Linear-Model Regularization

CS 530
Chapman

Spring 2021

1

Take 🏠 Message for rest of course: Linear Model Regularization

- ➢ What is regularization & why is it needed?
- ➢ How do we regularize linear models?

2

Table of Contents

3

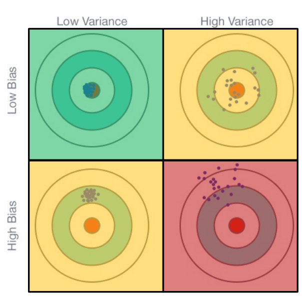
Regularization Methods

Subset selection can improve prediction accuracy when only a few of the $\{x_j\}$ variables have a strong relation with the dependent y variable. However, in other cases, it can overfit, making the test error rise.

As an alternative to subset selection, we can also try to shrink or *regularize* the coefficient estimates to reduce overfitting. We can do this by utilizing the bias-variance tradeoff to create a set of regularization methods.

4

Bias, Variance, and Regularization

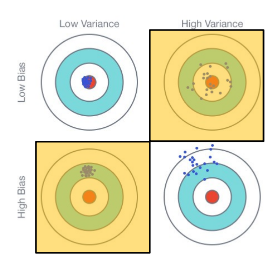


5

Bias, Variance, and Regularization

- Which situation is worse
  - – High bias, low variance,

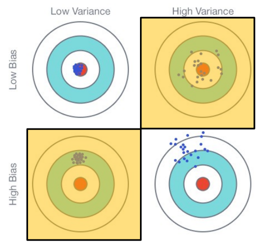
    or
  - – Low bias, high variance?



6

Bias, Variance, and Regularization

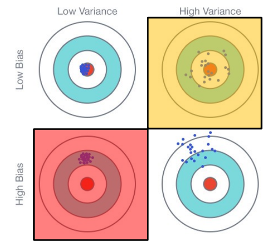
- Which situation worse
  ↑bias-↓var / ↓bias-↑var?



7

Bias, Variance, and Regularization

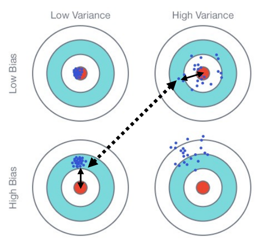
- Which situation worse
  ↑bias-↓var / ↓bias-↑var?
- Many think ↑bias-↓var worse: always off target



8

Bias, Variance, and Regularization

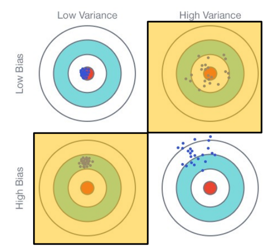
- Which situation worse
  ↑bias-↓var / ↓bias-↑var?
- Many think ↑bias-↓var worse: always off target
- But, in real life, we only get one dataset. So, ↑var could be as bad or worse than ↑bias



9

Bias, Variance, and Regularization

- Which situation worse
  ↑bias-↓var / ↓bias-↑var?
- Many think ↑bias-↓var worse: always off target
- But, in real life, we only get one dataset. So, ↑var could be as bad or worse than ↑bias
- So, <u>slightly ↑bias worth much ↓var</u>



10

Regularization Methods

How can we take advantage of the bias-variance tradeoff?

The regularization methods
- ➢ Ridge regression
- ➢ LASSO method
- ➢ Elastic Net regression

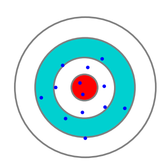
often reduce much variance at the cost of a little bias. So, overall, they reduce the MSE (on the test set):

$$\text{MSE} = \left[\text{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \text{Var}\left(\hat{f}(x_0)\right) + \text{Var}(\varepsilon)$$

11

Biased/Unbiased Estimation

- Ordinary least squares (OLS) estimation guarantees unbiased estimators $\{\beta_j\}$
- But there are no guarantees about the variance being small
- Sometimes a little bias decreases the variance by a lot
- So, biased estimators can result in overall smaller error on the test set



12

Biased estimator:

RIDGE REGRESSION

13

Ridge Regression

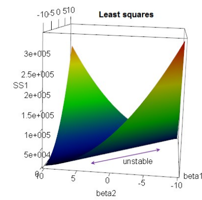


Image Source

Ridge regression gets its name from the above scenario. Multicollinearity in an OLS model can lead to a ridge/depression in the likelihood function, leading to a valley in the RSS (on the left).

14

Ridge Regression

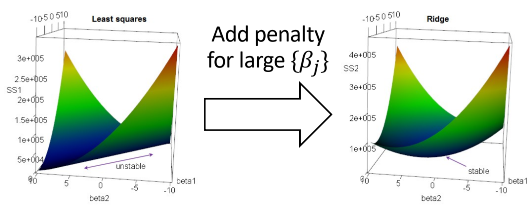


Image Source

On the right, ridge regression elminates the ridge by adding a penalty term, which turns the ridge/valley into a peak/depression.

