Ch 5.1.3-4: *k*-Fold Cross-Validation

Lecture 13 - CMSE 381

Prof. Elizabeth Munch

Michigan State University
::
Dept of Computational Mathematics, Science & Engineering

Wed, Oct 4, 2023

Announcements

Last time:

- Validation Set
- LOOCV

Announcements:

-
-

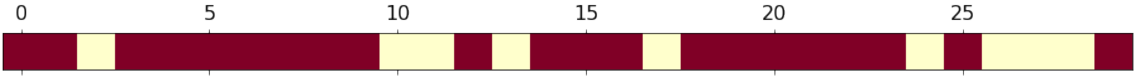
Covered in this lecture

- k-fold CV

Section 1

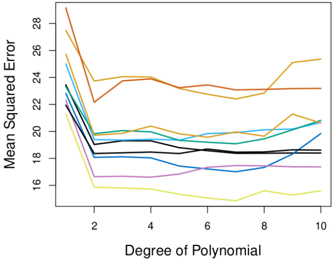
Last time

Validation set approach

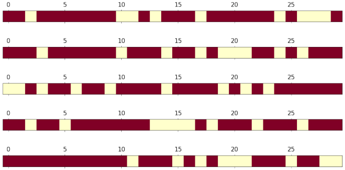


- Divide randomly into two parts:
  - Training set
  - Validation/Hold-out/Testing set
- Fit model on training set
- Use fitted model to predict response for observations in the test set
- Evaluate quality (e.g. MSE)

Problems



Ex. Predict `mpg` using `horsepower`



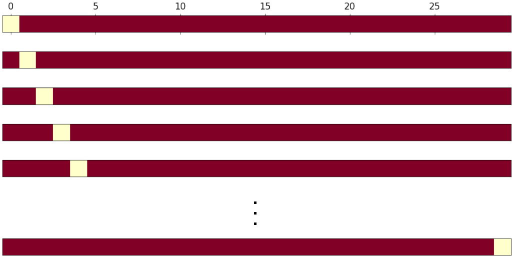
- Highly variable results, no consensus about the error
- Tends to overestimate test error rate

Leave One Out CV (LOOCV)

- Remove (x_1, y_1) for testing.
- Train the model on $n-1$ points: $\{(x_2, y_2), \cdots, (x_n, y_n)\}$
- Calculate $\mathrm{MSE}_1 = (y_1 - \hat{y}_1)^2$

- Remove (x_2, y_2) for testing.
- Train the model on $n-1$ points: $\{(x_1, y_1), (x_3, y_3), \cdots, (x_n, y_n)\}$
- Calculate $\mathrm{MSE}_2 = (y_2 - \hat{y}_2)^2$

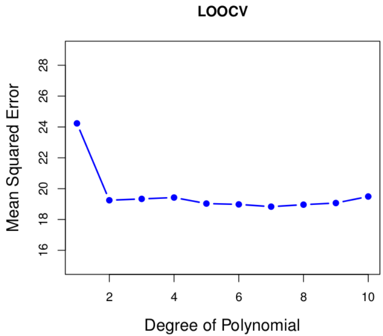
- Rinse and repeat



Return the score:

$$CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{MSE}_i$$

Pros and Cons



- No variance
- Higher computation cost

Speeding up LOOCV

Warning: This only works for least squares linear or polynomial regression.

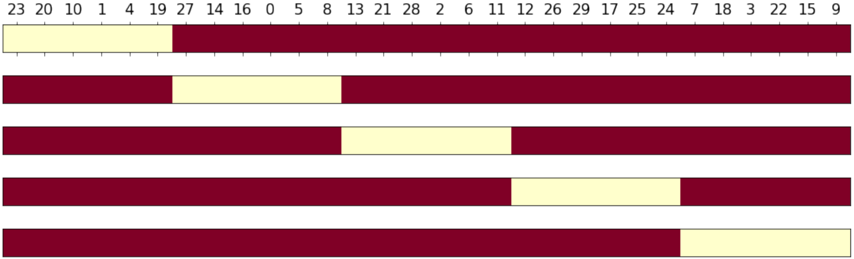
$$h_i = \frac{1}{n} + \frac{(x_i - \overline{x})^2}{\sum_{j=1}^{n}(x_j - \overline{x})^2}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathrm{MSE}_i = CV_{(n)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - \hat{y}_i}{1 - h_i}\right)^2$$

