

Ch 5.1.5: k -fold Cross-Validation for Classification

Lecture 15 - CMSE 381

Prof. Elizabeth Munch

Michigan State University

::

Dept of Computational Mathematics, Science & Engineering

Mon, Oct 7, 2024

Last time:

- k-fold CV

This lecture:

- CV for classification

Announcements:

- Homework #4 is posted, Due Wednesday
- Grades

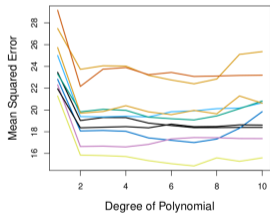
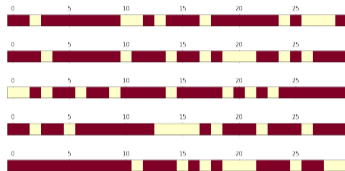
Percent	Convert
≥ 90%	4.0
≥ 85%	3.5
≥ 80%	3
≥ 75%	2.5
≥ 70%	2
≥ 65%	1.5
≥ 60%	1
< 60%	0

Section 1

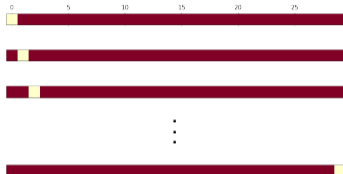
Last time

Approximations of Test Error

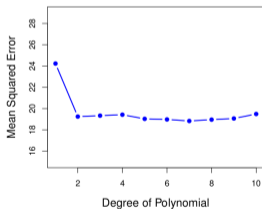
Validation Set



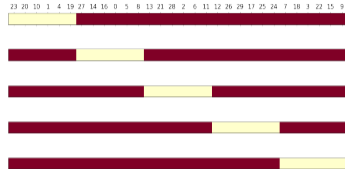
LOOCV



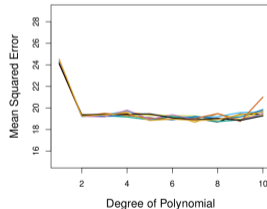
LOOCV



K-fold CV

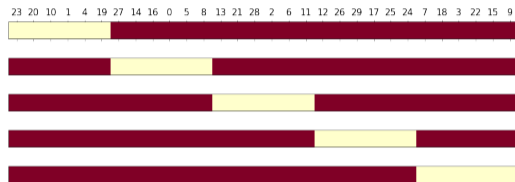


10-fold CV



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

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

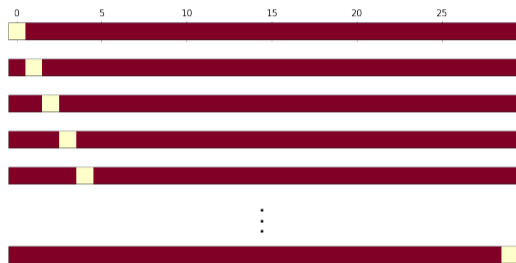
Section 2

CV for Classification

Setup: LOOCV

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

- Rinse and repeat

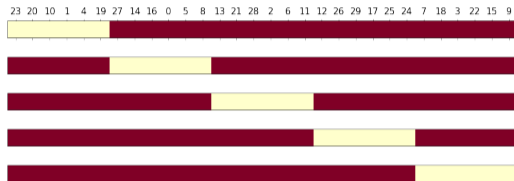


Return

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

Setup: k -fold

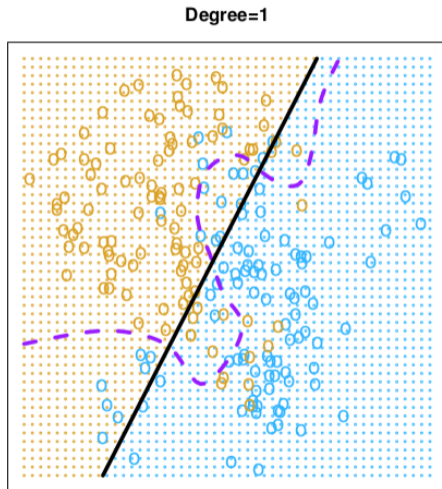
- Randomly split data into k -groups (folds)
- Approximately equal sized. For the sake of notation, say each set has ℓ points
- Remove i th fold U_i and reserve for testing.
- Train the model on remaining points
- Calculate
$$\text{Err}_i = \frac{1}{\ell} \sum_{(x_j, y_j) \in U_i} \mathbf{I}(y_j \neq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



Return

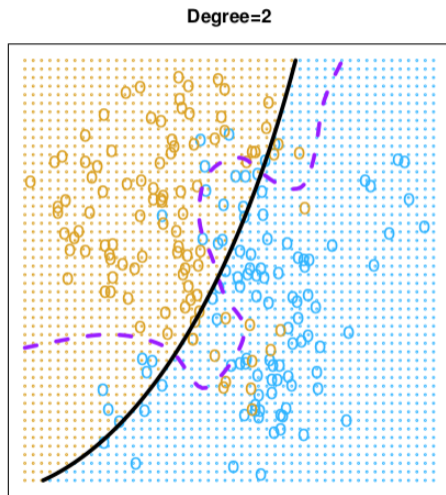
$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^k \text{Err}_i$$

