Introduction to Bootstrap

DS 301

Iowa State University

Resampling Techniques

1. Cross-validation:
    - Used to estimate supervised test error (prediction or classification error)
    - Can also help us to find an optimal tuning parameter (such as λ in regularized regression)
2. Bootstrap
    - Used to estimate uncertainty surrounding a statistical approach.
    - Common examples:
        - Estimate the standard error of parameter estimates
        - Construct confidence intervals.

Boostrap

- One of the most important techniques in all of data science/statistics.
- Widely applicable, extremely powerful, computationally intensive.
- Literally involves just resampling from the data.
- No distributional assumptions.
- Requires a moderately sized data set.

Recall `lm()` output

```
> summary(lm(medv~.,data=Boston))

Call:
lm(formula = medv ~ ., data = Boston)

Residuals:
    Min      1Q  Median      3Q     Max 
-15.595  -2.730  -0.518   1.777  26.199 

Coefficients:
              Estimate Std. Error t value Pr(>|t|)    
(Intercept)  3.646e+01  5.103e+00   7.144 3.28e-12 ***
crim        -1.080e-01  3.286e-02  -3.287 0.001087 ** 
zn           4.642e-02  1.373e-02   3.382 0.000778 ***
indus        2.056e-02  6.150e-02   0.334 0.738288    
chas         2.687e+00  8.616e-01   3.118 0.001925 ** 
nox         -1.777e+01  3.820e+00  -4.651 4.25e-06 ***
rm           3.810e+00  4.179e-01   9.116  < 2e-16 ***
age          6.922e-04  1.321e-02   0.052 0.958229    
dis         -1.476e+00  1.995e-01  -7.398 6.01e-13 ***
rad          3.060e-01  6.635e-02   4.613 5.07e-06 ***
tax         -1.233e-02  3.760e-03  -3.280 0.001112 ** 
ptratio     -9.527e-01  1.308e-01  -7.283 1.31e-12 ***
black        9.312e-03  2.686e-03   3.467 0.000573 ***
lstat       -5.248e-01  5.072e-02 -10.347  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

How did we obtain these standard errors and carry out inference?

(1) linear relationship btwn Y & X

(2) constant variance

σ^2

↳ $se(\hat{B})$

(3) normality

↳ inference (t-tests / F-tests)

How did we obtain these standard errors and carry out inference?

What happens if

- The modeling assumptions break down?
- There is no analytical formula that can be derived. $se(\hat{\beta})$?

Bootstrap standard error and confidence intervals!

- Requires no math.
- Requires no distributional assumptions.

Boostrap standard errors

Suppose we are working on the lasso model. We estimate our parameters $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_p$. (λ^*)

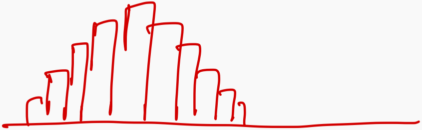
We want to quantify: how accurate are our estimates?

Let's say I knew the true sampling distribution of $\hat{\beta}_1$, what can we do? $\sim N(0,1)$.

$$\left\{ \begin{array}{l} \hat{B}_1 \sim N(0,1) \\ se(\hat{B}_1)\,? \end{array} \right. \longrightarrow$$ I could literally just re-sample from this sampling distributions.

$$\Rightarrow \sqrt{var(\hat{B}_1)}$$

$$= \sqrt{\frac{1}{99} \sum_{i=1}^{n} (\hat{B}_i - \bar{B}_i)^2}$$



sampling distr. of $\hat{B}_1$.

In reality, we do not know true sampling distr. of $\hat{B}_1$.

Bootstrap standard errors

Pick a large number: $B = 1000$ and repeat the following for $b = 1, \dots, B$

(1) Draw a bootstrap sample

original data:

Z_1
Z_2
Z_3
$\vdots$
Z_n

$\xrightarrow{\text{Draw } n \text{ observations w/ replacement}}$

bootstrap sample:

