Ch 6.1: Subset Selection

Lecture 16 - CMSE 381

Prof. Mengsen Zhang

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

Mon, Feb 24, 2025

Announcements

Last time

- *k*-fold CV for Classification

Covered in this lecture

- response to feedback (from Quiz 5)
- Subset selection
- Forward and Backward Selection

Announcements:

- HW #4 Due Sunday (3/2)

	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**		

Feedback collected in Quiz 5

- overall state of the class: ok to well
  - few find the class too slow (1 person wants more math!)
  - few find it too difficult
  - most people are in between, found ways to improve going forward
- common themes:
  - difficult to sustain attention throughout class & taking most relevant notes
  - want more sample problems
- what I can do:
  - I will go through jupyter solution after every coding segment.
  - I can pause more to ask more questions in class
  - I can write more keywords on slides slowly
- what you can do:
  - Read textbook chapter before class
  - Stay for entire class, take notes, ask questions
  - Respond to pre-review survey with your three burning questions (after spring break)

Section 1

Previously on linear regression ...

The problem of many features (p) relative to samples (n)

Up to now, we've focused on standard linear model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ and done least squares estimation.

Prediction accuracy

The problem of many features (p) relative to samples (n)

Up to now, we've focused on standard linear model: $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon$ and done least squares estimation.

Model Interpretability

Section 2

Best Subset Selection

Go through each combo of variables exhaustively (exhausting?)

All subsets of 4 variables ($2^4 = 16$)

- $\emptyset$

- X_1
- X_2
- X_3
- X_4

- $X_1 \; X_2$
- $X_1 \; X_3$
- $X_1 \; X_4$
- $X_2 \; X_3$
- $X_2 \; X_4$
- $X_3 \; X_4$

- $X_1 \; X_2 \; X_3$
- $X_1 \; X_2 \; X_4$
- $X_1 \; X_3 \; X_4$
- $X_2 \; X_3 \; X_4$

- $X_1 \; X_2 \; X_3 \; X_4$

One way of breaking this up

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.

2. For $k = 1, 2, \ldots p$:

   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.

   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Calculate by hand

We train a model using four variables, X_1, X_2, X_3, X_4. We're interested in getting a subset of the variables to use. The following table shows the mean squared error and the MSE value computed for the model learned using each possible subset of variables.

	Training MSE (x10^7)	k-fold CV Testing Error
Null model	8.76	10.08
X1	8.63	9.98
X2	7.42	8.01
X3	8.16	8.3
X4	8.33	9.06
X1,X2	4.33	7.47
X1,X3	5.82	5.22
X1,X4	3.17	4.23
X2,X3	4.07	3.78
X2,X4	3.31	4.01
X3,X4	3.06	4.16
X1,X2,X3	3.08	5.49
X1,X2,X4	3.55	4.02
X1,X3,X4	2.97	4.23
X2,X3,X4	2.98	3.17
X1,X2,X3,X4	2.16	4.39

1. What subset of variables is found for each of the sets $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ when using best subset selection?
2. What subset of variables is returned using best subset selection?

Extra work space if it helps

	Training MSE (x10^7)	k-fold CV Testing Error
Null model	8.76	10.08
X1	8.63	9.98
X2	7.42	8.01
X3	8.16	8.3
X4	8.33	9.06
X1,X2	4.33	7.47
X1,X3	5.82	5.22
X1,X4	3.17	4.23
X2,X3	4.07	3.78
X2,X4	3.31	4.01
X3,X4	3.06	4.16
X1,X2,X3	3.08	5.49
X1,X2,X4	3.55	4.02
X1,X3,X4	2.97	4.23
X2,X3,X4	2.98	3.17
X1,X2,X3,X4	2.16	4.39

- $\emptyset$

- X_1
- X_2
- X_3
- X_4

- X_1 X_2
- X_1 X_3
- X_1 X_4
- X_2 X_3
- X_2 X_4
- X_3 X_4

- X_1 X_2 X_3
- X_1 X_2 X_4
- X_1 X_3 X_4
- X_2 X_3 X_4

- X_1 X_2 X_3 X_4

Code to do this

Section 3

Forward Selection

What's the problem with best subset selection?

