Cross Validation

Cary Goltermann

Galvanize

2016

Overview

Cross Validation

- Purpose
- Types of Fit
- Kinds of Model Error
- Model Selection
- K-Fold Cross Validation

Overview

Cross Validation

- Purpose
- Types of Fit
- Kinds of Model Error
- Model Selection
- K-Fold Cross Validation

Purpose of Cross Validation

There are two main reasons we use cross validation:

1. Find the best model to use.

Purpose of Cross Validation

There are two main reasons we use cross validation:

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

Modeling in an Equation

The problem that we are faced with when trying to create at predictive model coming up with a function that turns data into targets. Or in math:

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

where ϵ is error that our model doesn't account for.

Modeling in an Equation

The problem that we are faced with when trying to create at predictive model coming up with a function that turns data into targets. Or in math:

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

where ϵ is error that our model doesn't account for.

Eventually we will have many ways to make our fs and ways to format our data into different looking Xs. Cross validation is the tool that we use to decide between all of these choices.

Comparing Linear Regressions

Imagine we have just a single variable x_1. We can create a linear regressions with different powers of this variable:

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

or

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

or

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

Comparing Linear Regressions

Imagine we have just a single variable x_1. We can create a linear regressions with different powers of this variable:

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

or

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

or

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

Question

Which one of these equations has the most flexibility in describing a relationship between x_1 and y?

Business Example

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:

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

$$Total\ expected = \sum_i price_{future,i} - price_{today,i}$$

Business Example

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:

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

$$Total\ expected = \sum_i price_{future,i} - price_{today,i}$$

What are the risks to your business scheme?

Linear Regression - How Do We Choose a Model?

- Coefficients of linear regression minimize square error for given X.
- P-values are for the test of H_0: $\beta_i = 0$.
- In assessing fit we have $\{R^2, AIC, BIC\} = f(X, \beta)$.

Linear Regression - How Do We Choose a Model?

- Coefficients of linear regression minimize square error for given X.
- P-values are for the test of H_0: $\beta_i = 0$.
- In assessing fit we have $\{R^2, AIC, BIC\} = f(X, \beta)$.

Which of these help us answer the question: ”How will my model perform on data that it hasn’t seen?”

Linear Regression - How Do We Choose a Model?

- Coefficients of linear regression minimize square error for given X.
- P-values are for the test of H_0: $\beta_i = 0$.
- In assessing fit we have $\{R^2, AIC, BIC\} = f(X, \beta)$.

Which of these help us answer the question: "How will my model perform on data that it hasn't seen?"

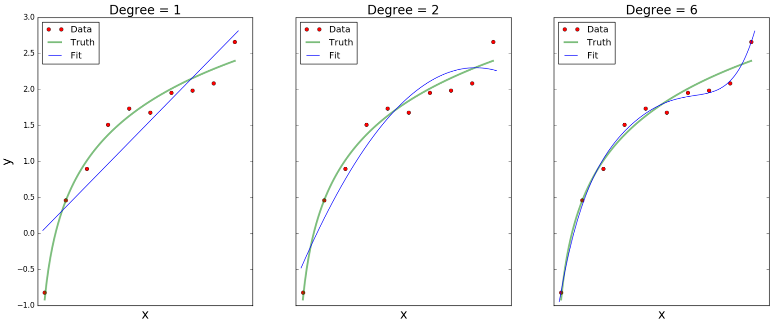
MSE is the only one that seems like it could be helpful here. Even then though, that measure is likely to be optimistic (Gauss-Markov Theorem).

