

Ch 5.1.3-4: k -Fold Cross-Validation

Lecture 13 - CMSE 381

Prof. Guanqun Cao

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Mon, Sep 29, 2025

Announcements

Last time:

- Validation Set
- LOOCV

CMSE381_F2025_Schedule : Schedule

	M	9/22	Project Day & Review		
	W	9/24	Midterm #1		
12	F	9/26	Leave one out CV	5.1.1, 5.1.2	
13	M	9/29	k-fold CV	5.1.3	
14	W	10/1	More k-fold CV	5.1.4-5	
15	F	10/3	k-fold CV for classification	5.1.5	
16	M	10/6	Subset selection	6.1	
17	W	10/8	Shrinkage: Ridge	6.2.1	
18	F	10/10	Shrinkage: Lasso	6.2.2	HW #4 Due Sun 10/12
19	M	10/13	PCA	6.3	
20	W	10/15	PCR	6.3	
	F	10/17	Review		
	M	10/20	Fall Break		
	W	10/22	Midterm #2		
21	F	10/24	Polynomial & Step Functions	7.1-7.2	HW #5 Due Sun 10/28
22	M	10/27	Step Functions; Basis functions; Start Splines	7.2-7.4	
23	W	10/29	Regression Splines	7.4	

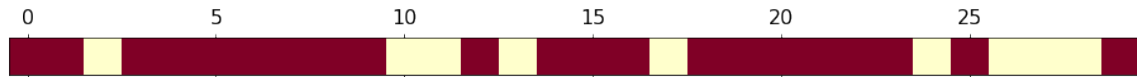
Covered in this lecture

- k -fold CV

Section 1

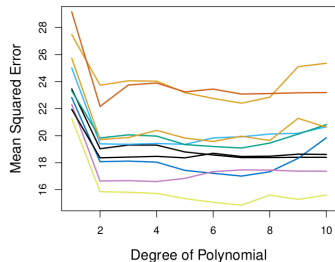
Last time

Validation set approach

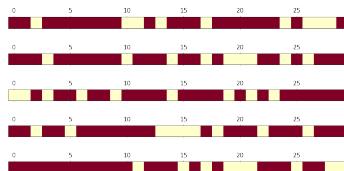


- Divide randomly into two parts:
 - ▶ Training set
 - ▶ Validation/Hold-out/Testing set
- Fit model on training set
- Use fitted model to predict response for observations in the test set
- Evaluate quality (e.g. MSE)

Problems



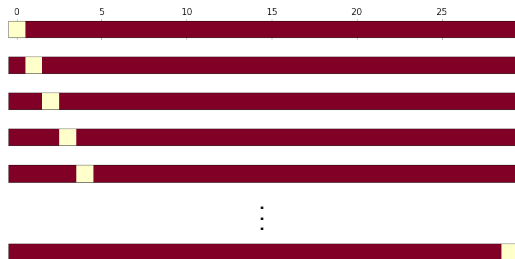
Ex. Predict mpg using
horsepower



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

Leave One Out CV (LOOCV)

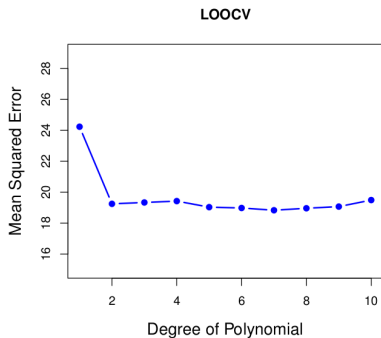
- Remove (x_1, y_1) for testing.
- Train the model on $n - 1$ points:
 $\{(x_2, y_2), \dots, (x_n, y_n)\}$
- Calculate $\text{MSE}_1 = (y_1 - \hat{y}_1)^2$
- Remove (x_2, y_2) for testing.
- Train the model on $n - 1$ points:
 $\{(x_1, y_1), (x_3, y_3), \dots, (x_n, y_n)\}$
- Calculate $\text{MSE}_2 = (y_2 - \hat{y}_2)^2$
- Rinse and repeat



Return the score:

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

Pros and Cons

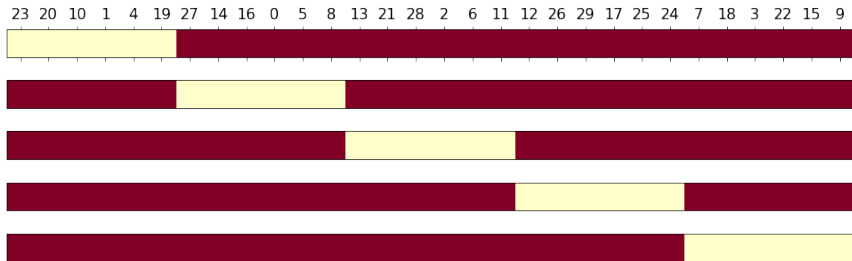


- No variance
- Higher computation cost

Section 2

k -Fold CV

The idea



Mathy version

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

Return

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

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

	Fold 1	Fold 2	Fold 3
• Damien			
• Alice			
• Greta			
• Jasmin			
• Benji			
• Inigo			
• Firas			
• Carina			
• Enrique			
• Hubert			

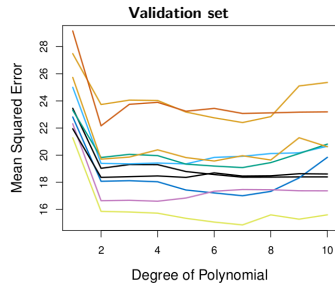
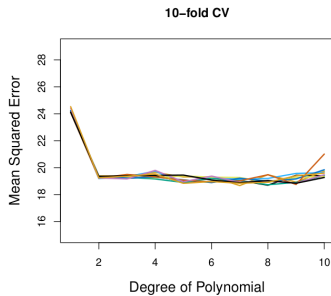
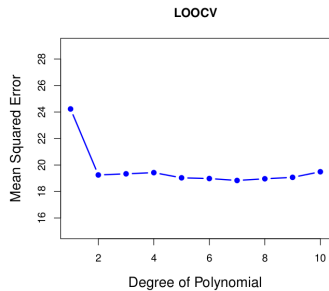
Coding - Building k -fold CV

Pros and Cons

Pros:

Cons:

Comparison



Next time

CMSE381_F2025_Schedule : Schedule

	M	9/22	Project Day & Review		
	W	9/24	Midterm #1		
12	F	9/26	Leave one out CV	5.1.1, 5.1.2	
13	M	9/29	k-fold CV	5.1.3	
14	W	10/1	More k-fold CV	5.1.4-5	
15	F	10/3	k-fold CV for classification	5.1.5	
16	M	10/6	Subset selection	6.1	
17	W	10/8	Shrinkage: Ridge	6.2.1	
18	F	10/10	Shrinkage: Lasso	6.2.2	HW #4 Due Sun 10/12
19	M	10/13	PCA	6.3	
20	W	10/15	PCR	6.3	
	F	10/17	Review		
	M	10/20	Fall Break		
	W	10/22	Midterm #2		
21	F	10/24	Polynomial & Step Functions	7.1-7.2	HW #5 Due Sun 10/28
22	M	10/27	Step Functions; Basis functions; Start Splines	7.2-7.4	
23	W	10/29	Regression Splines	7.4	