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, Feb 21, 2025

Announcements

Last time:

- k-fold CV

This lecture:

- CV for classification

Announcements:

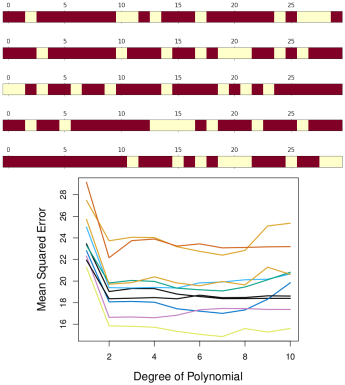
- Homework #4 is posted, Due Sunday (3/2)

Section 1

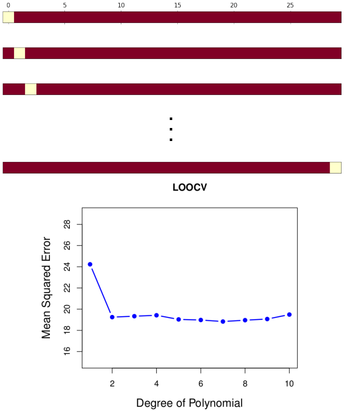
Last time

Approximations of Test Error

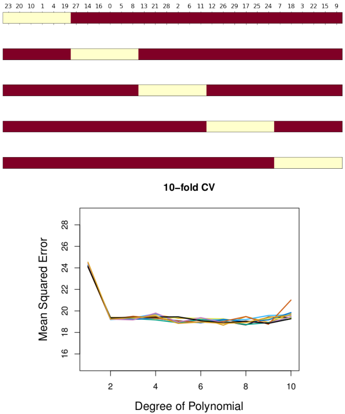
Validation Set



LOOCV

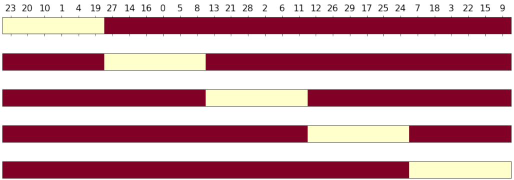


K-fold CV



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

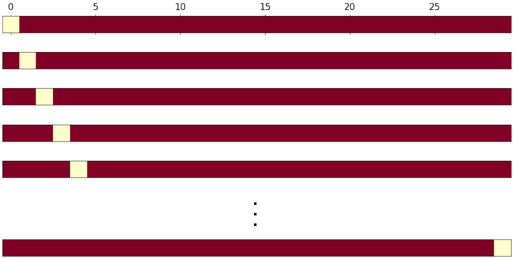
$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k}\mathrm{MSE}_i$$

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

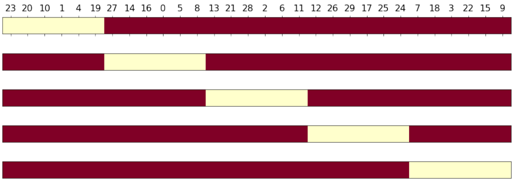


Return

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

Setup: k-fold

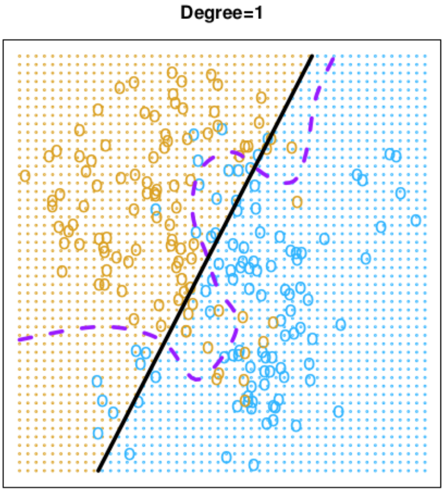
- Randomly split data into k-groups (folds)
- Approximately equal sized. For the sake of notation, say each set has ℓ points
- Remove ith fold U_i and reserve for testing.
- Train the model on remaining points
- Calculate $\mathrm{Err}_i = \frac{1}{\ell}\sum_{(x_j,y_j)\in U_i} \mathrm{I}(y_j \neq \hat{y}_j)$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



Return

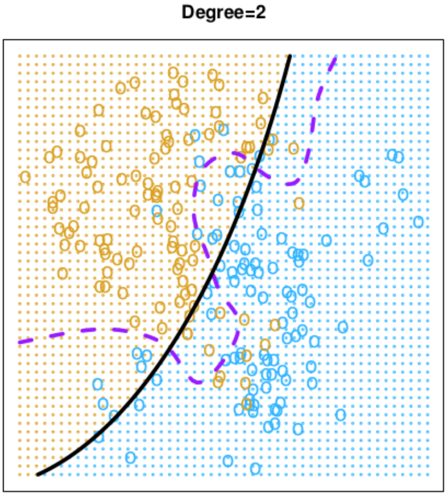
$$CV_{(k)} = \frac{1}{k}\sum_{i=1}^{k} \mathrm{Err}_i$$