Overview

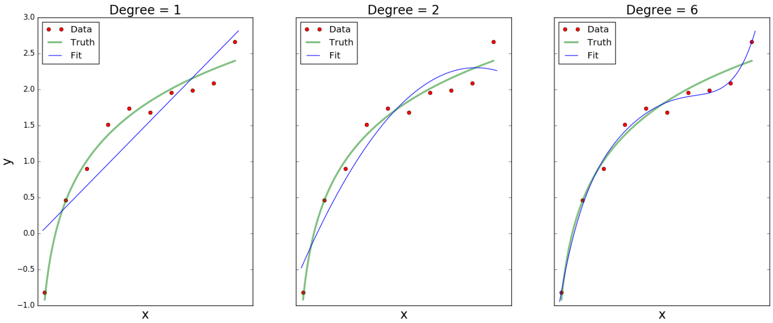
Cross Validation

- Purpose
- Types of Fit
- Kinds of Model Error
- Model Selection
- K-Fold Cross Validation

Types of Fit


Degree = 1
Degree = 2
Degree = 6
Data
Truth
Fit
y
x
3.0
2.5
2.0
1.5
1.0
0.5
0.0
−0.5
−1.0

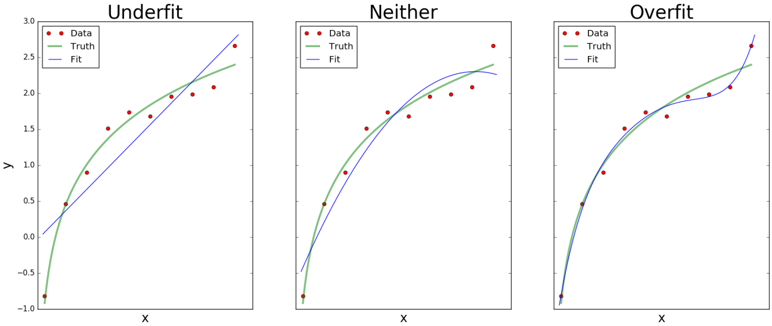
Types of Fit



Question

Describe the fit of each model above.

Types of Fit



Question

What is (potentially) wrong with each of these models?

Over and Underfitting

Both of these problems with fit come down to a failure of the true relationship between y and X.

Over and Underfitting

Both of these problems with fit come down to a failure of the true relationship between y and X.

Underfitting

- Model does not fully capture the signal in X.
- Model is not flexible enough.

Over and Underfitting

Both of these problems with fit come down to a failure of the true relationship between y and X.

Underfitting

- Model does not fully capture the signal in X.
- Model is not flexible enough.

Overfitting

- Model attributes to signal that which is truly noise.
- Model is too flexible.

Overview

Cross Validation

- Purpose
- Types of Fit
- Kinds of Model Error
- Model Selection
- K-Fold Cross Validation

Bias and Variance

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

Bias and Variance

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

One important thing to note is that together bias and variance make up all **reducible** sources of error in a model.

Bias and Variance

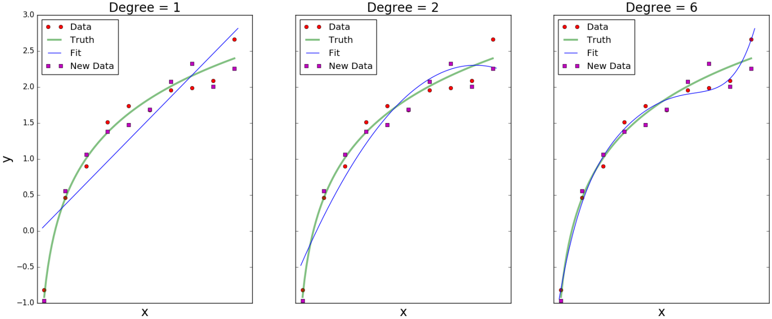
Typically we refer to the error caused by under and overfitting by their statistical names: **bias** and **variance**.

