Workspace 'Workspace' in 'lec14'
Page 1 (row 1, column 1)

MICHIGAN STATE UNIVERSITY

Module 6: Model Selection and Regularization

Lecture 14
Feb 20th, 2023
Ch 6.1- 6.2

SPARTANS WILL.

Feeling Sharing

MICHIGAN STATE UNIVERSITY

You can leave when and if you need to

Nothing on the syllabus is as important as **your well-beings**.

It is important to **ask for help** in this class and beyond

We should prioritize flexibility, grace, and care for each other

Recap

MICHIGAN STATE UNIVERSITY

Bootstrap vs CV

The goal: quantify the uncertainty associated with a given estimator (model) or statistical learning method.

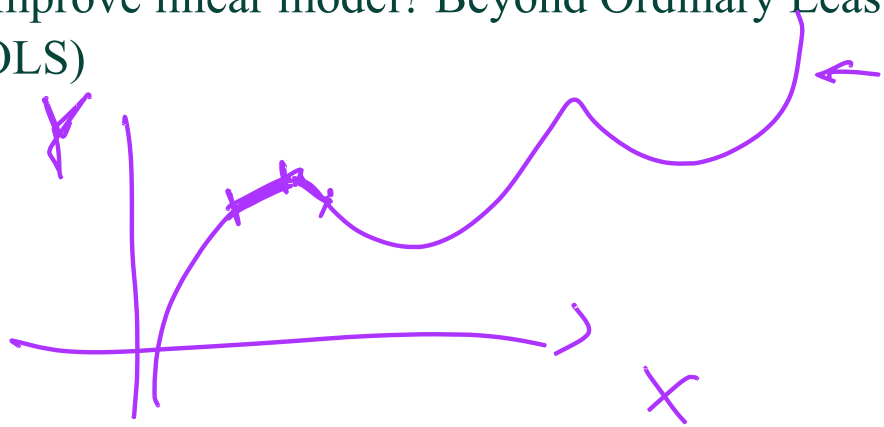
$$Y = \hat{\beta}_0 + \hat{\beta}_1 X + \epsilon$$

x is really associated with y?

Linear Models

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \epsilon.$$

- The linear model has advantages in terms of its interpretability and often shows good predictive performance.
- How to improve linear model? Beyond Ordinary Least Square(OLS)



locally
most time
it => linear.

Why Consider Alternatives to

- Prediction Accuracy: especially when $\boxed{p > n}$. Need to control variance
- Model Interpretability: Small number of features are easier to understand and design experiment.

$$Y = \begin{bmatrix} y_1 \\ \vdots \\ y_{2000} \end{bmatrix}$$

Three Classes of Methods

MICHIGAN STATE UNIVERSITY

- Subset Selection. We identify a subset of the p predictors then fit a model using least squares on the reduced set of variables.
- Shrinkage. We fit a model involving all p predictors, but the estimated coefficients are shrunken towards zero relative to the OLS estimates. This shrinkage (also known as regularization) has the effect of reducing variance and can also perform variable selection.
- Dimension Reduction.

Subset Selection

- Best subset selection
- Stepwise selection

Too many possible models

MICHIGAN STATE UNIVERSITY

M_0

- $\emptyset$

$Y = \beta_0 + \varepsilon$

M_1

- X_1
- X_2
- X_3
- X_4

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

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

M_2

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

M_3

- $X_1\ X_2\ X_3\ X_4$

φ

$2^4 + \cancel{4} + 1 = 2$

Best Subset Selection

MICHIGAN STATE UNIVERSITY

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.

$\frac{p(p-1)}{2}$

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

   p

   train err

   (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.

   $\mathcal{M}_1 \ldots$

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

testing error

$\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2, \ldots \mathcal{M}_p$

Group work:

MICHIGAN STATE UNIVERSITY

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?

