Cross-Validation, Regularized Regression & Bias-Variance Tradeoff

Isaac Laughlin

May 2, 2016

Today's schedule

11:30 - 12:30 This lecture

12:30 - 3:30 Individual Sprint and Lunch (in whatever order your stomach prefers)

3:30 - 4:30 Afternoon lecture

4:30 - 7? Paired Sprint (cross-cohort pairs are possible!)

Objectives

At the end of the lecture you should:

AM

- State the purpose of Cross Validation
- Be able to describe the two kinds of model error.
- Describe how to select a model using CV
- Give the reason for using k-fold CV
- Explain k-fold Cross Validation
- Explain the training, validation, testing data sets and what they are used for

PM

- Be able to state the purpose of Lasso and Ridge regression, and compare the two choices
- Build test error curves for regularized regression
- Build and interpret learning curves

The (two) purposes of CV

1. Find the best model to use.
2. Predict how well that model will perform on unseen data.

Our general problem:

$$y = f(X) + \epsilon$$

- Eventually we will have many fs and Xs to choose from
- So far, we only have one tool, linear regression, but still many choices

Comparing linear regression models

Imagine we have just a single variable x_1.

We can create a linear regression

$$y^{(1)} = \beta_0 + \beta_1 x_1$$

or

$$y^{(2)} = \beta_0 + \beta_1 x_1$$

or

$$y^{(3)} = \beta_0 + \beta_1 x_1 + \beta_2 x_1^2 + \beta_2 x_1^3 + \ldots$$

Poll:

http://pollev.com/galvanizedsi351

Business

Scenario

You are building a house-flipping company which will scrape zillow for undervalued houses and buy them to flip. You're going to make money like this:

$$price_{future} = f(X)$$

$$\sum_i price_{future,i} - price_{today,i}$$

What are the risks to your business scheme?

Issues

- Coefficients of linear regression minimize squared error for given X
- p-values tell us whether we can reject the idea that our coefficient could be 0
- $R^2 = f(X, \beta)$ ditto AIC, BIC

Of these mean squared error is the only one that seems able to answer the question "How will my model perform on data it hasn't seen?" But, the MSE the model achieves is likely to be *optimistic*.

Overfitting

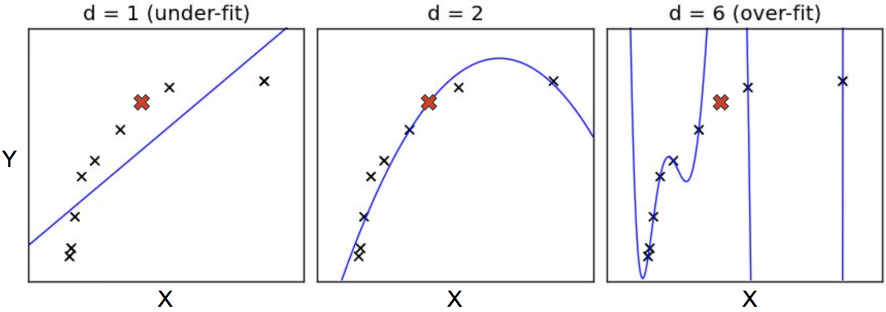

Figure 1:Overfitting example

Question:

What's wrong with each of the models above?

Underfitting and Overfitting

Both are a failure to capture the true relationship between y and X.

Underfitting

- Model does not fully capture the signal in X
- Insufficently flexible model

Overfitting

- Model erroneously interprets noise as signal
- Overly flexible model

Two kinds of model error: Bias and Variance

Typically we refer to the error caused by under/overfitting by their statistical names *bias* and *variance*.

Good news

Bias and Variance describe all reducible sources of error in a model

Bias and Variance

$$Y = f(X) + \epsilon$$

$$\hat{Y} = \hat{f}(X)$$

$$E[(y_0 - \hat{f}(x_0))^2] = ... = Var(\hat{f}(x_0)) + Bias^2(\hat{f}(x_0)) + Var(\epsilon)$$