15

Ridge Regression

Ridge Regression is similar to OLS regression, but with a modified cost function. OLS regression finds coefficient estimates $\hat{\beta}_j$ minimizing

$$\text{RSS} = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2$$

Ridge regression minimizes the RSS with an added penalty term

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\hat{\beta}_j^{\,2} = \text{RSS} + \lambda\sum_{j=1}^{p}\hat{\beta}_j^{\,2}$$

where $\lambda \geq 0$ is the tuning parameter.

16

Ridge Regression

Ridge Regression: $\underbrace{\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2}_{\text{RSS}} + \underbrace{\lambda\sum_{j=1}^{p}\hat{\beta}_j^{\,2}}_{\text{Penalty term}}$

Ridge regression differs from OLS regression by minimizing not just the RSS,

but also $\left(\lambda\sum_{j=1}^{p}\hat{\beta}_j^{\,2}\right)$, a term called the 'shrinkage penalty'.

Minimizing both the RSS and shrinkage penalty results in smaller $\{|\beta_j|\}$.

17

Ridge Regression

Ridge Regression: $\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\hat{\beta}_j^{\,2}$

The tuning parameter λ controls how much the coefficient estimates will be shrunk.

For $\lambda = 0$, Ridge Regression collapses back to OLS regression. The shrinkage term then has no effect.

As $\lambda \to \infty$, the coefficient estimates (except for $\hat{\beta}_0$) will all approach 0.

18

Biased estimator:

LASSO REGRESSION

19

The LASSO Method

The LASSO method (Least Absolute Shrinkage and Selection Operator), like ridge regression, attempts to shrink the model coefficients. LASSO often results in *feature selection* (variable elimination) and not just lower magnitude. This results in a potentially more interpretable model than ridge regression.

20

The LASSO Method

LASSO regularization is very similar to ridge regression, with a subtle—but important—difference:

Ridge Regression: (L_2-Norm)

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\boxed{\beta_j{}^2}$$

LASSO Method: (L_1-Norm)

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}\boxed{|\beta_j|}$$

- Ridge Regression penalizes based on the sum of the squares (L_2-norm) of the coefficient estimates.
- LASSO penalizes based on the absolute values (L_1-norm) of the coefficient estimates.

21

The LASSO Method

LASSO Method:

$$\sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

The key advantage of the LASSO method is that, on top of shrinking coefficient estimates, it can make some $\hat{\beta}_j$ terms disappear, resulting in variable selection.

22

LASSO vs. Ridge Methods

It can be shown that the Ridge formulation is equivalent to:

$$\underset{\vec{\beta}}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}\beta_j^2 \le s$$

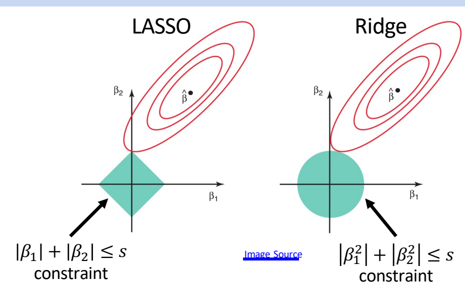
While the LASSO formulation is equivalent to:

$$\underset{\vec{\beta}}{\text{minimize}}\left\{\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2\right\} \text{ subject to } \sum_{j=1}^{p}|\beta_j| \le s$$

For every λ there exists an s for which the above holds.

23

LASSO vs. Ridge Methods



Meeting point of red ellipses (constant RSSs contours) & green regions is minimization under constraints. Ellipse is much more likely to hit constraint on axis ($\beta_1 = 0$ above) for LASSO than for Ridge. So LASSO much more likely to eliminate variables.

24

Biased estimator:

ELASTIC-NET REGRESSION

25

When not to use LASSO?

- LASSO tends to zero out variables as λ grows. This is often good. But may not be desired sometimes.
- Also, LASSO always arbitrarily chooses one among a group of correlated variables
  - What if we know that these are not as heavily correlated on the test set and want to keep them all?
  - What is we want to choose the variable to be dropped?
  - What if we must not drop any variables?
- For large number of variables (p) & few data points (n), LASSO will never pick more variables than data points

We sometimes should not use LASSO

27

Elastic Net: A Mix of Ridge and LASSO

Elastic-net regression overcomes the limitations of the LASSO method by imposing a combination of Ridge and LASSO penalties (L_1 and L_2 norm penalties) on the cost function. This also makes the RSS function convex with a unique minimum (like in the diagram for ridge regression).

