DS-UA 112
Introduction to Data Science

Week 13: Lecture 1

Regularization – Ridge and Lasso

How does shrinking parameters
help us to determine models?

DS-UA 112
Introduction to Data Science

Week 13: Lecture 1
Regularization – Ridge and Lasso

Adapted from Nolan, Speed, Gonzalez, Lau

Announcements

▶ Please check Week 12 agenda on NYU Classes

  ▶ Lab 12

    ▶ Due on Friday April 24 at 11:59PM EST

  ▶ Homework 5

    ▶ Due on Friday May 1 at 11:59PM EST

Review

▶ We can use backward feature selection to eliminate features. However we need to fit many different models to determine the irrelevant features

▶ Instead we can use regularization. We replace the average loss for linear regression

$$\frac{1}{n} \sum_{i=1}^{n} (a + bx_i - y_i)^2$$

with regularized average loss for lasso regression

$$\lambda (|a| + |b|) + \frac{1}{n} \sum_{i=1}^{n} (a + bx_i - y_i)^2$$

Lasso Regression Weights

4
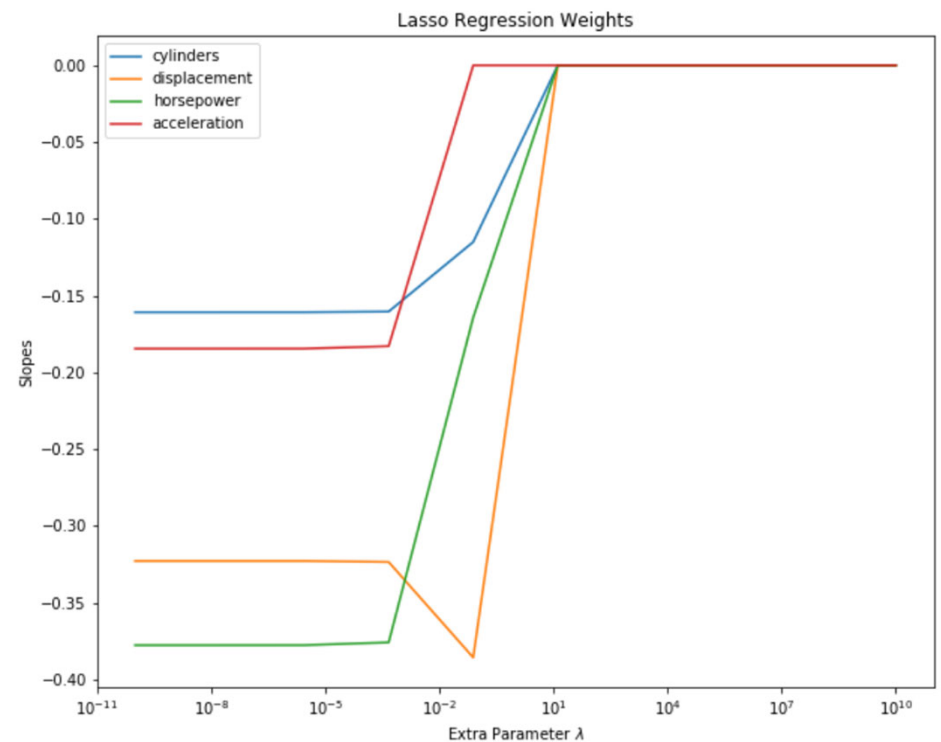
Agenda

▶ Lasso Regression

▶ Ridge Regression

Extra Parameter

▶ We need to add an extra parameter to a model for regularization.

▶ Unlike the other parameters, we cannot use a method like gradient descent to determine the value of the extra parameter.

▶ Instead we can use cross validation. We select the extra parameter that gives good prediction across validation sets.