CV for Model Selection

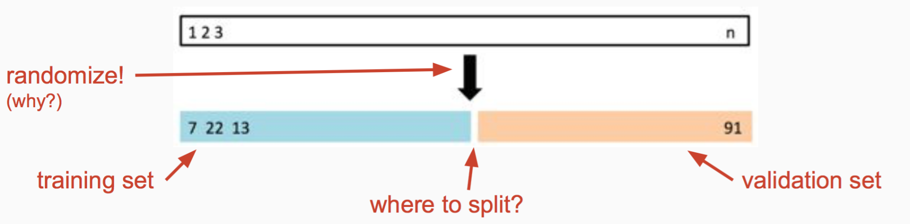

Figure 2:Training Validation Split Diagram

CV for Model Selection

Basic procedure:

1. Split into training/validation sets
2. Use training set to train several models of varying complexity
3. Evaluate each model using the validation set
4. Keep the model that performs best over the validation set

So what should we do?

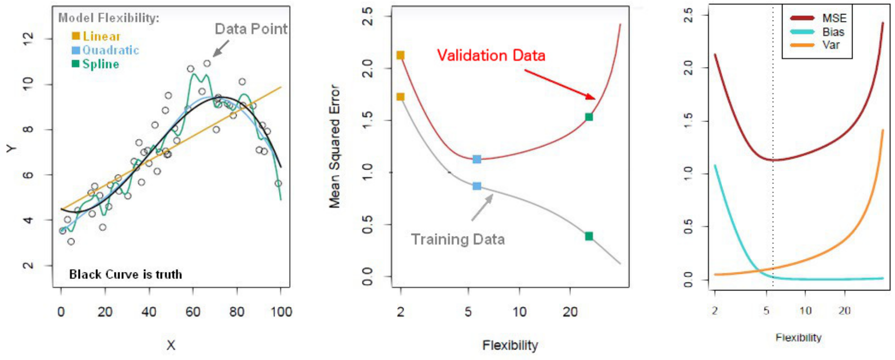

Figure 3:Bias Variance

Example use on cars data

Figure 4:HP vs. MPG

Train-Test error curves

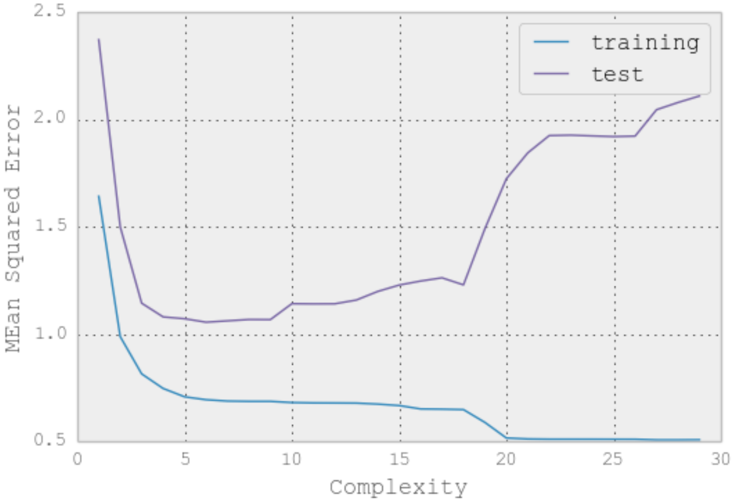

Figure 5:Train-Test Errors

Train-Test Errors

http://pollev.com/galvanizedsi351

Potential Problem

Discuss

Given the train-validation split described, why might we doubt that our chosen model is truly the best? *Hint: what if we're unlucky?*

