Ch 6.1: Subset Selection

Lecture 15 - CMSE 381

Prof. Elizabeth Munch

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

Fri, Oct 14, 2022

Announcements

Last time

- Boostrapping

Covered in this lecture

- Subset selection
- Forward and Backward Selection
- Adjusted training MSE scores: C_p, AIC, BIC, Adjusted R^2

Announcements:

- No jupyter notebook for this lecture
- HW #5 posted and due Monday

Section 1

Last time

Goals of fitting a given model

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 **Model Interpretability**

Goal of next chapter

Section 2

Best Subset Selection

Too many variables

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.

Group work: 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 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 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

Section 3

Forward Selection

What's the problem?

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.

An example for Backward 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 backward

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 Backward Stepwise

Pros:

Cons:

Section 5

Alternatives for Approximating Test Error

Remembering what we're doing

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.

Algorithm 6.2 *Forward stepwise selection*

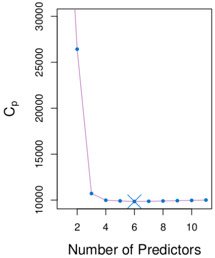
1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$:
   (a) Consider all $p - k$ models that augment the predictors in $\mathcal{M}_k$ with one additional predictor.
   (b) Choose the *best* among these $p - k$ models, and call it $\mathcal{M}_{k+1}$. 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.

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.

The C_p estimate

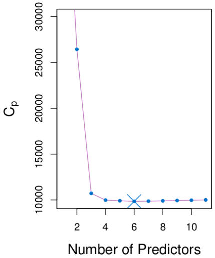
$$C_p = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2)$$



Example using `Credit`

The AIC criterion

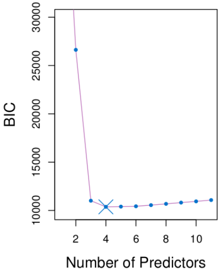
$$\mathrm{AIC} = \frac{1}{n}(\mathrm{RSS} + 2d\hat{\sigma}^2)$$



Example using `Credit`

The BIC

$$\mathrm{BIC} = \frac{1}{n}(\mathrm{RSS} + \log(n) d\hat{\sigma}^2)$$

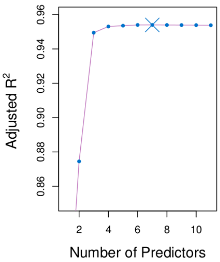


Example using `Credit`

Adjusted R^2

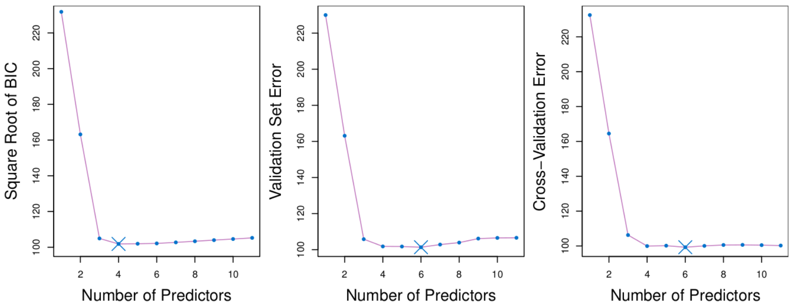
$$R^2 = 1 - \frac{\mathrm{RSS}}{\mathrm{TSS}}$$

$$\text{adjusted } R^2 = 1 - \frac{\mathrm{RSS}/(n-d-1)}{\mathrm{TSS}/(n-1)}$$



Comparisons

All this vs. Validation and Cross Validation



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

10	M	Oct 3	Leave one out CV	5.1.1, 5.1.2	
11	W	Oct 5	k-fold CV	5.1.3	
12	F	Oct 7	More k-fold CV,	5.1.4-5	
13	M	Oct 10	k-fold CV for classification	5.1.5	HW #4 Due
14	W	Oct 12	Resampling methods: Bootstrap	5.2	
15	F	Oct 14	Subset selection	6.1	
16	M	Oct 17	Shrinkage: Ridge	6.2.1	HW #5 Due
17	W	Oct 19	Shrinkage: Lasso	6.2.2	
18	F	Oct 21	[No class, Dr Munch out of town]		
	M	Oct 24	No class - Fall break		
19	W	Oct 26	Dimension Reduction	6.3	
20	F	Oct 28	More dimension reduction; High dimensions	6.4	HW #6 Due
	M	Oct 31	Review		
	W	Nov 2	***Midterm #2***		