Regularization

Ryan Miller

Outline

1. Regularization and the bias-variance trade-off
2. Ridge regression
3. Lasso regression

Review

Consider the basic linear regression model:

$$Y = w_o + w_1 X_1 + w_2 X_2 + \ldots + w_p X_p + \epsilon$$

We've previously estimated **w**, the vector of weights, by optimizing the following cost function:

$$\text{Cost} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_i)^2 = \frac{1}{n} (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^T (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})$$

Regularized Regression

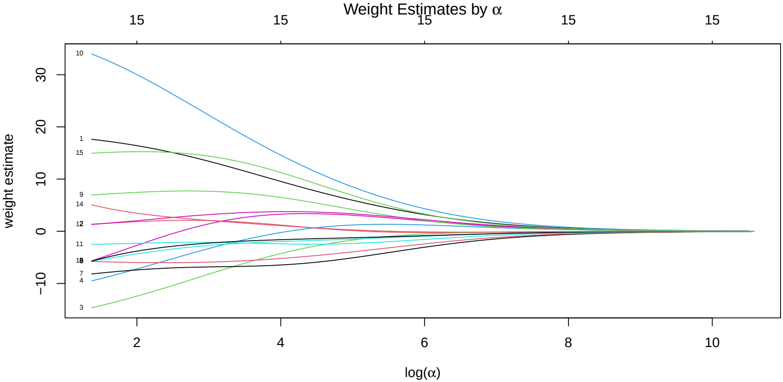
Regularized regression adds a penalty term to the cost function that shrinks weight estimates towards zero:

$$\text{Cost} = \frac{1}{n}(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) + P_\alpha(\hat{\mathbf{w}})$$

- $P()$ is a *penalty function* involving α, a **regularization parameter** that controls the trade-off between each term in the cost function

Example

When the regularization parameter, α, is large, the penalty term dominates the cost function and weights are estimated to be zero. When *alpha* is zero, cost function reduces to squared error loss.



Benefits of Regularization

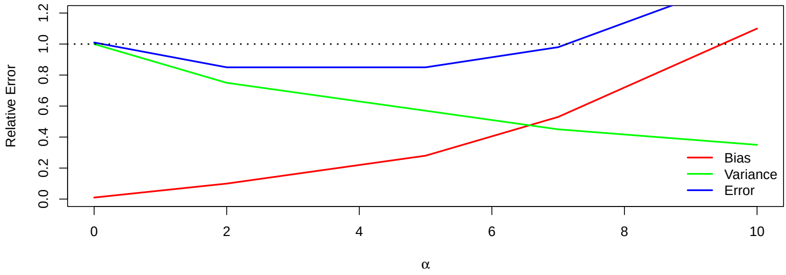
- Intuitively, regularization operates under the belief that across many predictors small weights should be more likely than large weights
  - Thus, overfitting can be prevented by using penalization to discourage larger weight estimates

Benefits of Regularization

- Intuitively, regularization operates under the belief that across many predictors small weights should be more likely than large weights
  - Thus, overfitting can be prevented by using penalization to discourage larger weight estimates
- In 1970, it was shown by Hoerl and Kennard that *ridge regression* (a type of regularized regression) can *always* produce a lower *RMSE* than ordinary (unpenalized) regression

Benefits of Regularization

Mathematically, it's possible to decompose mean-squared error (MSE) into bias and variance terms. Here's a heuristic look at how these components might look as α is varied:



Ridge Regression

Ridge regression uses the penalty function:

$$P_\alpha(\mathbf{w}) = \alpha \sum_{j=1}^{p} w_i^2$$

This makes the ridge regression cost function:

$$\text{Cost} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \mathbf{x}_i^T \mathbf{w})^2 + \alpha \sum_{j=1}^{p} w_i^2$$

In matrix form, this looks like:

$$\text{Cost} = \frac{1}{n} (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}})^T (\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) + \alpha \hat{\mathbf{w}}^T \hat{\mathbf{w}}$$

Note that $\hat{\mathbf{w}}^T \hat{\mathbf{w}}$ is the squared *L2 Norm* of the weight vector (or $||\hat{\mathbf{w}}||_2^2$), so the ridge penalty is often called *L2 regularization*

Ridge Regression

Similar to ordinary linear regression, minimizing the ridge regression cost function has a closed-form solution:

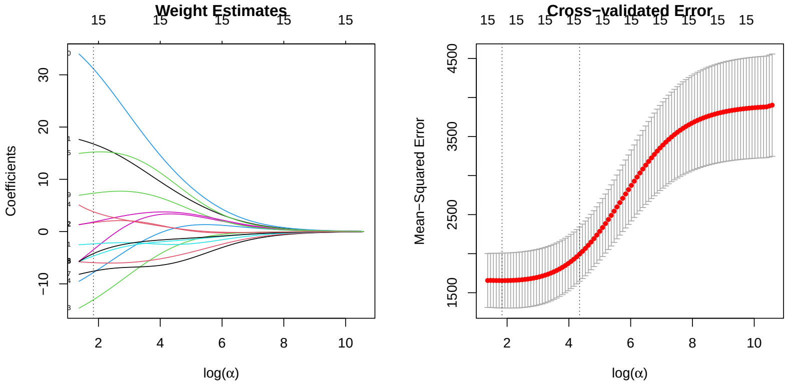
$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \alpha\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$

The method gets its name from the “ridge” added to the diagonal of $\mathbf{X}^T\mathbf{X}$ prior to inversion