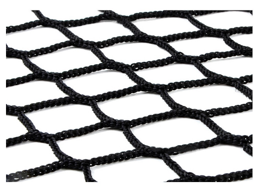


Image Source

28

Elastic Net: A Mix of Ridge and LASSO

$$RSS = \sum_{i=1}^{n}\left(y_i - \hat{\beta}_0 - \sum_{j=1}^{p}\hat{\beta}_j x_{ij}\right)^2$$

Ridge Regression: Minimize $RSS + \lambda\sum_{j=1}^{p}\hat{\beta}_j^{\,2}$

LASSO Method: Minimize $RSS + \lambda\sum_{j=1}^{p}|\hat{\beta}_j|$

Elastic Net: Minimize $RSS + \lambda\sum_{j=1}^{p}(\alpha\beta_j^{\,2} + (1-\alpha)\,|\beta_j|)$

Elastic net regression gives us a combination of ridge and LASSO by penalizing on both the L_1 and L_2 norms. The α term determines the balance between ridge and LASSO ($0 \le \alpha \le 1$; $\alpha = 1$: Ridge, $\alpha = 0$: LASSO).

29

Biased estimator:

TUNING RIDGE, LASSO, & ELASTIC NET

34

The Tuning Parameter λ

How to determine the value of λ (termed α in Python) to use for these regularization methods?

We can try a range of different λ values, build models based on them, and use cross-validation and measures of model fit to choose the λ value that will lead to a good model.

35

Tuning Parameter λ for Ridge Regression

This plot comes from applying ridge regression with several values of λ to a data set. Notice how the coefficient values for each variable gradually shrink and approach 0.

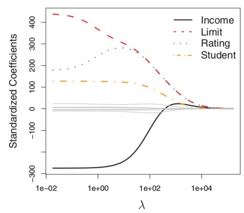


Image Source

36

Tuning Parameter λ for LASSO

Here, we have the same data set, but instead of ridge regression, we apply the LASSO method with different values of λ. Notice how the variables more sharply and abruptly go to 0 as λ increases.

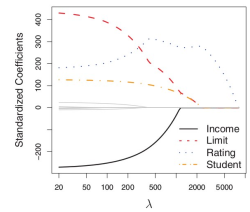


Image Source

37

The Tuning Parameter λ—Ridge

For Ridge Regression:

- Black line: squared bias
- Green line: variance
- Purple line: test MSE

The 'X' marks where the minimum test MSE is located. You can see the bias-variance tradeoff at work as we adjust λ.

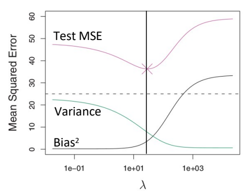


Image Source

$$\text{MSE} = \left[\text{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \text{Var}\left(\hat{f}(x_0)\right) + \text{Var}(\varepsilon)$$

38

The Tuning Parameter λ—LASSO

For the LASSO method:

- Black line: squared bias
- Green line: variance
- Purple line: test MSE

The 'X' marks where the minimum test MSE is located.

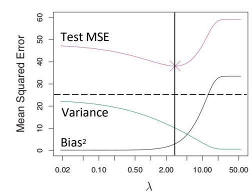


Image Source

$$\text{MSE} = \left[\text{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \text{Var}\left(\hat{f}(x_0)\right) + \text{Var}(\varepsilon)$$

39

The Tuning Parameter λ

- Black: squared bias
- Green: variance
- Purple: test MSE
    - Solid line: LASSO
    - Dashed line: Ridge

The 'X' indicates the LASSO method with the smallest test MSE. This graph shows the value of cross-validation. A model that fits the training set perfectly typically overfits the test set.

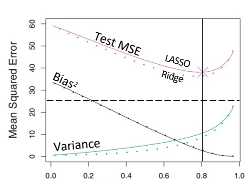


Image Source

$$\text{MSE} = \left[\text{Bias}\left(\hat{f}(x_0)\right)\right]^2 + \text{Var}\left(\hat{f}(x_0)\right) + \text{Var}(\varepsilon)$$

40

The Tuning Parameter λ

To avoid overfitting, we can look at a range of values of λ, create a model with that λ and the training set, and use that model on the test set to calculate cross-validation error. The λ value with the smallest cross-validation error is indicated with the vertical dashed line.

This is the λ value we should use for a predictive model. Similarly, cross-validation can help us determine which regularization method (ridge-LASSO mixing parameter, α, in elastic net) should be used.

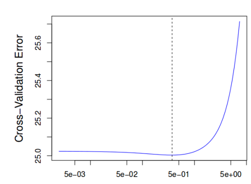


Image Source

41

Conclusion: Linear-Model Regularization

- Regularization: Methods for decreasing variance by decreasing coefficient estimate magnitudes (Ridge & LASSO) or eliminating coefficients (a.k.a. variable selection; LASSO only)
  - This increases bias, though often less than the decrease in variance, so overall MSE decreases
- Choosing the right tuning parameter for regularization is difficult and typically carried out using cross validation

42

Take 🏠 message for rest of course: Regularization

- Regularization ("shrinkage") of models to minimize variance introduced
- Cross-validation for tuning parameters mentioned

43