Sample

Training Set → Training Error

Validation Set → Validation Error

Test Set → Test Error
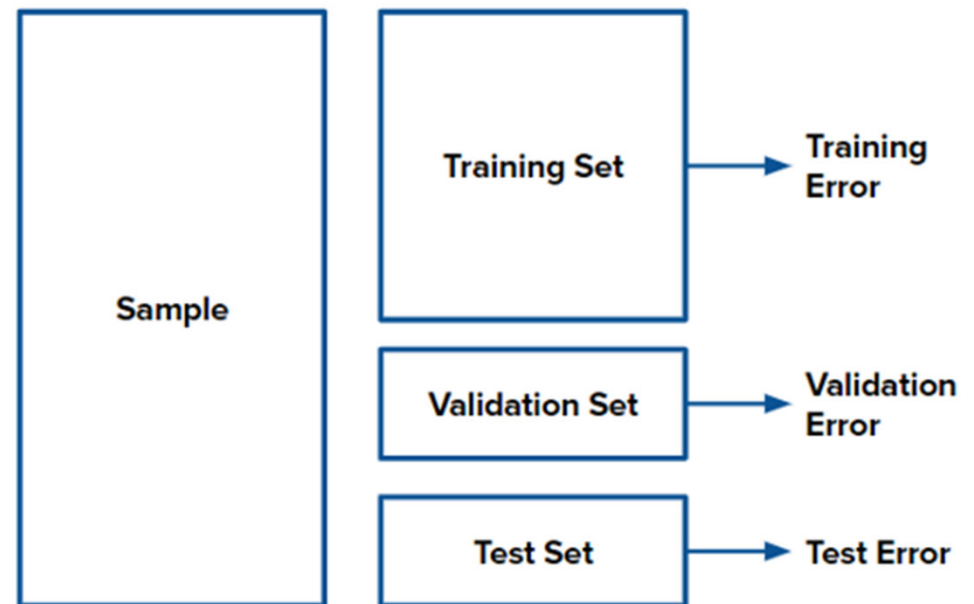
Error Bars

▶ We want to choose models that are both accurate and consistent.

▶ With **cross validation** we measure the difference between predictions and observations on many datasets.

  ▶ Small errors give us accuracy

  ▶ Similarity between errors give us consistency

▶ We can visualize both the accuracy and consistency through a line chart with error bars
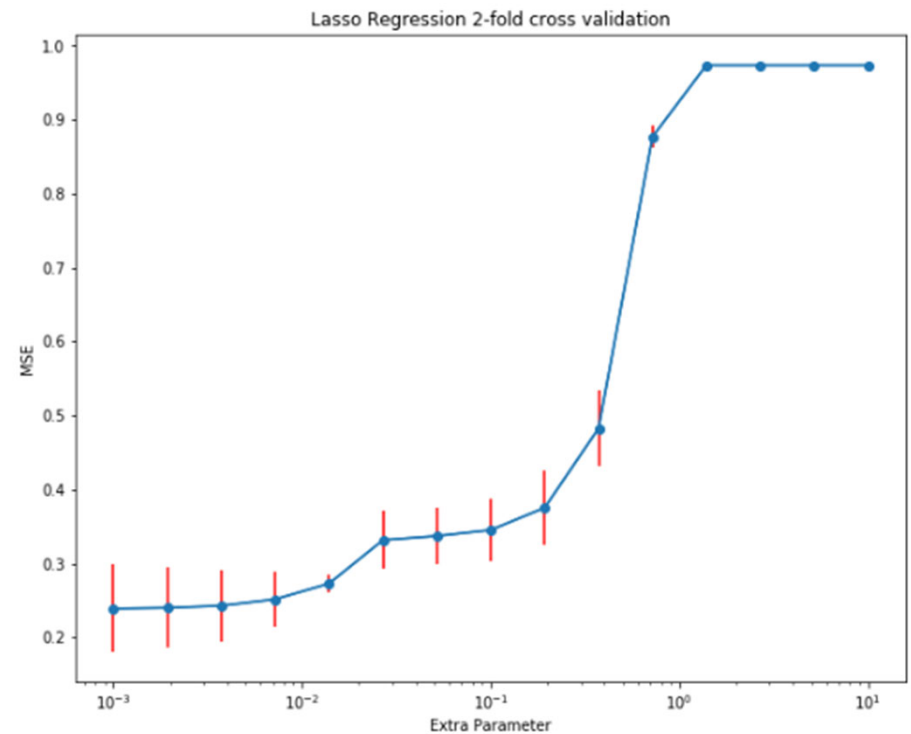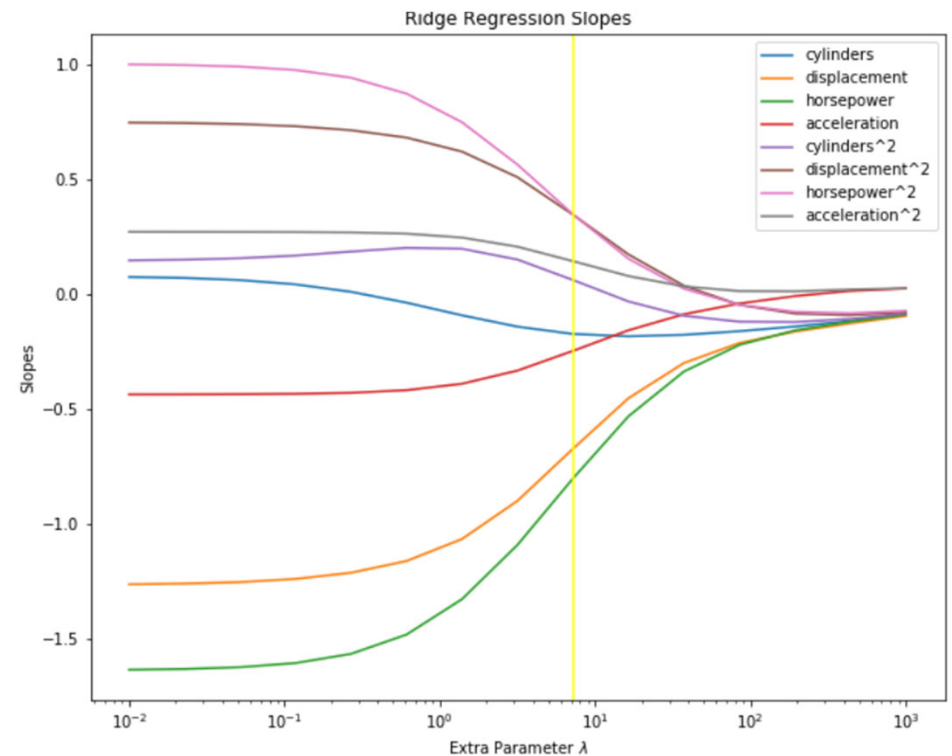
Ridge Regression

