25. Even More Model Selection (Bootstrap, CV, GCV)

Sunday, March 10, 2024 5:25 PM

Technique #9: Bootstrap Resampling

← idea: imitate a validation set

Bootstrap: Given dataset $S = (z_i)_{i=1}^m$ and we have a statistic γ, $\gamma(S)$, and we want to estimate distribution of γ (especially the variance)

for $b = 1, \dots, B$

Draw $m' \leq m$ samples from S, with replacement. Call this $S^{(b)}$

Compute $\gamma(S^{(b)})$

Then $\widehat{\mathrm{Var}(\gamma(S))} = \frac{1}{B-1} \sum_{b=1}^{B} \left(\gamma(S^{(b)}) - \underbrace{\hat{\mathbb{E}}\gamma(S)}_{\text{sample mean}} \right)^2$

turns one dataset into many! We're getting something from nothing, ie. lifting ourselves by our bootstraps?!

Too good to be true? Sort of, not completely.

To use for a criterion to estimate true risk, there's some subtlety on how to do it (otherwise it's like a worse version of cross-validation)

and a 0.632 heuristic. See Hastie, Tibshirani et al. § 7.11

25a. Even More Model Selection (Bootstrap, CV, GCV)a

Sunday, March 10, 2024 5:34 PM

Technique #10: Cross-Validation (CV)

Recall we already talked about (plain) validation

(train/validation/test split, Hoeffding ...)

⚠ Don't forget to adjust Hoeffding with $\delta \to \delta/k$ if you apply to k models simultaneously. (union bound)

known as the Bonferroni Correction, i.e. divide your p-value by # of regressors

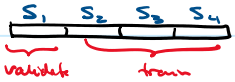
otherwise you're p-hacking!

(if use 0.05 significance level but run 20 independent test w/ chance results, then 64% chance at least 1 is declared significant)

k-fold CV

Better than plain validation if you don't have much data... but unnecessary if $m \geq 10^5$ or so

Split data into k-"folds"



$|S_i| = m/k$

$k = 4$

For $i = 1, \ldots, k$] step 1

Train h_i based on training data $S \setminus S_i$

"Validate" h_i using S_i, estimate empirical risk

Return avg. of estimates, $\frac{1}{k}\sum \hat{L}_{S_i}(h_i)$

For a fixed i, this is just plain validation since $h_i \perp S_i$

but since we average over i, ① analysis not easy

② ... but works better than a single plain validation

Typically $k = 5$ or 10

$k = 1$ doesn't work!

$K = m$ is aka "Leave-one-out" CV ... computationally expensive! ~ m times more expensive than a single train.

LOO

See §7.10.2 Hastie, Tibshirani et al.

for Right and Wrong ways to do CV

i.e. Feature Selection should be part of CV loop!

Use nice automated CV software frameworks (Scikit-Learn....) to prevent other kinds of bugs.

25b. Even More Model Selection (Bootstrap, CV, GCV)

Sunday, March 10, 2024 5:52 PM

Technique #11: Generalized CV (GCV)

For Ordinary Least Squares, this is a computationally efficient approximation to LOO C.V.

Take linear models $\hat{y} = Py$, P usually parameterized, depends on data X and parameter α

Define $\hat{y}_i$ = i^{th} output, aka $\hat{y}_i = h_S(x_i)$ (estimated model)

S = dataset $= \{x_i\}_{i=1}^m$

and $\hat{y}_i^{(-i)}$ = i^{th} output if trained on $S^{(-i)} = S \setminus \{x_i\}$

$= h_{S^{(-i)}}(x_i)$

$x_i \in \mathbb{R}^d$

For some models (OLS, Tikhonov, cubic smoothing splines)

then $\hat{y}_i^{(-i)} - y_i \overset{(*)}{=} \frac{\hat{y}_i - y_i}{1 - P_{ii}}$ $\quad P_{ij} = (i,j)^{th}$ entry of P

So LOO CV is

$$CV_{LOO}(\alpha) = \frac{1}{m}\sum_{i=1}^m \left(\hat{y}_i^{(-i)} - y_i\right)^2 = \frac{1}{m}\sum_{i=1}^m \left(\frac{\hat{y}_i - y_i}{1 - P_{ii}}\right)^2$$

Slow! Solve m OLS problems

Fast! Solve 1 OLS problem

Recall $P = P(\alpha)$

and GCV simplifies even further:

$$GCV(\alpha) = \frac{1}{m}\sum_{i=1}^m \left(\frac{\hat{y}_i - y_i}{1 - \mathrm{tr}(P)/d}\right)^2$$

$$= \underbrace{\frac{1}{(1 - \mathrm{tr}(P)/d)^2}}_{\text{penalize large effective dim.}} \cdot \underbrace{\frac{1}{m}\sum_{i=1}^m (\hat{y}_i - y_i)^2}_{\text{ERM}}$$

For Tikhonov

Recall from UPRE section

$$\mathrm{tr}(P_\alpha) = d - \sum_{i=1}^d \frac{\alpha}{\alpha + \lambda_i(X^TX)} \quad \begin{cases} \to 0 & \text{as } \alpha \to \infty \\ \to d & \text{as } \alpha \to 0 \end{cases}$$

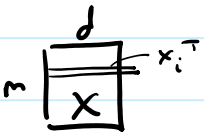
25c. Even More Model Selection (Bootstrap, CV, GCV)