Example on simulated data: Linear



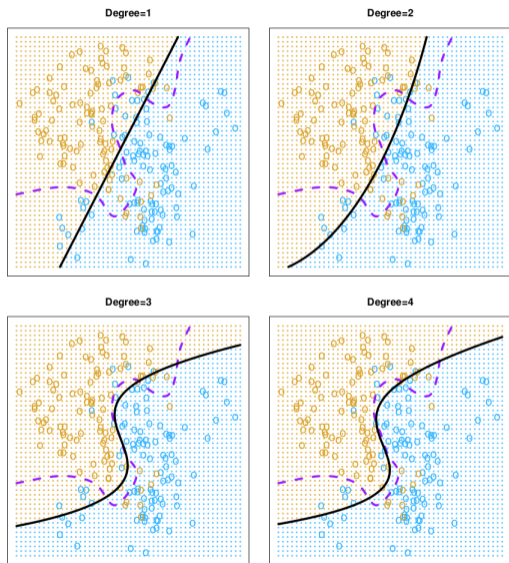
- Purple: Bayes decision boundary.
 - ▶ Error rate: 0.133
- Black: Logistic regression
 - ▶ $\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
 - ▶ Error rate: 0.201

Example on simulated data: Quadratic logistic regression



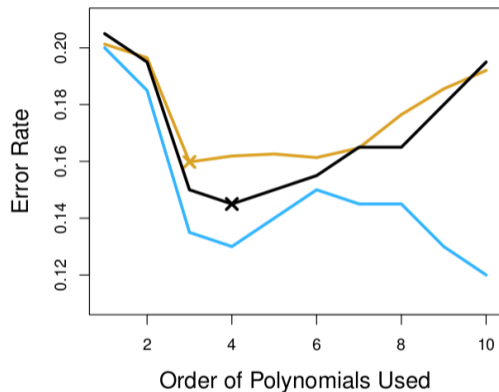
- Purple: Bayes decision boundary.
 - ▶ Error rate: 0.133
- Black: Logistic regression
 - ▶ $\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$
 - ▶ Error rate: 0.197

Example on simulated data: all the polynomials!



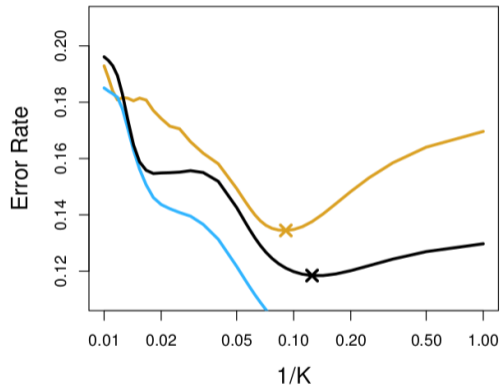
- Purple: Bayes decision boundary.
 - ▶ Error rate: 0.133
- Black: Logistic regression
 - ▶ Deg 1 Error rate: 0.201
 - ▶ Deg 2 Error rate: 0.197
 - ▶ Deg 3 Error rate: 0.160
 - ▶ Deg 4 Error rate: 0.162

Decide degree based on CV



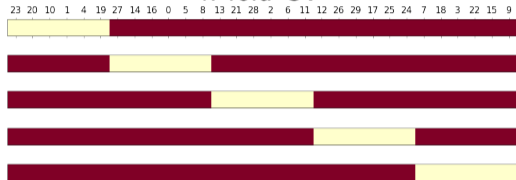
- Test error (brown)
- Training error (blue)
- 10-fold CV error (black)

Similar game for KNN



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

Coding - k-fold for penguin classification section

k -fold CV

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

Use $k = 5$ or 10 usually

k -fold CV for classification

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

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

Next time

Lec #	Date		Reading	HW	
12	Mon	9/30	Leave one out CV	5.1.1, 5.1.2	
13	Wed	10/2	k-fold CV	5.1.3	
14	Fri	10/4	More k-fold CV,	5.1.4-5	
15	Mon	10/7	k-fold CV for classification	5.1.5	
16	Wed	10/9	Resampling methods: Bootstrap	5.2	HW #4 Due Weds 10/9
17	Fri	10/11	Subset selection	6.1	
18	Mon	10/14	Shrinkage: Ridge	6.2.1	
19	Wed	10/16	Shrinkage: Lasso	6.2.2	
20	Fri	10/18	Dimension Reduction	6.3	HW #5 Due Fri 10/18
	Mon	10/21	No class - Fall break		
	Wed	10/23	Review		
	Fri	10/25	Midterm #2		