One important thing to note is that together bias and variance make up all **reducible** sources of error in a model.

$$
\begin{aligned}
y &= f(X) + \epsilon \\
\hat{y} &= \hat{f}(X) \\
E[(y - \hat{f}(x))^2] &= Var(\hat{f}(x)) + Bias^2(\hat{f}(x)) + Var(\epsilon)
\end{aligned}
$$

$Bias(\hat{f}(x)) = E[\hat{f}(x) - f(x)]$ $\qquad$ $Var(\hat{f}(x)) = E[\hat{f}(x)^2] - E[\hat{f}(x)]^2$

Bias-Variance Graphically

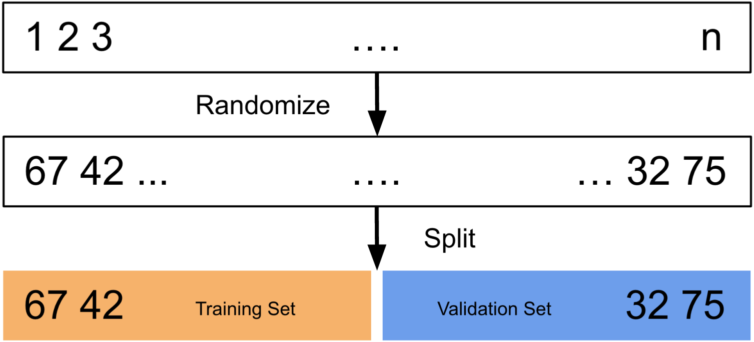


Overview

Cross Validation

- Purpose
- Types of Fit
- Kinds of Model Error
- Model Selection
- K-Fold Cross Validation

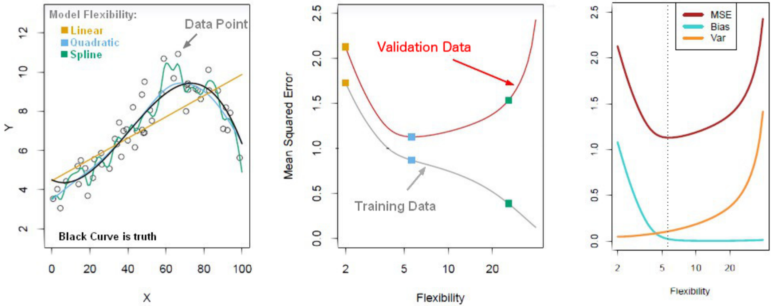
Training/Validation Split



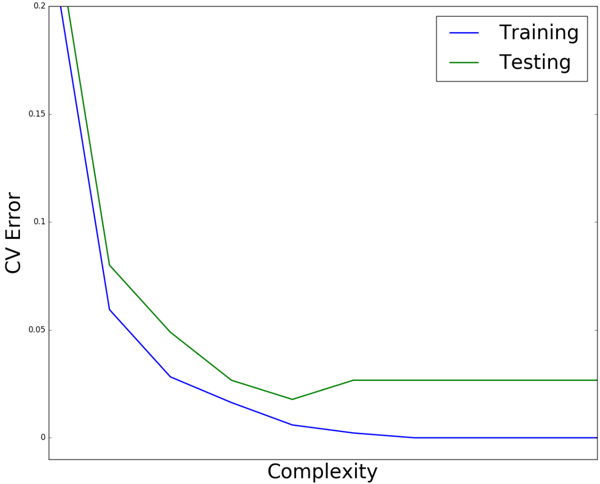
Procedure

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

How to Use


Model Flexibility:
■ Linear
■ Quadratic
■ Spline
Data Point
Black Curve is truth
Y
2
4
6
8
10
12
0
20
40
60
80
100
X
Mean Squared Error
0.0
0.5
1.0
1.5
2.0
2.5
Validation Data
Training Data
2
5
10
20
Flexibility
MSE
Bias
Var
0.0
0.5
1.0
1.5
2.0
2.5
2
5
10
20
Flexibility

Train Test Errors



Discussion

Question

Given the train-validation split procedure just described, why might we doubt that our chosen model is truly the best?

Discussion

Question

Given the train-validation split procedure just described, why might we doubt that our chosen model is truly the best?