Group work:

M_k

$2^p \cdot K$

	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

$\leftarrow M_0$

$\leftarrow M_1$

$\leftarrow M_2$ ✓

$\leftarrow M_3$

$\leftarrow M_4$

- $\emptyset$

- X_1
- X_2
- X_3
- X_4

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

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

- $X_1\ X_2\ X_3\ X_4$

Challenges for Best Subset selection

MICHIGAN STATE UNIVERSITY

2^{p}

$p = 40$

$2^{20,000}$

Forward Stepwise Selection

MICHIGAN STATE UNIVERSITY

Algorithm 6.2 *Forward stepwise selection*

1. Let $\mathcal{M}_0$ denote the *null* model, which contains no predictors.
2. For $k = 0, \ldots, p-1$: $k=2$ $\boxed{X_2}+$
   (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.

$\boxed{\mathcal{M}_0, \mathcal{M}_1 \ldots \mathcal{M}_p} \rightarrow$ CV

An Example for Forward Stepwise

MICHIGAN STATE UNIVERSITY

1

- $\emptyset$

M_0

4

- X_1
- X_2
- X_3
- X_4

M_1

3

- $X_1\ X_2$
- $\cancel{X_1\ X_3}$
- $\cancel{X_1\ X_4}$
- $X_2\ X_3$
- $X_2\ X_4$
- $\cancel{X_3\ X_4}$

M_2

2

M_3

- $X_1\ X_2\ X_3$
- $\cancel{X_1\ X_2\ X_4}$
- $\cancel{X_1\ X_3\ X_4}$
- $X_2\ X_3\ X_4$

1

M_4

- $X_1\ X_2\ X_3\ X_4$

$$11 \approx p^2 \ll 2^p$$

Group Work

MICHIGAN STATE UNIVERSITY

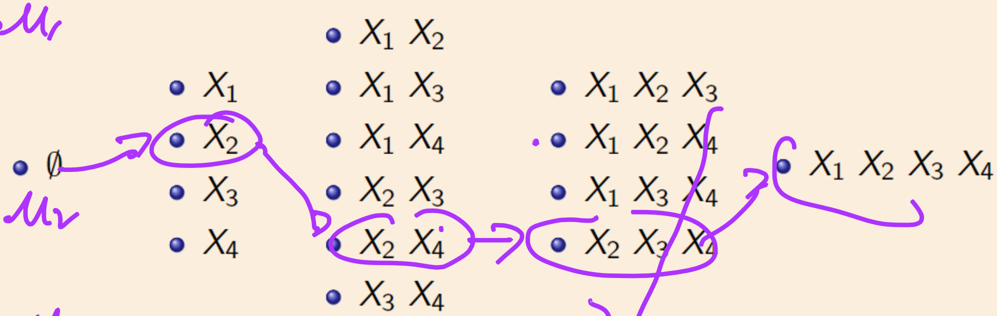
We train a model using four variables, X_1, X_2, X_3, X_4. We're interested in getting a subset of the variables to use. The following table shows the mean squared error and the R^2 value computed for the model learned using each possible subset of variables.

	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?

Group Work

	Training MSE (x10^7)	k-fold CV Testing Error	
Null model	8.76	10.08	M_0
X1	8.63	9.98	
X2	7.42	8.01	$\to M_1$
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	$\boxed{3.31}$	4.01	$\to M_2$
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	√ $\to M_3$
X1,X2,X3,X4	2.16	4.39	√ $\to M_4$



Pros and Cons of Forward Selection

MICHIGAN STATE UNIVERSITY

Pros

less models to be tested

Computationally feasible.

Cons

stuck on local solution

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

1	2	3	4	1
		$\cancel{X_1\ X_2}$		
	$\cancel{X_1}$	$X_1\ X_3$	$X_1\ X_2\ X_3$	
$\emptyset$	X_2	$X_1\ X_4$	$X_1\ X_2\ X_4$	$X_1\ X_2\ X_3\ X_4$
	$\cancel{X_3}$	$\cancel{X_2\ X_3}$	$X_1\ X_3\ X_4$	
	X_4	$\cancel{X_2\ X_4}$	$X_2\ X_3\ X_4$	
		$X_3\ X_4$		

11 models

Group work

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?