Example on simulated data: Linear



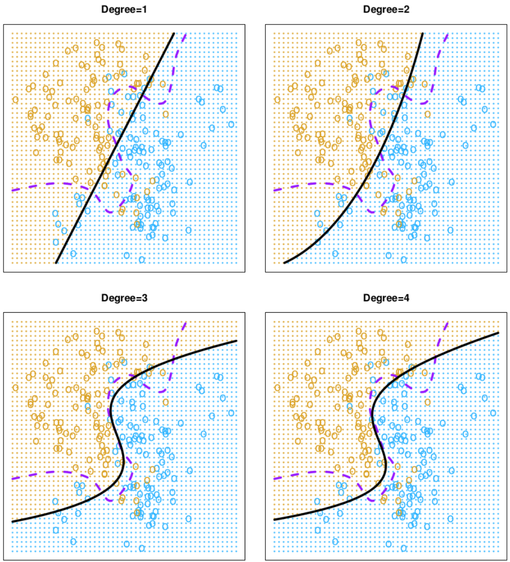
- Purple: Bayes decision boundary.
  - Error rate: 0.133
- Black: Logistic regression
  - $\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_2$
  - Error rate: 0.201

Example on simulated data: Quadratic logistic regression



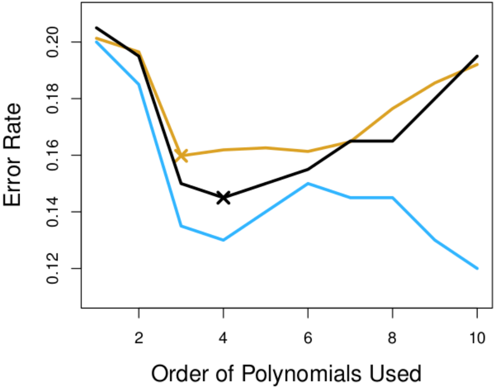
- Purple: Bayes decision boundary.
  - Error rate: 0.133
- Black: Logistic regression
  - $\log(p/(1-p)) = \beta_0 + \beta_1 X_1 + \beta_2 X_1^2 + \beta_3 X_2 + \beta_4 X_2^2$
  - Error rate: 0.197

Example on simulated data: all the polynomials!



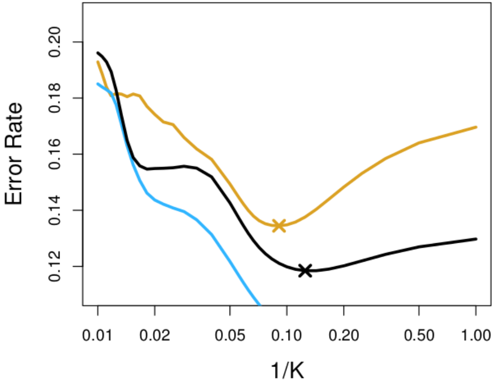
- Purple: Bayes decision boundary.
  - Error rate: 0.133
- Black: Logistic regression
  - Deg 1 Error rate: 0.201
  - Deg 2 Error rate: 0.197
  - Deg 3 Error rate: 0.160
  - Deg 4 Error rate: 0.162

Decide degree based on CV



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

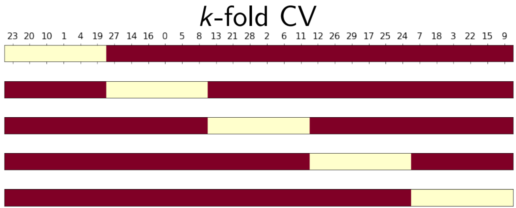
Similar game for KNN



- 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

	W	2/12	**Midterm #1**		
12	F	2/14	Leave one out CV	5.1.1, 5.1.2	
13	M	2/17	k-fold CV	5.1.3	
14	W	2/19	More k-fold CV	5.1.4-5	
15	F	2/21	k-fold CV for classification	5.1.5	
16	M	2/24	Subset selection	6.1	
17	W	2/26	Shrinkage: Ridge	6.2.1	
18	F	2/28	Shrinkage: Lasso	6.2.2	HW #4 Due Sun 3/2
	M	3/3	Spring Break		
	W	3/5	Spring Break		
	F	3/7	Spring Break		
19	M	3/10	PCA	6.3	
20	W	3/12	PCR	6.3	
	F	3/14	***Review***		HW #5 Due Sun 3/16
	M	3/17	**Midterm #2**		