K-Fold Cross Validation

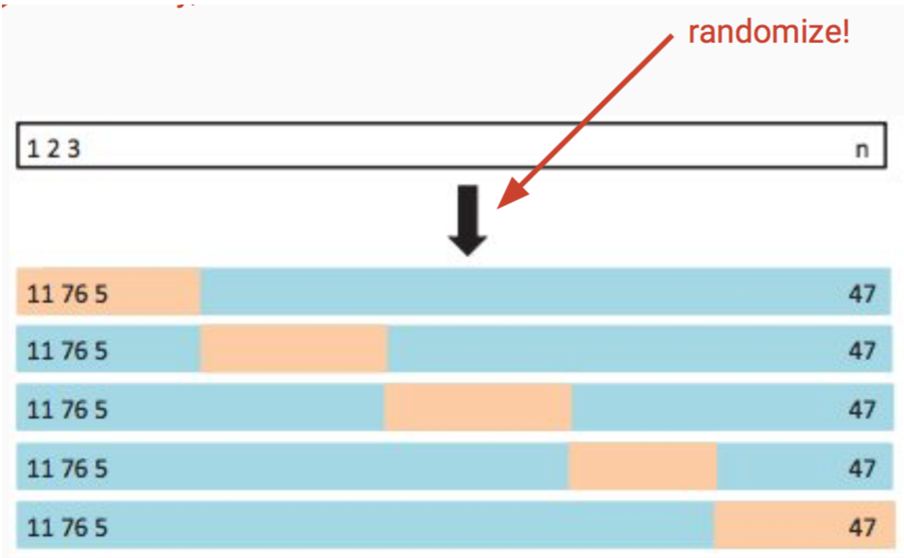

Figure 6:K-Fold CV Diagram

K=3-Fold Example

1. Given a dataset D, partition D into 3 parts, D_1, D_2, D_3.
2. For each D_i
    1. Mark D_i as the validation set
    2. Mark the remaining $D_{j \neq i}$ as the training set.
    3. Train candidate models on the training set.
    4. Append the model errors to a list.
3. Compute the mean model errors.
4. Select the model with the lowest mean model error.
5. Retrain the model on all data.

A subtle, but important problem

Another problem

Is the error from the validation sets actually the error that I can expect on unseen data? *Hint: if I iteratively try many models, and choose the ones that have the best error on the validation data, is my validation set representative of unseen data?*

Train-Validation-Test

No.

- Just as the errors observed in training are conservative, because those errors are realized on data the model had an opportunity to learn
- *Similarly* the errors observed in CV are conservative, because those errors are realized on data that the model-selection procedure had an opportunity to learn

CV Workflow

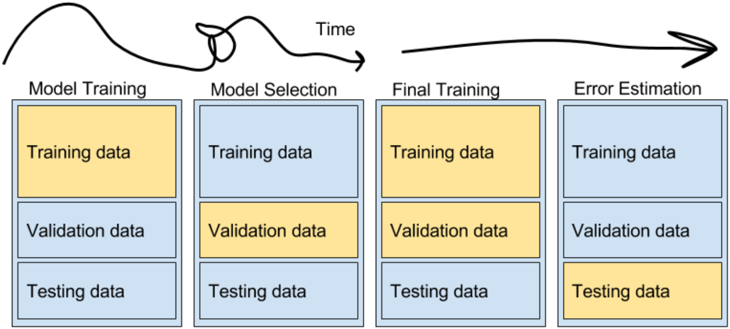

Figure 7:CV Workflow

Advanced CV techniques:

- Leave One Out CV (like k-fold, where k=n)
- Stratified CV (random within subgroups)
- Time-Series CV (useful for cases when observations are not independent in time)

Final Word

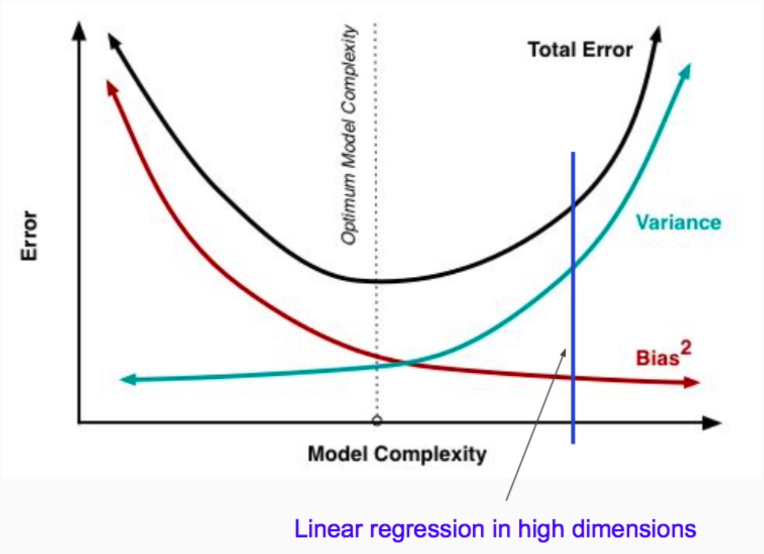

Figure 8:Optimum Model Complexity

Regression for Prediction