Assessing Model Accuracy

Nate Wells

Math 243: Stat Learning

September 9th, 2020

Outline

In today's class, we will. . .

Outline

In today's class, we will. . .

- Discuss the Mean Squared Error as measure of model accuracy

Outline

In today's class, we will...

- Discuss the Mean Squared Error as measure of model accuracy
- Investigate the Bias-Variance trade-off

Outline

In today's class, we will. . .

- Discuss the Mean Squared Error as measure of model accuracy
- Investigate the Bias-Variance trade-off
- Analyze data from the 'guess my age' activity

Section 1

Mean Squared Error

How do we measure quality of a model?

Goal: Devise a quantitative measurement of error for a model. Then develop a general algorithm for finding the model that minimizes this measure of error.

How do we measure quality of a model?

Goal: Devise a quantitative measurement of error for a model. Then develop a general algorithm for finding the model that minimizes this measure of error.

- For regression, the most common measure of error is the **Mean Squared Error** (MSE):

$$\mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

  where $\hat{f}$ is the model, the x_i are the observed predictor values, and the y_i are the corresponding observed response values.

How do we measure quality of a model?

Goal: Devise a quantitative measurement of error for a model. Then develop a general algorithm for finding the model that minimizes this measure of error.

- For regression, the most common measure of error is the **Mean Squared Error** (MSE):

$$\mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

  where $\hat{f}$ is the model, the x_i are the observed predictor values, and the y_i are the corresponding observed response values.

- Under what circumstances is MSE small?

How do we measure quality of a model?

Goal: Devise a quantitative measurement of error for a model. Then develop a general algorithm for finding the model that minimizes this measure of error.

- For regression, the most common measure of error is the **Mean Squared Error** (MSE):

$$\mathrm{MSE}(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \hat{f}(x_i)\right)^2$$

  where $\hat{f}$ is the model, the x_i are the observed predictor values, and the y_i are the corresponding observed response values.

- Under what circumstances is MSE small?
- What are the problems with trying to minimize MSE on the set of observed data $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$?

Training and Test Data

Training Data is the collection of data we use to build our model. Often, it is a subset of all data we have available.

Training and Test Data

Training Data is the collection of data we use to build our model. Often, it is a subset of all data we have available.

Test Data is the collection of data on which we assess the accuracy of our model. It should be distinct from the training data.

Training and Test Data

Training Data is the collection of data we use to build our model. Often, it is a subset of all data we have available.

Test Data is the collection of data on which we assess the accuracy of our model. It should be distinct from the training data.

Use a model-building algorithm that uses **training data** in order to minimize MSE on a large number of previously unobserved **test data** points (x_0, y_0), i.e. minimize

$$\mathrm{Ave}\left(y_0 - \hat{f}(x_0)\right)^2$$

Training and Test Data

Training Data is the collection of data we use to build our model. Often, it is a subset of all data we have available.

Test Data is the collection of data on which we assess the accuracy of our model. It should be distinct from the training data.

Use a model-building algorithm that uses **training data** in order to minimize MSE on a large number of previously unobserved **test data** points (x_0, y_0), i.e. minimize

$$\mathrm{Ave}\left(y_0 - \hat{f}(x_0)\right)^2$$

- If we have training and test data, we can construct a number of models and compare their performance on the test data in order to select the best model

An Example

Suppose we wish to predict students' final exam scores Y based on their first midterm scores X. We have data from two previous classes.

An Example

Suppose we wish to predict students' final exam scores Y based on their first midterm scores X. We have data from two previous classes.

We don't care about how well our model predicts exam scores for the previous classes. We want to know how well it predicts future scores.

An Example

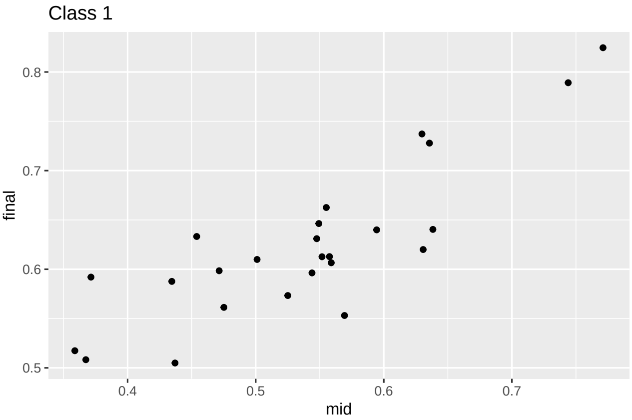
Suppose we wish to predict students' final exam scores Y based on their first midterm scores X. We have data from two previous classes.

We don't care about how well our model predicts exam scores for the previous classes. We want to know how well it predicts future scores.