Friday, March 22, 2024 3:04 PM

Prove:

$$\hat{y}_i^{(-i)} - y_i \overset{(*)}{=} \frac{\hat{y}_i - y_i}{1 - P_{ii}}$$

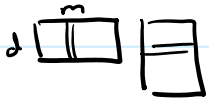
For Tikhonov:



$$\hat{y} = X \cdot \hat{w}, \qquad \hat{w} = \underbrace{(X^TX + \alpha I)^{-1}}_{G} X^T y$$

so $P = XGX^T$, $P_{ij} = x_i^T G x_j$

$$XX^T = \sum_{i=1}^{m} x_i x_i^T$$



so... $\hat{w}^{(-i)} = \underbrace{(X^TX - x_i x_i^T + \alpha I)^{-1}}_{G^{(-i)}} \cdot (X^T y - x_i y_i)$

Sherman-Morrison-Woodbury matrix inversion lemma

$$(A + UCV^T)^{-1} = A^{-1} - A^{-1}U(C^{-1} + V^TA^{-1}U)^{-1}V^TA^{-1}$$

so ($U = -x_i$, $C = 1$, $V = x_i$)

$$G^{(-i)} = (X^TX + \alpha I - x_i x_i^T)^{-1} = G + \frac{1}{\underbrace{1 - x_i^T G x_i}_{P_{ii}}} \cdot G x_i x_i^T G$$

hence

$$\hat{w}^{(-i)} = G^{(-i)}(X^T y - x_i y_i) = \left(G + \frac{1}{1-P_{ii}} G x_i x_i^T G\right)\cdot(X^T y - x_i y_i)$$

$$= GX^T y - G x_i y_i + \frac{1}{1-P_{ii}} G x_i x_i^T G X^T y - \frac{1}{1-P_{ii}} G x_i x_i^T G x_i y_i$$

$$= \hat{w} + G x_i \left(-y_i + \frac{1}{1-P_{ii}} \underbrace{x_i^T G X^T y}_{\hat{y}_i} - \frac{1}{1-P_{ii}} \underbrace{x_i^T G x_i}_{P_{ii}} y_i\right)$$

$$= \hat{w} + G x_i \left(\frac{-y_i + P_{ii}\hat{y}_i}{1-P_{ii}} + \frac{1}{1-P_{ii}}\hat{y}_i - \frac{1}{1-P_{ii}} P_{ii} y_i\right)$$

$$= \hat{w} + G x_i \cdot \frac{(\hat{y}_i - y_i)}{1-P_{ii}}$$

so

$$\hat{y}_i^{(-i)} = x_i^T \hat{w}^{(-i)} = x_i^T \hat{w} + \underbrace{x_i^T G x_i}_{P_{ii}} \frac{(\hat{y}_i - y_i)}{1-P_{ii}} = \hat{y}_i + \frac{P_{ii}}{1-P_{ii}}(\hat{y}_i - y_i)$$

so $y_i^{(-i)} - y_i = \hat{y}_i + \frac{P_{ii}}{1-P_{ii}}(\hat{y}_i - y_i) - y_i = \frac{1}{1-P_{ii}}\left(\hat{y}_i - P_{ii}\hat{y}_i + P_{ii}(\hat{y}_i - y_i) - y_i + P_{ii} y_i\right)$

$$= \frac{1}{1-P_{ii}}(\hat{y}_i - y_i) \qquad \square$$

ASIDE: $(A + \varepsilon B)^{-1}$

(1) Is inverse continuous? Yes, if bounded away from being singular, eg. plot x^{-1} cts on $\mathbb{R} \setminus (-\varepsilon, \varepsilon)$

(2) How to compute? (20) Sherman-Morrison-Woodbury

(26) Neumann-Series

25d. Even More Model Selection (Bootstrap, CV, GCV)

Friday, March 22, 2024 3:27 PM

Neumann Series

Assume A^{-1} is reference (i.e. A nonsingular)

$(A-\varepsilon B)^{-1}$ → ① convert to nicer form

$$(A-\varepsilon B)x = y \iff A^{-1}(A-\varepsilon B)x = A^{-1}y$$

$$\Rightarrow I - \varepsilon \underbrace{A^{-1}B}_{\tilde{B}} x = \underbrace{A^{-1}y}_{\tilde{y}}$$

or, if need to preserve symmetry,

$$I - \varepsilon A^{-1/2} B \underbrace{A^{-1/2}\tilde{x}}_{x = A^{-1/2}\tilde{x}} = A^{-1/2}y$$

② so wlog, look at perturbations of identity

$(I-\varepsilon B)^{-1}$

Scalar case: $\frac{1}{1-x} = 1 + x + x^2 + x^3 + \dots$ if $|x| < 1$

(i.e. $\sum_{k=0}^{n} r^k = \frac{1-r^{n+1}}{1-r} \xrightarrow{n\to\infty} \frac{1}{1-r}$ Calc 1)

by analogy

$$(I-\varepsilon B)^{-1} = \underbrace{\sum_{k=0}^{\infty} (\varepsilon B)^k}_{\text{Neumann Series}}$$

Converges (in operator norm), even in Banach space!, if

$\|\varepsilon B\| < 1$ i.e. $\varepsilon < 1/\|B\|$

↑ Spectral (operator) norm, not Euclidean / Frobenius

$$\|B\| = \sup_{\|x\|_2 = 1} \|Bx\|_2$$

= max singular value