sampling distribution of an estimator can be **difficult** to obtain
- Asymptotic expansions are sometimes easier but expressions for standard errors based on large sample theory may not perform well in finite samples



https://online.stat.psu.edu/stat555/node/119/

The bootstrap principle

Suppose $X = \{X_1, \ldots, X_n\}$ is a sample used to estimate some parameter $\theta = T(P)$ of the underlying distribution P. To make inference on θ, we are interested in the properties of our estimator $\hat{\theta} = S(X)$ for θ.

The bootstrap principle

Suppose $X = \{X_1, \ldots, X_n\}$ is a sample used to estimate some parameter $\theta = T(P)$ of the underlying distribution P. To make inference on θ, we are interested in the properties of our estimator $\hat{\theta} = S(X)$ for θ.

- If we knew P,
  - we could obtain $\left\{X^b \mid b = 1, \ldots B\right\}$ from P and use Monte-Carlo to estimate the sampling distribution of $\hat{\theta}$

The bootstrap principle

Suppose $X = \{X_1, \ldots, X_n\}$ is a sample used to estimate some parameter $\theta = T(P)$ of the underlying distribution P. To make inference on θ, we are interested in the properties of our estimator $\hat{\theta} = S(X)$ for θ.

- If we knew P,
  - we could obtain $\left\{X^b \mid b = 1, \ldots B\right\}$ from P and use Monte-Carlo to estimate the sampling distribution of $\hat{\theta}$
- However, we don't,
  - we do the next best thing and resample from original sample, i.e. the empirical distribution, $\hat{P}$
  - we expect the empirical distribution to estimate the underlying distribution well by the *Glivenko-Cantelli* Theorem

3 forms of bootstrap

Based on how the population is estimated,

1. Nonparametric bootstrap
2. Semiparametric bootstrap
3. Parametric bootstrap

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.
- Step 1: Select B independent bootstrap resamples $\mathbf{D}(b)$, each consisting of N data values drawn with replacement from the data.

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.
- Step 1: Select B independent bootstrap resamples $\mathbf{D}(b)$, each consisting of N data values drawn with replacement from the data.
- Step 2: Compute estimates from each bootstrap resample

$$\hat{\theta}^*(b) = S\left(\mathbf{D}^*(b)\right) \quad b = 1, \ldots, B$$

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.
- Step 1: Select B independent bootstrap resamples $\mathbf{D}(b)$, each consisting of N data values drawn with replacement from the data.
- Step 2: Compute estimates from each bootstrap resample

$$\hat{\theta}^*(b) = S\left(\mathbf{D}^*(b)\right) \quad b = 1, \ldots, B$$

- Step 3: Estimate the standard error se $(\hat{\theta})$ by the sample standard deviation of the B replications of $\hat{\theta}^*(b)$

Nonparametric bootstrap (resampling)

Reproduce the items that were in the original sample (sample with replacement)

- Example: estimate the standard error and confidence interval for some $\hat{\theta} = S(\mathbf{D})$ where $\mathbf{D}$ encodes our observed data.
- Step 1: Select B independent bootstrap resamples $\mathbf{D}(b)$, each consisting of N data values drawn with replacement from the data.
- Step 2: Compute estimates from each bootstrap resample

$$\hat{\theta}^*(b) = S\left(\mathbf{D}^*(b)\right) \quad b = 1, \ldots, B$$

- Step 3: Estimate the standard error se $(\hat{\theta})$ by the sample standard deviation of the B replications of $\hat{\theta}^*(b)$
- Step 4: Estimate the confidence interval by finding the $100(1-\alpha)$ percentile bootstrap CI,

$$\left(\hat{\theta}_L, \hat{\theta}_U\right) = \left(\hat{\theta^*}^{\alpha/2}, \hat{\theta^*}^{1-\alpha/2}\right)$$

Semiparametric bootstrap (adding noise)

- Assumes the population includes other items are similar to the observed sample by sampling from a smoothed version of the sample histogram

Parametric bootstrap

- Assumes the data comes from a known distribution with unknown parameters
- First estimate the parameters from the data and then use the estimated distribution to simulate the samples

StatQuest videos

Check out these videos made by Josh Starmer with vivid illustration for the boostrap!

- Bootstrapping Main Ideas [link]
- Using Bootstrapping to Calculate p-values [link]

Resources

This tutorial is based on

- PennState STAT555 Statistical Analysis of Genomics Data [links].
- Harvard's Biostatistics Preparatory Course Methods [links].