Group work

	Training MSE (x10^7)	k-fold CV Testing Error	
Null model	8.76	10.08	→ M_0
X1	8.63	9.98	
X2	7.42	8.01	
X3	8.16	8.3	→ M_1
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	→ M_2 ✓
X1,X2,X3	3.08	5.49	
X1,X2,X4	3.55	4.02	
X1,X3,X4	2.97	4.23	→ M_3
X2,X3,X4	2.98	3.17	
X1,X2,X3,X4	2.16	4.39	← M_4

- $\emptyset$

- X_1
- X_2
- X_3
- X_4

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

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

- $X_1\ X_2\ X_3\ X_4$

$$Y = \beta_0 + \beta_1 X_3 + \beta_2 X_4 + \varepsilon$$

Pros and Cons of Backward Selection

MICHIGAN STATE UNIVERSITY

pro

① computational feasible

Cons

② local solution

Forward vs Backwark

$p > n$

✓

X

p^2

Alternatives for Approximating Test Error

Estimating Test Error

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.

- CV is good but very time consuming.

- We can indirectly estimate test error by making an adjustment to the training error to account for the bias due overfitting

Estimating Test Error

MICHIGAN STATE UNIVERSITY

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 $k = 1, 2, \dots p$:
   (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 $\mathcal{M}_0, \dots, \mathcal{M}_p$ 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 $k = 0, \dots, p-1$:
   (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 $\mathcal{M}_0, \dots, \mathcal{M}_p$ 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 predict
2. For $k = p, p-1, \dots, 1$:
   (a) Consider all k models that contain all but one of the p in $\mathcal{M}_k$, for a total of $k - 1$ predictors.
   (b) Choose the *best* among these k models, and call it $\mathcal{M}$ *best* is defined as having smallest RSS or highest R^2.
3. Select a single best model from among $\mathcal{M}_0, \dots, \mathcal{M}_p$ usi validated prediction error, C_p (AIC), BIC, or adjusted R^2.

Cp = AIC

complexity ↑

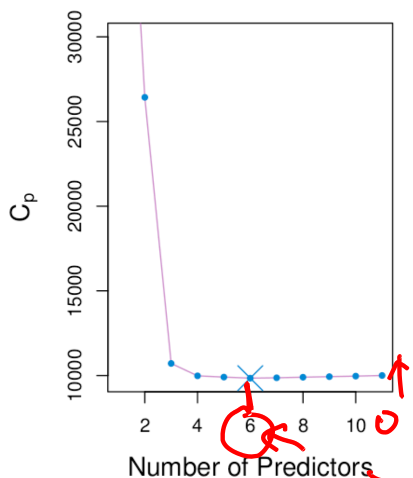
training ↓

- *Mallow's C_p*:

$$C_p = \frac{1}{n}\left(\text{RSS} + \boxed{2d\hat{\sigma}^2}\right),$$

→ penalized complexity

where d is the total # of parameters used and $\hat{\sigma}^2$ is an estimate of the variance of the error ϵ associated with each response measurement.



$2\hat{\sigma}^2 \cdot \boxed{d}$

$Y = f(x) + \epsilon$

$\epsilon \sim N(0, \sigma^2)$

AIC

MICHIGAN STATE UNIVERSITY

- The *AIC* criterion is defined for a large class of models fit by maximum likelihood:

$$\text{AIC} = \overbrace{\boxed{-2\log L}}^{\approx RSS} + \boxed{2 \cdot d}$$

where L is the maximized value of the likelihood function for the estimated model.

