Ch 5.1.5: k-fold Cross-Validation for Classification

Lecture 15 - CMSE 381

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

Fri, Feb 20, 2026

Announcements

Last time:

- k-fold CV

This lecture:

- CV for classification

Announcements:

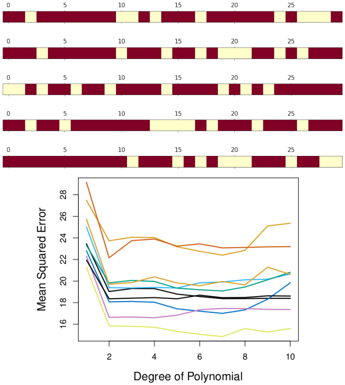
- Homework #3 is due Sunday 3/8

Section 1

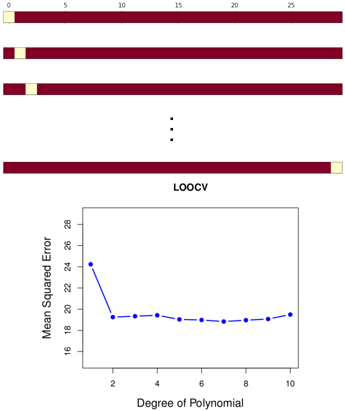
Last time

Approximations of Test Error

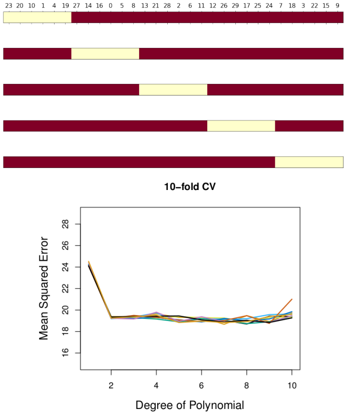
Validation Set



LOOCV

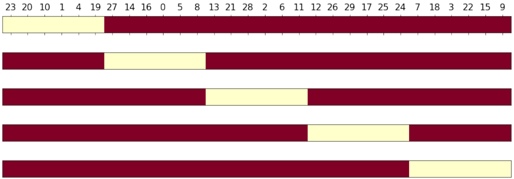


K-fold CV



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 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

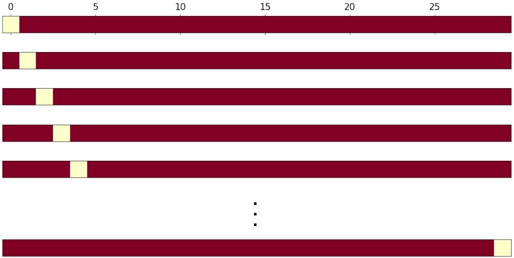
$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{MSE}_i$$

Section 2

CV for Classification

Setup: LOOCV

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

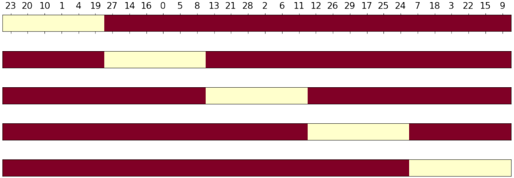


Return

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

Setup: k-fold

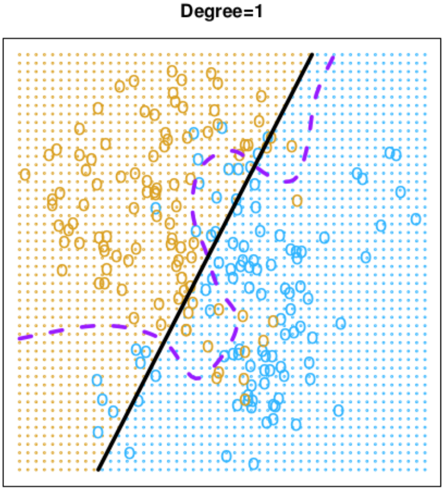
- Randomly split data into k-groups (folds)
- Approximately equal sized. For the sake of notation, say each set has ℓ points
- Remove ith fold U_i and reserve for testing.
- Train the model on remaining points
- Calculate $\mathrm{Err}_i = \frac{1}{\ell} \sum_{(x_j,y_j)\in U_i} \mathrm{I}(y_j \neq \hat{y}_j)$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



Return

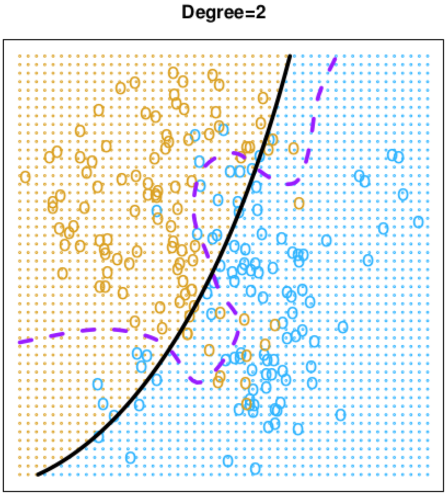
$$CV_{(k)} = \frac{1}{k} \sum_{i=1}^{k} \mathrm{Err}_i$$