Section 2

k-Fold CV

The idea


23 20 10 1 4 19 27 14 16 0 5 8 13 21 28 2 6 11 12 26 29 17 25 24 7 18 3 22 15 9

Mathy version

- Randomly split data into k-groups (folds)
- Approximately equal sized. For the sake of notation, say each set has ℓ points
- Remove ith fold U_i and reserve for testing.
- Train the model on remaining points
- Calculate $\mathrm{MSE}_i = \frac{1}{\ell} \sum_{(x_j,y_j)\in U_i} (y_j - \hat{y}_j)^2$
- Rinse and repeat

Return

$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{MSE}_i$$

By hand first!

There are 10 students in the class, and we have data points for each. They have already been randomly permuted below. Write down the training/testing sets for a 3-fold CV

- Damien
- Alice
- Greta
- Jasmin
- Benji
- Inigo
- Firas
- Carina
- Enrique
- Hubert

Fold 1

Fold 2

Fold 3

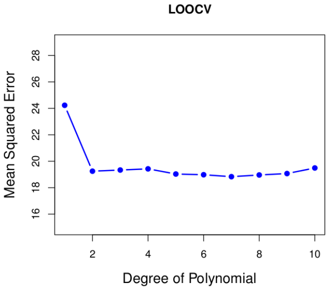
Coding - Building *k*-fold CV

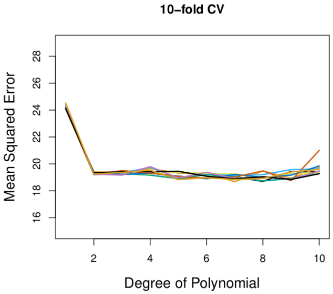
Pros and Cons

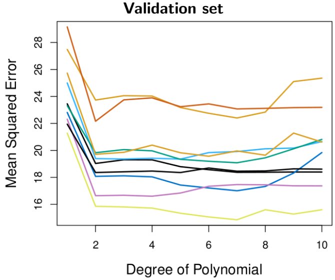
Pros:

Cons:

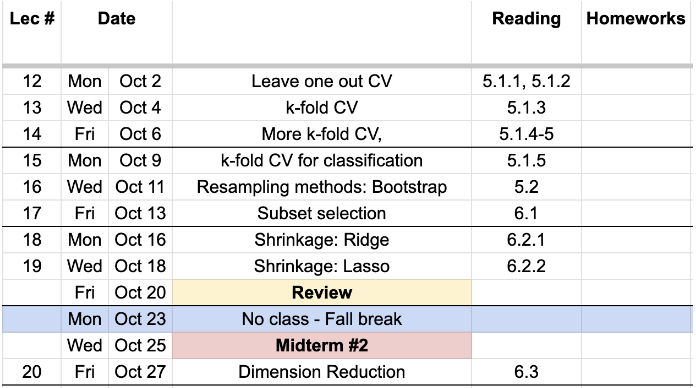
Comparison


LOOCV


10–fold CV


Validation set

Next time



Lec #	Date			Reading	Homeworks
12	Mon	Oct 2	Leave one out CV	5.1.1, 5.1.2	
13	Wed	Oct 4	k-fold CV	5.1.3	
14	Fri	Oct 6	More k-fold CV,	5.1.4-5	
15	Mon	Oct 9	k-fold CV for classification	5.1.5	
16	Wed	Oct 11	Resampling methods: Bootstrap	5.2	
17	Fri	Oct 13	Subset selection	6.1	
18	Mon	Oct 16	Shrinkage: Ridge	6.2.1	
19	Wed	Oct 18	Shrinkage: Lasso	6.2.2	
	Fri	Oct 20	**Review**		
	Mon	Oct 23	No class - Fall break		
	Wed	Oct 25	**Midterm #2**		
20	Fri	Oct 27	Dimension Reduction	6.3	