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+

The jackknife

- like Leave-one-out CV

- $\mathcal{D}_{-i} = \mathcal{D} \setminus \{(x_i, y_i)\}$ or $\mathcal{D} \setminus \{x_i\}$
- θ is parameter of interest
  - $\hat{\theta}$ = estimate of θ from $\mathcal{D}$
  - $\hat{\theta}_{-i}$ = estimate of θ from $\mathcal{D}_{-i}$, $i = 1 : n$

- **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 $\mathcal{D}_{-i}$
  3. Use $\hat{F}(\hat{\theta}) \approx \hat{F}(\hat{\theta}_{-i},\ i = 1 : n)$ to estimate bias, variance, ... $\hat{\theta}$

Bag of little bootstraps [arXiv:1112.5016]

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

- **Bag of little Bootstraps Algorithm**

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

  $\hat{\text{Var}}(\hat{\theta}) = \frac{1}{K}\sum_{k=1}^{K} V^{*(k)}$

Bag of little Boostraps

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

- In practice $K \approx 50$ okay

- Computation $K \times B \times n'$ when estimation algorithm can use weighted data points efficiently

Conformal prediction

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

Given data set $\mathcal{D} = \{(x_i, y_i), i = 1 : N\}$
Training algorithm $\mathcal{A}$, so that $\mathcal{A}(\mathcal{D}) = f$ the predictor
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.

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

jackknife+

jackknife+ Algorithm

In data set $\mathcal{D} = \{(x_i, y_i), i = 1 : n\}$
Training algorithm $\mathcal{A}$, so that $\mathcal{A}(\mathcal{D}) = f$ the predictor
Confidence level $1 - \alpha$

Want CI for $\hat{y} = f(x)$ where x is a new data point

1. Precompute $f_{-i} \leftarrow \mathcal{A}(\mathcal{D}_{-i})$ for $i = 1 : n$
2. Compute "leave one out" residuals $R_i = |y_i - f_{-i}(x_i)|$, for $i = 1 : n$
3. For every new x: compute $f(x)$, then get $1 - \alpha$ Prediction Interval $[a, b]$ for $f(x)$ by
   - 3.1 Compute lower bounds $a_i = f_{-i}(x) - R_i$, for $i = 1 : n$
   - 3.2 Sort $a_{1:n}$
   - 3.3 Set $a \leftarrow \lfloor \frac{\alpha}{2} n \rfloor$ quantile of $a_{1:n}$
   - 3.4 Compute upper bounds $b_i = f_{-i}(x) + R_i$, for $i = 1 : n$
   - 3.5 Sort $b_{1:n}$
   - 3.6 Set $b \leftarrow \lceil \left(1 - \frac{\alpha}{2}\right) n \rceil$ quantile of $b_{1:n}$
   - 3.7 Output $CI^{\alpha} = [a, b]$
4. **Theorem** $P[y(x) \in CI^{\alpha}] \geq 1 - \alpha$

jackknife+