Example on simulated data: Linear



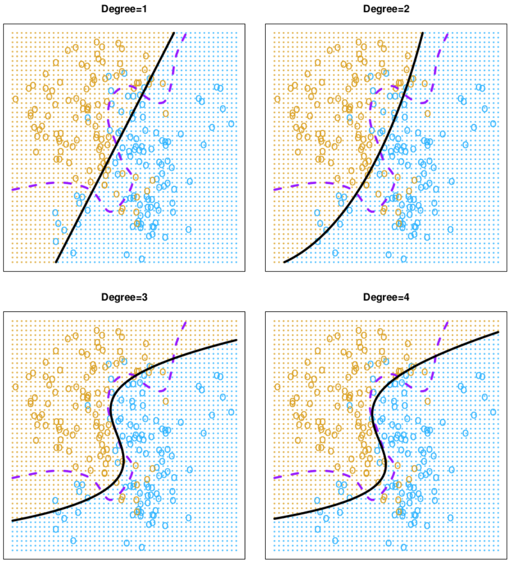
- Purple: Bayes decision boundary.
  - Error rate: 0.133
- Black: Logistic regression
  - $\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - Error rate: 0.201

Example on simulated data: Quadratic logistic regression



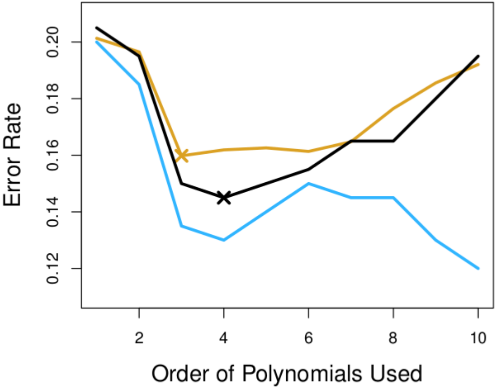
- Purple: Bayes decision boundary.
  - Error rate: 0.133
- Black: Logistic regression
  - $\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$
  - Error rate: 0.197

Example on simulated data: all the polynomials!



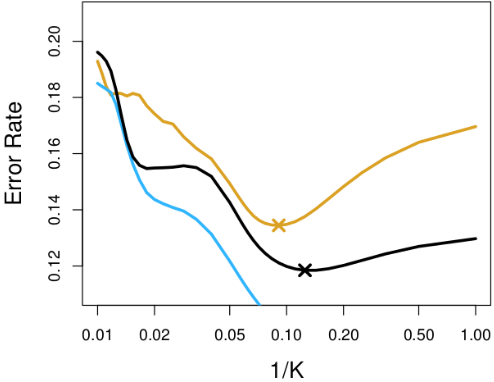
- Purple: Bayes decision boundary.
    - Error rate: 0.133
- Black: Logistic regression
    - Deg 1 Error rate: 0.201
    - Deg 2 Error rate: 0.197
    - Deg 3 Error rate: 0.160
    - Deg 4 Error rate: 0.162

Decide degree based on CV



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

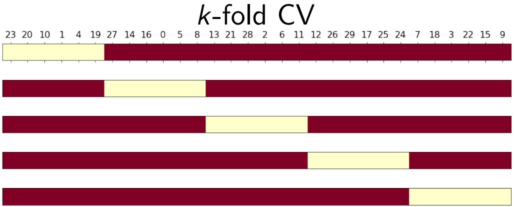
Similar game for KNN



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

Coding - k-fold for penguin classification section

TL;DR



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

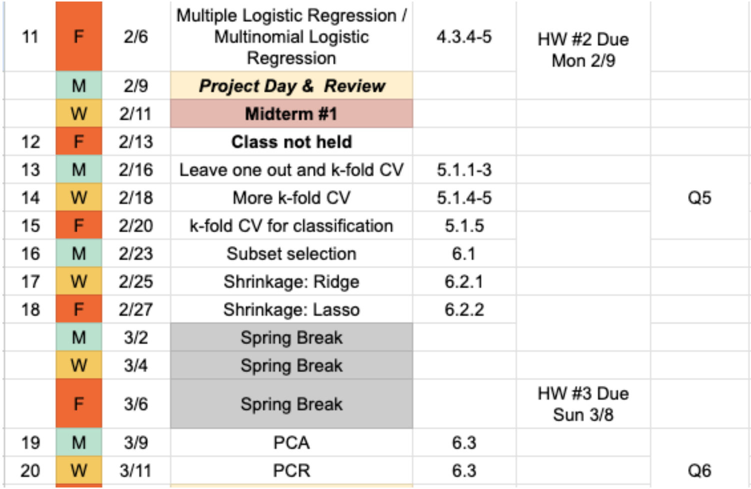
Use $k = 5$ or 10 usually

k-fold CV for classification

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

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

Next time



11	F	2/6	Multiple Logistic Regression / Multinomial Logistic Regression	4.3.4-5	HW #2 Due Mon 2/9	
	M	2/9	***Project Day & Review***			
	W	2/11	**Midterm #1**			
12	F	2/13	**Class not held**			
13	M	2/16	Leave one out and k-fold CV	5.1.1-3		Q5
14	W	2/18	More k-fold CV	5.1.4-5		
15	F	2/20	k-fold CV for classification	5.1.5		
16	M	2/23	Subset selection	6.1		
17	W	2/25	Shrinkage: Ridge	6.2.1		
18	F	2/27	Shrinkage: Lasso	6.2.2		
	M	3/2	Spring Break			
	W	3/4	Spring Break			
	F	3/6	Spring Break		HW #3 Due Sun 3/8	
19	M	3/9	PCA	6.3		Q6
20	W	3/11	PCR	6.3		