

Ch 5.1.5: k -fold Cross-Validation for Classification

Lecture 15 - CMSE 381

Prof. Mengsen Zhang

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Fri, Oct 3, 2025

Last time:

- k-fold CV

This lecture:

- CV for classification

Announcements:

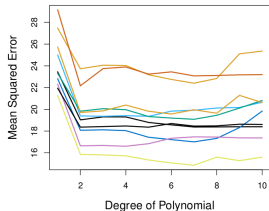
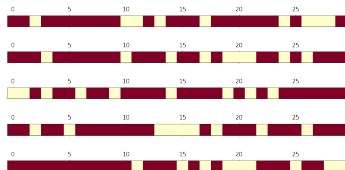
- Homework #4 is posted, Due Sunday (10/12)

Section 1

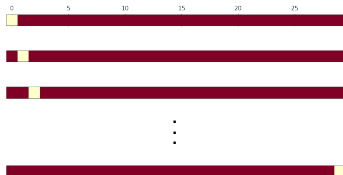
Last time

Approximations of Test Error

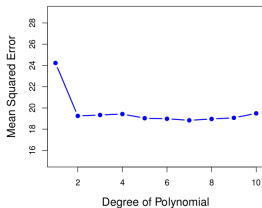
Validation Set



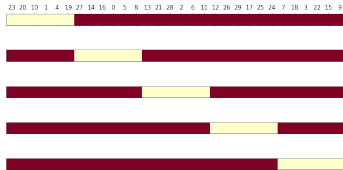
LOOCV



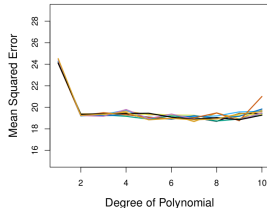
LOOCV



K-fold CV

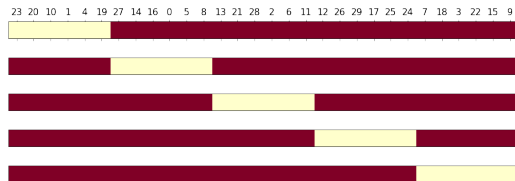


10-fold CV



Definition of k -fold CV

- 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

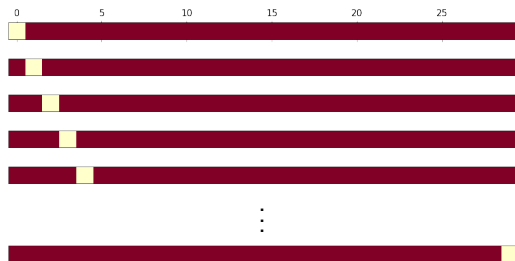
$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k \text{MSE}_i$$

Section 2

CV for Classification

Setup: LOOCV

- Remove i th point (x_i, y_i) and reserve for testing.
- Train the model on remaining points
- Calculate $\text{Err}_i = I(y_i \neq \hat{y}_i)$
- Rinse and repeat

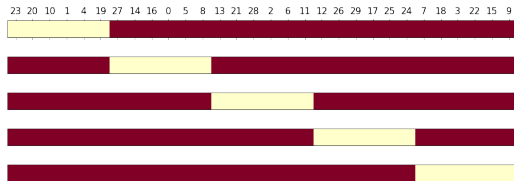


Return

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

Setup: k -fold

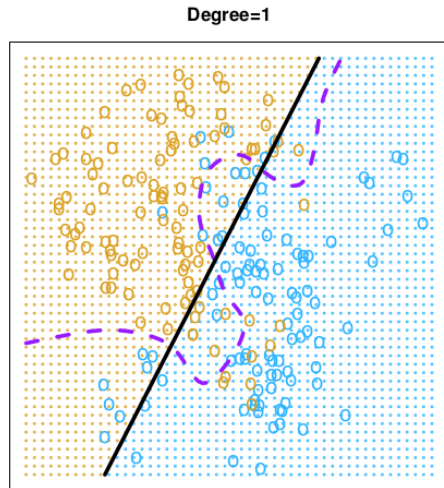
- Randomly split data into k -groups (folds)
- Approximately equal sized. For the sake of notation, say each set has ℓ points
- Remove i th fold U_i and reserve for testing.
- Train the model on remaining points
- Calculate
$$\text{Err}_i = \frac{1}{\ell} \sum_{(x_j, y_j) \in U_i} \mathbb{I}(y_j \neq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