Forward Stepwise Selection

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.

2. For $k = 0, \ldots, p-1$:

   (a) Consider all $p-k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.

   (b) Choose the *best* among these $p-k$ models, and call it $\mathcal{M}_{k+1}$. Here *best* is defined as having smallest RSS or highest R^2.

3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

An example for Forward Stepwise Selection

- $\emptyset$

- X_1
- X_2
- X_3
- X_4

- X_1 X_2
- X_1 X_3
- X_1 X_4
- X_2 X_3
- X_2 X_4
- X_3 X_4

- X_1 X_2 X_3
- X_1 X_2 X_4
- X_1 X_3 X_4
- X_2 X_3 X_4

- X_1 X_2 X_3 X_4

Group work: by hand same example with forward example

We train a model using four variables, X_1, X_2, X_3, X_4. We're interested in getting a subset of the variables to use. The following table shows the mean squared error and the R^2 value computed for the model learned using each possible subset of variables.

	Training MSE (x10^7)	k-fold CV Testing Error
Null model	8.76	10.08
X1	8.63	9.98
X2	7.42	8.01
X3	8.16	8.3
X4	8.33	9.06
X1,X2	4.33	7.47
X1,X3	5.82	5.22
X1,X4	3.17	4.23
X2,X3	4.07	3.78
X2,X4	3.31	4.01
X3,X4	3.06	4.16
X1,X2,X3	3.08	5.49
X1,X2,X4	3.55	4.02
X1,X3,X4	2.97	4.23
X2,X3,X4	2.98	3.17
X1,X2,X3,X4	2.16	4.39

1. What subset of variables is found for each of the sets $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3, \mathcal{M}_4$ when using forward selection?
2. What subset of variables is returned using forward subset selection?

Extra work space if it helps

	Training MSE (x10^7)	k-fold CV Testing Error
Null model	8.76	10.08
X1	8.63	9.98
X2	7.42	8.01
X3	8.16	8.3
X4	8.33	9.06
X1,X2	4.33	7.47
X1,X3	5.82	5.22
X1,X4	3.17	4.23
X2,X3	4.07	3.78
X2,X4	3.31	4.01
X3,X4	3.06	4.16
X1,X2,X3	3.08	5.49
X1,X2,X4	3.55	4.02
X1,X3,X4	2.97	4.23
X2,X3,X4	2.98	3.17
X1,X2,X3,X4	2.16	4.39

- $\emptyset$

- X_1
- X_2
- X_3
- X_4

- X_1 X_2
- X_1 X_3
- X_1 X_4
- X_2 X_3
- X_2 X_4
- X_3 X_4

- X_1 X_2 X_3
- X_1 X_2 X_4
- X_1 X_3 X_4
- X_2 X_3 X_4

- X_1 X_2 X_3 X_4

Pros and Cons of Forward Stepwise

Pros:

Cons:

Section 4

Backward Selection

Backward stepwise selection

Algorithm 6.3 *Backward stepwise selection*

1. Let $\mathcal{M}_p$ denote the *full* model, which contains all p predictors.
2. For $k = p, p-1, \ldots, 1$:
   (a) Consider all k models that contain all but one of the predictors in $\mathcal{M}_k$, for a total of $k-1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}_{k-1}$. Here *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Pros and Cons of Backward Stepwise

Pros:

Cons:

TL;DR

Algorithm 6.1 *Best subset selection*

1. Let $\mathcal{M}_0$ denote the *null model*, which contains no predictors. This model simply predicts the sample mean for each observation.
2. For $k = 1, 2, \ldots p$:
   (a) Fit all $\binom{p}{k}$ models that contain exactly k predictors.
   (b) Pick the best among these $\binom{p}{k}$ models, and call it $\mathcal{M}_k$. Here *best* is defined as having the smallest RSS, or equivalently largest R^2.
3. Select a single best model from among $\mathcal{M}_0, \ldots, \mathcal{M}_p$ using cross-validated prediction error, C_p (AIC), BIC, or adjusted R^2.

- Modify step 2 with forward or backward selection
- Choose best model in step 3 using one of our adjusted training scores or CV

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**		