▶ Suppose we have associated features. If we want to shrink their parameters, then we should shrink the same amount.

▶ We can replace the average loss from linear regression

$$\frac{1}{n} \sum_{i=1}^{n} (a + bx_i - y_i)^2$$

with regularized average loss for ridge regression

$$\lambda \left(a^2 + b^2 \right) + \frac{1}{n} \sum_{i=1}^{n} (a + bx_i - y_i)^2$$
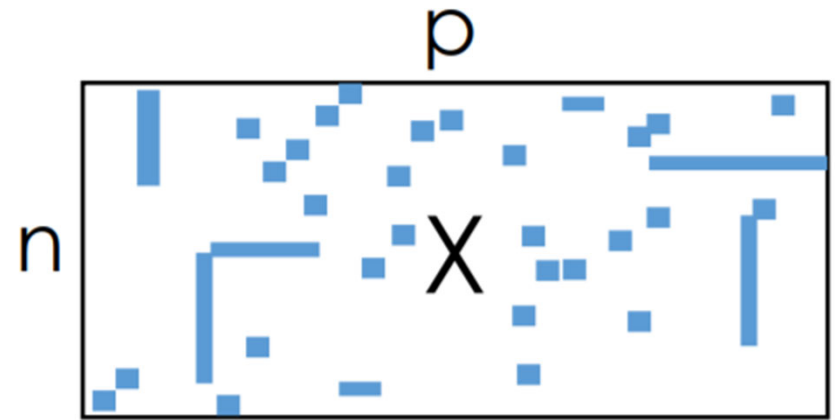

Ridge Regression Slopes

Shrinking Weights

▶ Lasso Regression tends to shrink parameters down to zero.

  ▶ Helpful to eliminate features from the model

  ▶ Erratically chooses between associated features

▶ Ridge Regression tends to shrink parameters close to zero.

