STAT 403

4/28/25

Lecture 12

Project ~ TBPosted

No Quizzes

Jacknife

LOO CV

BLB

Conformal Prediction

Lecture Notes VI – Modern resampling methods. Conformal prediction

Marina Meila
mmp@stat.washington.edu

Department of Statistics
University of Washington

May 2025

Jackknife

Bag of little bootstraps

Conformal prediction. Jackknife+

Boostrap - when? future lecture

1) The jackknife

Cross Validation

LOO CV

- like Leave-one-out CV
- $\mathcal{D}_{-i} = \mathcal{D} \setminus \{(x_i, y_i)\}$ or $\mathcal{D} \setminus \{x_i\}$ ← Leave One Out (i)
- θ is parameter of interest
  - $\hat{\theta}$ = estimate of θ from $\mathcal{D}$
  - $\boxed{\hat{\theta}_{-i}}$ = estimate of θ from $\mathcal{D}_{-i}$, $i = 1 : n$

training set $\mathcal{D}_{-i}$ $|\mathcal{D}_{-i}| = n-1$

test set $\{i\}$ $|\mathcal{D}^{test}| = 1$

- **jackknife Algorithm** estimates $F(\hat{\theta})$ and from it bias and variance of $\hat{\theta}$.
  1. Estimate $\hat{\theta}$ from $\mathcal{D}$
  2. for $i = 1 : n$
     estimate $\hat{\theta}_{-i}$ from $\mathcal{D}_{-i}$ $\Rightarrow \{\hat{\theta}_{-i}\}_{i=1:n} \longrightarrow \hat{F}_{\hat{\theta}_{-i}}$ Empirical CDF
  3. Use $\hat{F}(\hat{\theta}) \approx \hat{F}(\hat{\theta}_{-i}, \ i = 1 : n)$ to estimate bias, variance, ... $\hat{\theta}$

$\hat{\theta}$ = param of interest

want $F(\hat{\theta})$

$\sim E[\hat{\theta}]$

$\rightarrow \mathrm{Var}[\hat{\theta}]$

K times

2) # Bag of little bootstraps [arXiv:1112.5016]

$n' < n$

- For large n, sampling, estimating $\hat{\theta}^*$ are expensive! Can we use $n' = |\mathcal{D}^*| \ll n$?
- **Bag of little Bootstraps Algorithm**

Bag of K sets of size $n' \ll n$

$\{\mathcal{D}^{*k}\}$

for $k = 1 : K$

1. sample $\mathcal{D}^{*(k)}$ of size n' from $\mathcal{D}$ without replacement
2. do boostrap on $\mathcal{D}^{*(k)}$ with sample size n

   for $b = 1 : B$

   2.1 sample $(n_{i,k,b}, i = 1 : n') \sim$ multinomial $\left(n, \left[\frac{1}{n'}, \ldots \frac{1}{n'}\right]\right)$, for $i = 1 : n'$

   we sample multiplicities of points in $\mathcal{D}^{*(k)}$

   2.2 estimate $\hat{\theta}^{*(k,b)}$ from $\mathcal{D}^{*(k,b)}$

   (fast because only n' distinct samples)

3. estimate $V^{*(k)} = \hat{\mathrm{Var}}\hat{\theta}^{*(k)}$ from $\mathcal{D}^{*(k,1:B)}$

little bootstrap

few samples, big w_i's

← from bootstrap of $\mathcal{D}^{*(k)}$

$\hat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{K}\sum_{k=1}^{K} V^{*(k)}$ ← avg Var over Bag

$B \times n$ slow for large n!

Boostrap

for $b = 1 : B$

$\times B$ Resample $\mathcal{D}^{*(b)}$ ← n

Estimate $\hat{\theta}^{*(b)}$ ← (n)

$< n$ if repetitions

BLB

$\mathcal{D}^{*(k)} = \{1, 2, \dots n'\}$ $\qquad n' \ll n$

$\frac{1}{n'} \quad \frac{1}{n'} \quad \frac{1}{n}$

$x^*_{1:n} \sim \text{unif}\{1:n'\}$

j

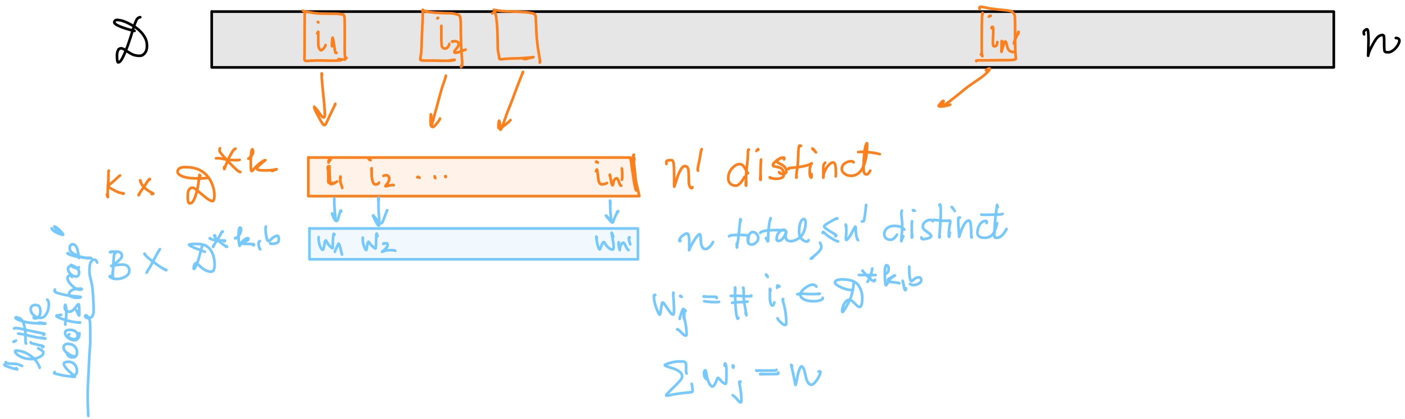
$z^*_{1:n'} = \#\{x^*_i = j\} \Leftarrow$ $\qquad \Pr[j] = \frac{1}{n}$