$\tilde{Z}_1^{(b)}$
$\tilde{Z}_2^{(b)}$
$\vdots$
$\tilde{Z}_n^{(b)}$

can have repeated observation

(2) for this bootstrap sample, compute your estimate of interest using $\tilde{Z}_1^{(b)}, \tilde{Z}_2^{(b)}, \dots, \tilde{Z}_n^{(b)}$.

$\Longrightarrow$ Repeat this process $B = 1000$ times.

Bootstrap standard errors

ex: $\hat{B}_1$

$\hat{B}_1^{(1)}$

$\hat{B}_1^{(2)}$

$\hat{B}_1^{(3)}$

$\vdots$

$\hat{B}_1^{(1000)}$

$$\Longrightarrow \quad se_B(\hat{B}_1) = \sqrt{\frac{1}{B-1}\sum_{i=1}^{b}\left(\hat{B}_1^{(b)} - \bar{B}_1\right)^2}$$

each bootstrap estimate

avg of all your bootstrap estimates.

$\Longrightarrow$ This can be applied to (almost) any parameter θ and it's estimate. Does not need to be in context of a model.

↳ correlation → se

↳ percentile → se

↳ $\hat{Y}$ → se

Bootstrap confidence intervals

- Classically, confidence intervals require distributional assumptions.
- Recall from linear regression, if I wanted to construct a confidence interval for $\hat{\beta}$, I needed one of the following:

(1) $\varepsilon_i \sim N(0,1)$

(2) CLT kicks in for large enough n.

(3) Bootstrap

What is a reasonable range β_1 could be in?

$\beta_1 \rightarrow \hat{\beta}_1$ (l.s. estimate) : $1-\alpha$ CI

$E(Y) \rightarrow \hat{Y}$

$\hat{\beta}_1 \pm \boxed{t_{1-\frac{\alpha}{2}; df}} \times se(\hat{\beta}_1)$

quantile from t distr. w/ df.

(1) $\varepsilon_i \sim N(0,1)$

(2) CLT

This quantity $t_{1-\alpha/2; df}$ is based on one of these 2 assumptions holding.

If these assumptions are questionable, then the validity of our CI may no longer hold.

2 levels of bootstrap:

outer loop & inner loop

$$\frac{\hat{B}_1 - 0}{se(\hat{B}_1)} \overset{H_0}{\sim} t_{df} \Rightarrow \frac{\hat{B}_1 - B_1}{se(\hat{B}_1)}$$ we want to bootstrap this

our target: $\dfrac{\tilde{B}_1^{(b)} - \hat{B}_1}{se(\tilde{B}_1^{(b)})}$

outer loop: $\tilde{B}_1^{(b)}$

inner loop: $se(\tilde{B}_1^{(b)})$

(1) outer loop.

for $b = 1, \ldots, B$ $(B = 500)$

original data: $Z_1, Z_2, \ldots, Z_n$.

Draw a bootstrap sample from original data:

$\tilde{Z}_1^{(b)}, \tilde{Z}_2^{(b)}, \ldots, \tilde{Z}_n^{(b)}$.

$\hookrightarrow \tilde{B}_1^{(b)}$ $\quad b = 1 \Rightarrow \tilde{B}_1^{(1)} = 3$.

(2) inner loop:

target $se(\tilde{\beta}_1^{(b)})$ for a specific iteration b.

repeat for $m = 1, \ldots, M$ $(M = 100)$

Draw a bootstrap sample from $\tilde{Z}_1^{(b)}, \tilde{Z}_2^{(b)}, \ldots \tilde{Z}_n^{(b)}$.

$\Rightarrow$ call this: $\tilde{Z}_1^{(b,m)}, \tilde{Z}_2^{(b,m)}, \ldots, \tilde{Z}_n^{(b,m)}$

obtain estimates for $\tilde{\beta}_1^{(b,m)}$

when $b = 1$: for $m = 1, \ldots 100$:

$\tilde{\beta}_1^{(1,1)}$ $\tilde{\beta}_1^{(1,2)}$ $\tilde{\beta}_1^{(1,3)}$ $\ldots$ $\tilde{\beta}_1^{(1,100)}$.

↳ compute standard error $se(\tilde{\beta}_1^{(b)})$.

final output: $\dfrac{\tilde{\beta}_1^{(b)} - \hat{\beta}_1}{se(\tilde{\beta}_1^{(b)})}$ (you'll have B of these).

$$\frac{\tilde{B}_1^{(b)} - \hat{B}_1}{se(\tilde{B}_1^{(b)})} \rightarrow \tilde{F}^{(b)}$$

$B = 500$

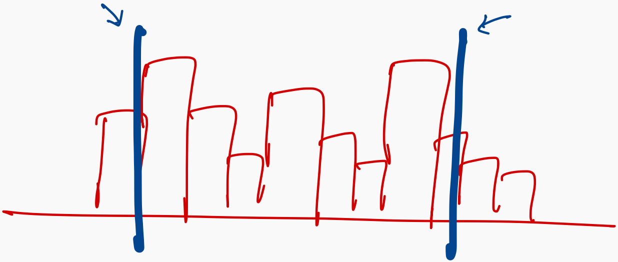
$\tilde{F}^{(1)}, \tilde{F}^{(2)}, \ldots, \tilde{F}^{(500)}$

$\cancel{\hat{B}_1 \pm t \times se(\hat{B}_1)}$

$\hat{B}_1 + q_{0.975} \times se(\hat{B}_1)$

$\hat{B}_1 - q_{0.025} \times se(\hat{B}_1)$

$\alpha = 0.05$



$q_{\alpha/2}$ $\quad$ $q_{1-\alpha/2}$

$q_{0.025}$ $\quad$ $q_{0.975}$