BIC

MICHIGAN STATE UNIVERSITY

$M_1 \to$ AIC

$M_2 \to$ BIC

$d \to$ model complexity

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

$n = 100$

$\log(n) \hat{\sigma}^2 \cdot d$

$2 \cdot \hat{\sigma}^2 \cdot d$

- Like C_p, the BIC will tend to take on a small value for a model with a low test error, and so generally we select the model that has the lowest BIC value.
- Notice that BIC replaces the $2d\hat{\sigma}^2$ used by C_p with a $\log(n)d\hat{\sigma}^2$ term, where n is the number of observations.
- Since $\log n > 2$ for any $n > 7$, the BIC statistic generally places a heavier penalty on models with many variables, and hence results in the selection of smaller models than

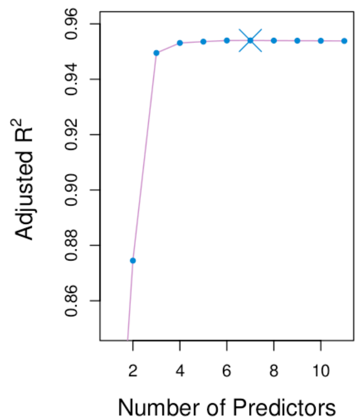
Adjusted $R2$

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

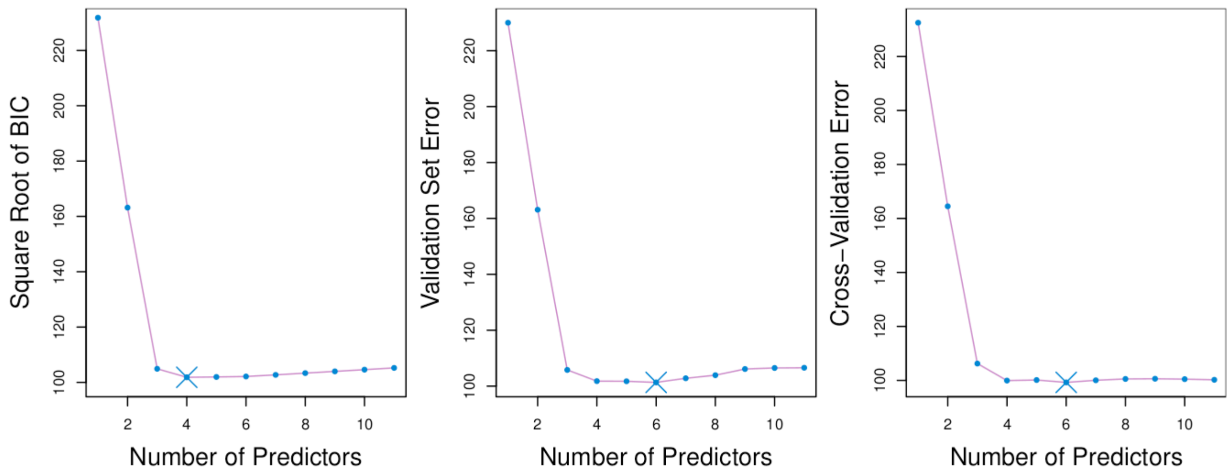
$$\text{Adjusted } R^2 = 1 - \frac{\text{RSS}/(n-d-1)}{\text{TSS}/(n-1)}.$$

where TSS is the total sum of squares.

- Unlike C_p, AIC, and BIC, for which a *small* value indicates a model with a low test error, a *large* value of adjusted R^2 indicates a model with a small test error.



Comparisons



Bonus Quiz 11

MICHIGAN STATE UNIVERSITY

We have two training points (x1 = 1, x2 = 1, y = 1) and (x1 = 2, x2 = 3, y = 2). We want to fit a linear model to minimize the MSE

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

What is the minimun MSE? What is the corresponding model?

$$\begin{cases} 1 = \beta_0 + \beta_1 \cdot 1 + \beta_2 \cdot 1 \\ 2 = \beta_0 + 2 \cdot \beta_1 + 3 \cdot \beta_2 \end{cases}$$