$(z^*_{1:n'}) \sim \text{multinomial}\left(n, \left(\frac{1}{n'}, \dots \frac{1}{n'}\right)\right)$

Ex: $\left.\begin{array}{l} n' = 6 \\ n = 100 \end{array}\right\} \to x^*_j = \text{die roll}$

$(z_j) \sim \text{multinomial}\left(n = 100, \left(\frac{1}{6} \dots \frac{1}{6}\right)\right)$

$j = 1:6$



$w_j = \# i_j \in \mathcal{D}^{*k,b}$

$\sum w_j = n$

$\hookrightarrow \text{Var}^{(k)} \hat{\theta}^* \approx V^k$

samples with weights

$\mathcal{D}^{*(k)} = \{x_1, \dots x_5\}$ $n' = 5$

Ex $\{x_1, x_5, x_1, x_3, x_2, x_2\} = \mathcal{D}^{*b}$ $n = 6$

$= \{\underbrace{x_1}_{\times 2}, \underbrace{x_2}_{\times 2}, \underbrace{x_3}_{\times 1}, \underbrace{x_5}_{\times 1}\} = \{(x_1, w_1 = 2), (x_2, w_2 = 2), (x_3, w_3 = 1), \dots\}$

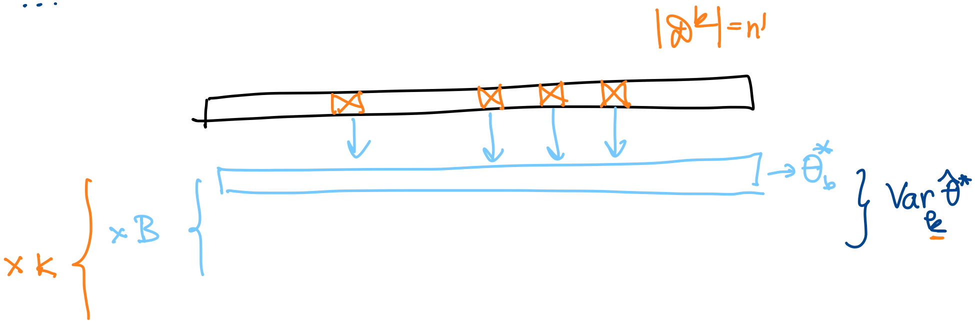
$\tilde{n} = 4$, $\tilde{\mathcal{D}} = \{(x_i, w_i)\}$

$w_i \to$ weight of x_i

$$\boxed{\sum_{i=1}^{n} w_i = n}$$

For some models, training $\sim \tilde{n}$ = # unique datapoints in $\mathcal{D}^*$

- Linear Regression
- Logistic Regression, any gradient ascent algorithm
- any algo that depends on sufficient statistics
- Decision Trees
- ...



Bag of little Boostraps

Ex $n = 10^6$

$n' = \sqrt{n} = 1000$

$(K = 1000)$

- Theoretical results
- $n' \sim \sqrt{n}$
- $K \sim \frac{n}{n'}$
- B as usual for boostrap, e.g. 50–100

$\Rightarrow K n' \sim n$

- In practice $K \approx 50$ okay
- Computation $K \times B \times n'$ when estimation algorithm can use weighted data points efficiently

3) Conformal prediction ← CI for $\hat{y}$

- **Conformal prediction:** CI for a single prediction $\hat{y} = f(x)$

Given data set $\mathcal{D} = \{(x_i, y_i), i = 1 : N\}$ ← Regression
Training algorithm $\mathcal{A}$, so that $\mathcal{A}(\mathcal{D}) = f$ the predictor e.g Linear Regression
Want CI for $\hat{y} = f(x)$ where x is a new data point
so that the CI is NOT dependent on $\mathcal{A}$ being statistically correct (e.g. $\mathcal{A}$ overfits, . . .)

- jackknife+ is a simple algorithm for CP
- More advanced algorithms exist. This is an active area of research in statistics.

$\boxed{\begin{array}{l}\mathcal{D}\\ \mathcal{A}\\ \mathcal{A}(\mathcal{D}) = f\end{array}}$

new x → $\boxed{f(x) = \hat{y}}$ prediction → CI = ? at level $1-\alpha$

$\alpha = 5\%$

!!!! Do NOT use CP to "crossvalidate" your algorithm!