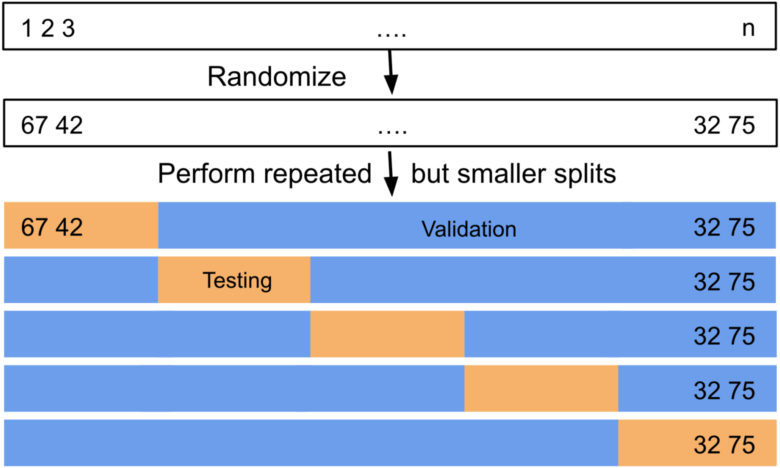
Hint: what if we're "unlucky"?

Overview

Cross Validation

- Purpose
- Types of Fit
- Kinds of Model Error
- Model Selection
- K-Fold Cross Validation

K-Fold Cross Validation



K-Fold Example, $K = 3$

Given a dataset D,

1. For each candidate model:
    1. 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 errors for each of the model error lists created in the last step.
2. Select the model with the lowest mean error.
3. Retrain the model on entire dataset, D.

How Good is K-Fold CV?

Question

How comparable is the error metric we get from K-Fold CV error we can expect on unseen data?

How Good is K-Fold CV?

Question

How comparable is the error metric we get from K-Fold CV error we can expect on unseen data?

Hint: In train-validation split, what happened when our validation set wasn't representative of unseen data?

Train-Validation-Test

Just as the errors observed in training are conservative because those errors apply to the data that the model had the opportunity to learn from during training.

Similarly, the errors observed in cross validation are conservative because those errors are realized on data that the model selection process got to see during training.

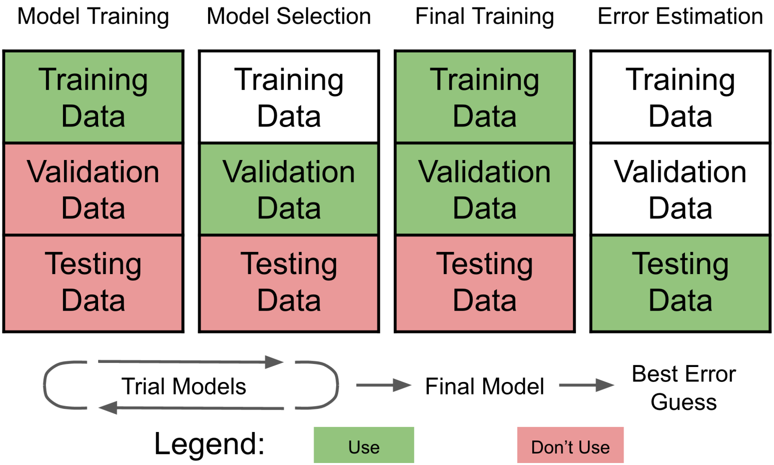
Train-Validation-Test

Just as the errors observed in training are conservative because those errors apply to the data that the model had the opportunity to learn from during training.

Similarly, the errors observed in cross validation are conservative because those errors are realized on data that the model selection process got to see during training.

How, then, do we get a good idea on how our model will perform "in the wild"?

Cross Validation Workflow



Model Training	Model Selection	Final Training	Error Estimation
🟩 Training Data	Training Data	🟩 Training Data	Training Data
🟥 Validation Data	🟩 Validation Data	🟩 Validation Data	Validation Data
🟥 Testing Data	🟥 Testing Data	🟥 Testing Data	🟩 Testing Data

Legend: 🟩 Use 🟥 Don't Use

Other CV Techniques

Leave One Out CV — Like K-Fold, except we set $k = n$.

Stratified K-Fold — Like K-Fold, except proportion of subgroups is maintained within each fold.

Time-Series CV — Never train on data from the future.