Return

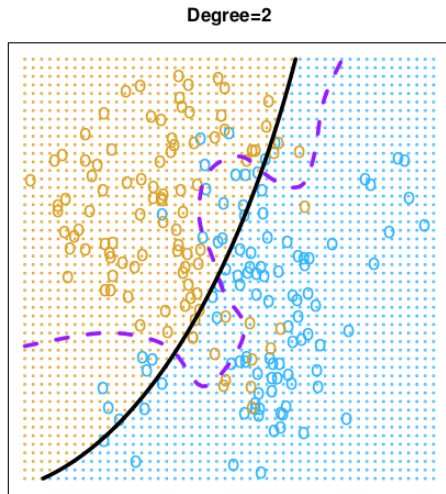
$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k \text{Err}_i$$

Example on simulated data: Linear



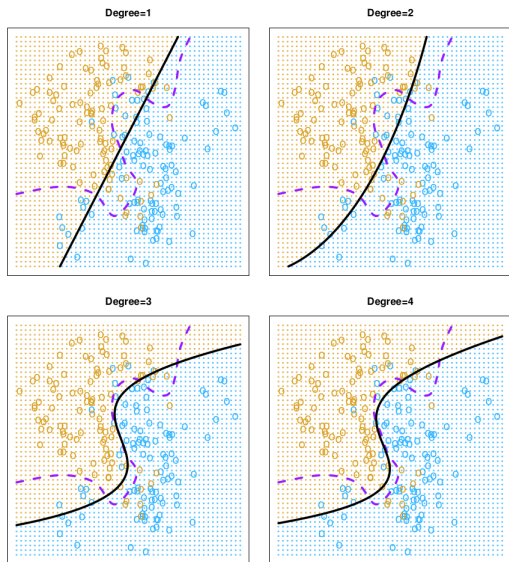
- Purple: Bayes decision boundary.
 - ▶ Error rate: 0.133
- Black: Logistic regression
 - ▶ $\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
 - ▶ Error rate: 0.201

Example on simulated data: Quadratic logistic regression



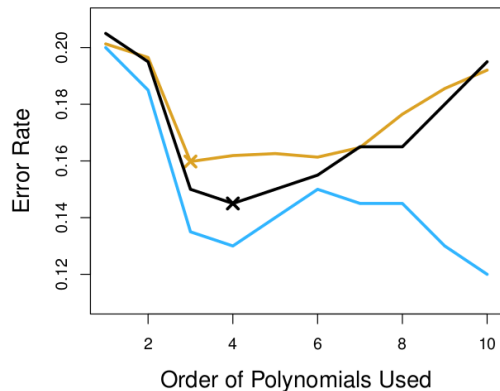
- Purple: Bayes decision boundary.
 - ▶ Error rate: 0.133
- Black: Logistic regression
 - ▶ $\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$
 - ▶ Error rate: 0.197

Example on simulated data: all the polynomials!



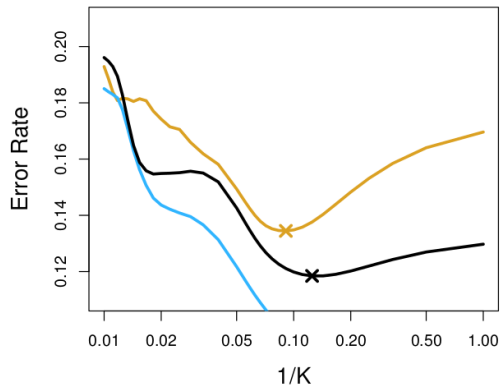
- Purple: Bayes decision boundary.
 - ▶ Error rate: 0.133
- Black: Logistic regression
 - ▶ Deg 1 Error rate: 0.201
 - ▶ Deg 2 Error rate: 0.197
 - ▶ Deg 3 Error rate: 0.160
 - ▶ Deg 4 Error rate: 0.162

Decide degree based on CV



- Test error (brown)
- Training error (blue)
- 10-fold CV error (black)

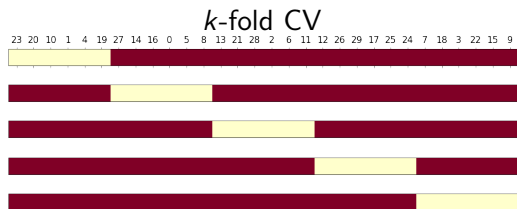
Similar game for KNN



- Test error (brown)
- Training error (blue)
- 10-fold CV error (black)

Test your understanding: [PollEv](#)

Coding - k-fold for penguin classification section



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

Use $k = 5$ or 10 usually

k-fold CV for classification

$$\text{Err}_i = \mathbb{I}(y_j \neq \hat{y}_j)$$

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

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	