Choosing α

- In penalized regression, α is a tuning parameter, with different values leading to different weight estimates
  - Larger values of α shrink the weights closer to zero (introducing more bias while reducing variance)
  - When $\alpha = 0$, the ridge regression estimates are the same those of ordinary linear regression
- Because penalization is proportional to the magnitude of w_j, it is important to *standardize* each variable as a pre-processing step when using regularization

Below are results for data that uses pollution and demographic variables of 60 US metro areas to their predict age-adjusted mortality:



Lasso

- The ridge penalty provides *stability* (ie: reduces variance) at the expense of adding *bias*
  - However, it doesn't truly reduce the complexity of the model (the number of non-zero weights is the same, regardless of the amount of penalization)

Lasso

- The ridge penalty provides *stability* (ie: reduces variance) at the expense of adding *bias*
  - However, it doesn't truly reduce the complexity of the model (the number of non-zero weights is the same, regardless of the amount of penalization)
- The lasso (least absolute shrinkage and selection operator) addresses this shortcoming by promoting *sparsity* in the estimated weight vector
  - The lasso cost function is shown below:

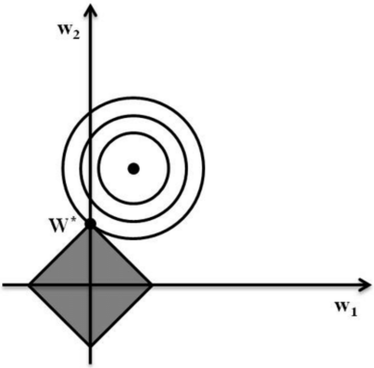
$$\frac{1}{n}\text{Cost} = \sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\mathbf{w})^2 + \alpha\sum_{j=1}^{p}|w_i|$$

- The lasso penalty involves the absolute value function, which is not strictly differentiable at its minimum
  - This leads to weight estimates of exactly zero being optimal in less important dimensions

Lasso

- To better understand why the lasso penalty promotes sparse weight estimates, we can view minimizing the lasso cost function as a constrained optimization problem
  - That is, the lasso's estimate of $\mathbf{w}$ minimizes $\frac{1}{n}\sum_{i=1}^{n}(y_i - \mathbf{x}_i^T\mathbf{w})^2$ subject to the constraint $\sum_{j=1}^{p}|w_j| < c$ where c describes a fixed amount of penalization (a function of α)
  - For comparison, the ridge estimate is similar but with the constraint $\sum_{j=1}^{p} w_j^2 < c$
- The next slide provides a geometric illustration of why the lasso constraint promotes sparsity, but the ridge constraint does not

Lasso vs. Ridge

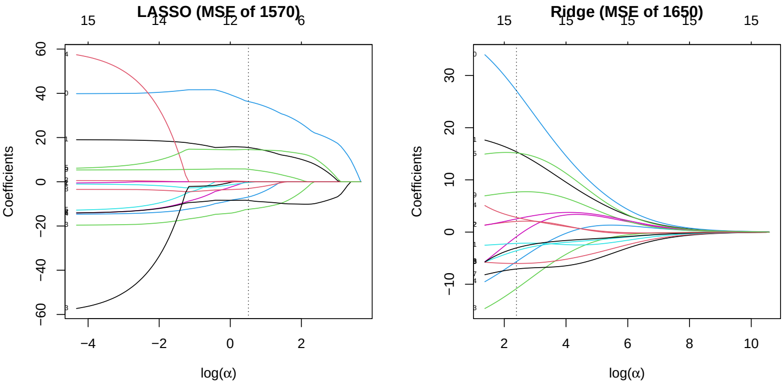


In two dimensions, weight estimates satisfying $\sum_{j=1}^{p} |w_j| < c$ exist within a diamond, while those satisfying $\sum_{j=1}^{p} w_j^2 < c$ exist within an ellipse. The former is likely to intersect contours of the squared error cost function at a corner (a weight estimate of exactly zero).

image credit: https://www.researchgate.net/figure/Plot-demonstrating-the-Sparsity-caused-by-the-LASSO-Penalty-The-plot-shows-the_fig1_317357840

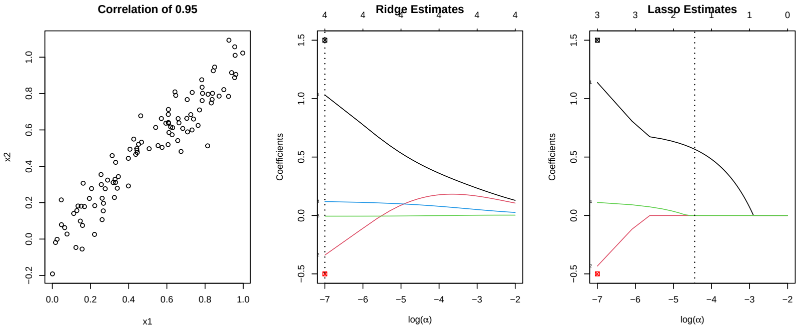
Lasso

For pollution example, lasso achieves a minimum cross-validated mean-squared error of around 1570, while ridge regression's minimum error (shown in an earlier slide) is around 1650 for these data.



Ridge Regression and Multicollinearity

In the presence of *multicollinearity*, lasso favors a single representative, while ridge will split the weight estimates in a more balanced manner:



Estimates found using ridge regularization can be more generalizable to new data for this reason.

Final Remarks on Regularization

- The lasso and ridge penalties can be used for the regularization of nearly any estimator
  - In general, regularization is an effect means of calibrating highly flexible/complex models so that they do not overfit the training data
  - Most advanced machine learning models involve some form of regularization