  ▶ Helpful to average out the values of the parameters among associated features

  ▶ Cannot eliminate features from the model

9
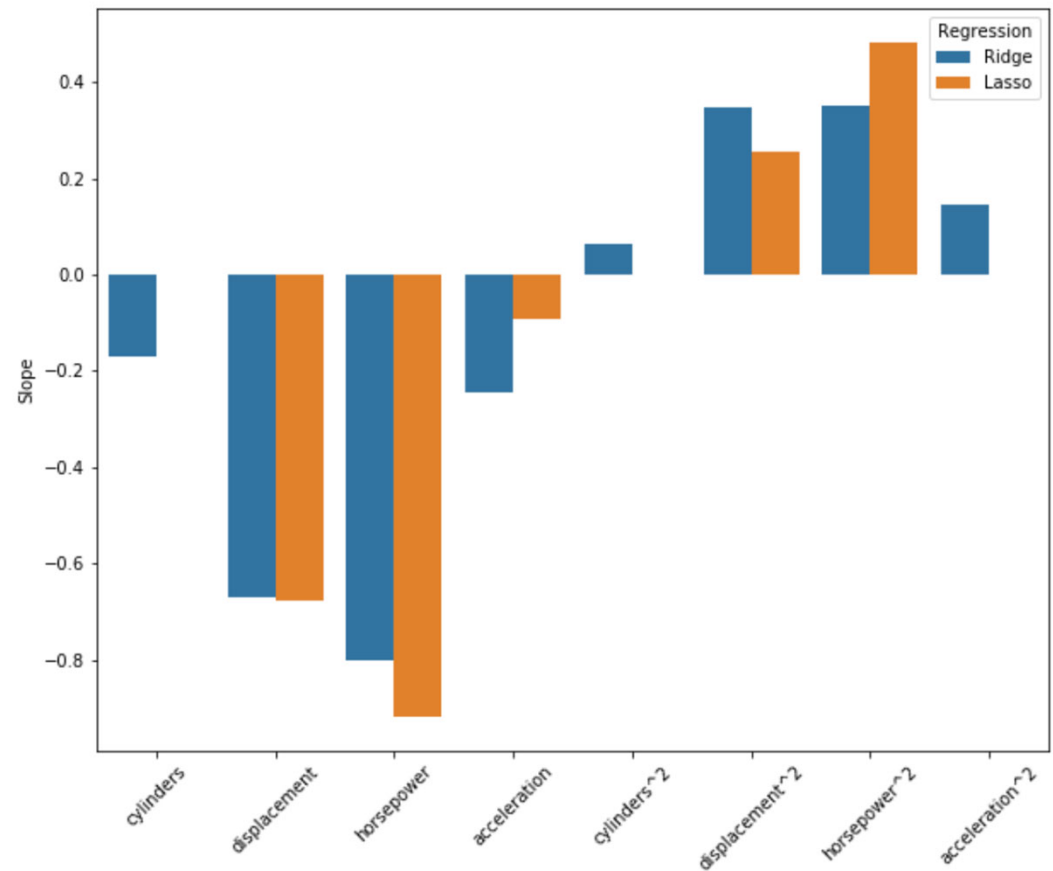
Shrinking Weights

Try Ridge Regression on dense tables with few columns

p

n X

p

n X

Try Lasso Regression on sparse tables with many columns

Summary

▶ Lasso Regression

▶ Ridge Regression

Goals

▶ Transform data to standard units for regularization

▶ Use cross validation to determine extra parameter

▶ Understand similarities and differences between Lasso Regression and Ridge Regression

Questions

- ▶ Questions on Piazza?
  - ▶ Please provide your feedback along with questions
- ▶ Question for You!
  - ▶ Should we use knowledge about a population to determine models?

NEW SCIENTIST

Specialist Knowledge Is Useless and Unhelpful

When data prediction is a game, the experts lose out.

By PETER ALDHOUS   DEC 08, 2012

PA: That sounds very different from the traditional approach to building predictive models. How have experts reacted?

JH: The messages are uncomfortable for a lot of people. It's controversial because we're telling them: "Your decades of specialist knowledge are not only useless, they're actually unhelpful; your sophisticated techniques are worse than generic methods." It's difficult for people who are used to that old type of science. They spend so much time discussing whether an idea makes sense. They check the visualizations and noodle over it. That is all actively unhelpful.