ISYE 7406: Data Mining and Statistical Learning

W

March 30, 2024

Contents

1 Bootstrapping algorithm

9.1.3

1.1 Motivating example

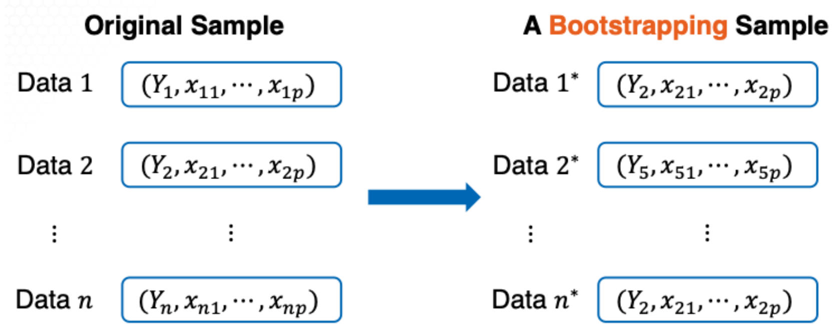
- Data: $z = (z_1, ..., z_n)$ where $z_i = (Y_i, x_{i1}, ..., x_{ip}), i = 1, ..., n$
- Parameter estimation: we derived real-valued summary statistics $S(z) = S(z_1, ..., z_n)$ which is en estimator of population parameter θ (e.g. mean, median, correlation coefficient, regression coefficient, etc.)
- Objective: derive a **robust** estimator of the confidence interval of θ or the standard error of $S(z)$
- Challenges: we don't know the distribution of data, and are unable to obtain additional training datasets

1.2 Idea in Bootstrapping algorithm

- Intuitive idea:
  - Estimate the standard error of $S(z) = S(z_1, ..., z_n)$ based on the sample standard deviation if we have many values of $S(z)$ or many independent copies of training data

– from the original training dataset ($z = z_1, z_2, ..., z_n$) to generate many copies of "new" training dataset. This allows us to compute many values of $S(z)$

1.3 Resample with replacement



- High level:
  - Input Data: $z = (z_1, ..., z_n)$ where $z_i = (Y_i, x_{i1}, ..., x_{ip}), i = 1, ..., n$ and estimator of $S(z)$ for θ
  - For $b = 1, ..., B$
    - **Sample with replacement** to get bootstrap sample $z^{*b} = (z_1^{*b}, ..., z_n^{*b})$
    - Compute the value $S(z^{*b}) = S(z_1^{*b}, ..., z_n^{*b})$ for this bootstrap sample
  - Once the B values of $S(z^{*b})$ have been computed,
    - The quantiles of $S(z^{*b})$'s provide an empirical distribution of $S(z)$
    - They can be used to provide confidence intervals of θ