- Use the first class as training data
- Use the second class as test data

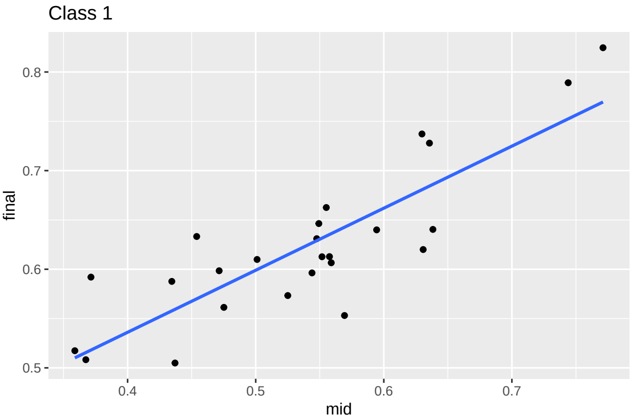
Training Set

```
##
##
scores %>% ggplot( aes(x = mid, y = final)) +
  geom_point()+labs(title = "Class 1")
```



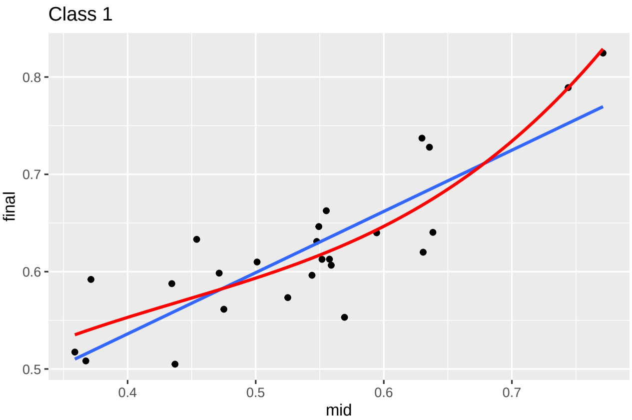
Model 1

```
##
scores %>% ggplot( aes(x = mid, y = final)) + geom_point()+
  labs(title = "Class 1") +
  geom_smooth( method = "lm" , se = FALSE)
```

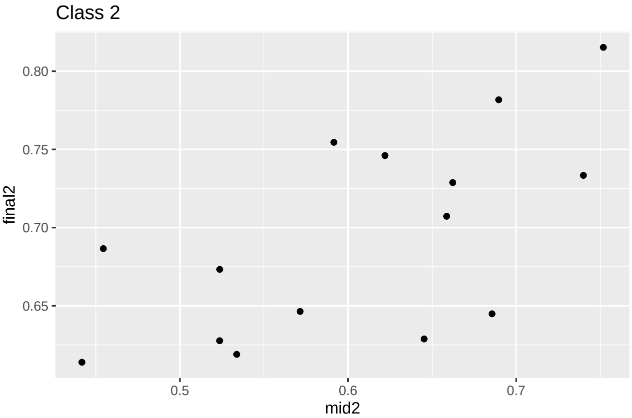


Model 1 and 2

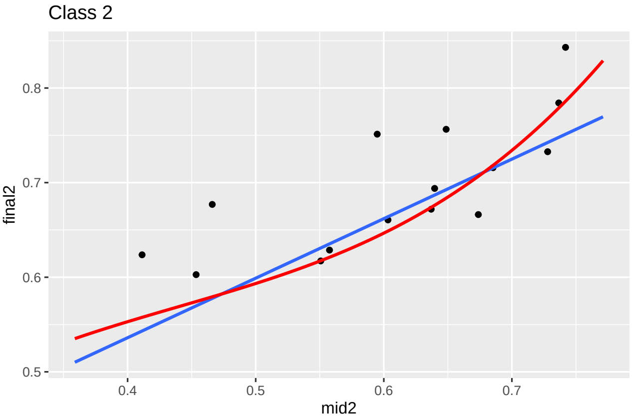
```
scores %>% ggplot( aes(x = mid, y = final)) + geom_point() +
  labs(title = "Class 1") +
  geom_smooth( method = "lm" , se = FALSE) +
  geom_smooth( method = "lm" ,formula = y ~ poly(x, 3), se = FALSE, color = "red")
```



Test Set



Test Set with models



Minimize MSE subject to model shape

What if no test is available?

Minimize MSE subject to model shape

What if no test is available?

- Recall the setting of simple linear regression from Math 141.

Minimize MSE subject to model shape

What if no test is available?

- Recall the setting of simple linear regression from Math 141.

We can choose a model that minimizes MSE on the training set, subject to constraints (i.e. restricting to linear, quadratic, exponential models)

Minimize MSE subject to model shape

What if no test is available?

- Recall the setting of simple linear regression from Math 141.

We can choose a model that minimizes MSE on the training set, subject to constraints (i.e. restricting to linear, quadratic, exponential models)

But no guarantee that model which minimizes MSE on training data will also do so on test data.

Minimize MSE subject to model shape

What if no test is available?

- Recall the setting of simple linear regression from Math 141.

We can choose a model that minimizes MSE on the training set, subject to constraints (i.e. restricting to linear, quadratic, exponential models)

But no guarantee that model which minimizes MSE on training data will also do so on test data.

In fact, when selecting a complex model that minimizes MSE on the training data, the test MSE will often be very large!

Demo in RStudio

See .Rmd file (Wednesday 9-9 Demo) on the schedule page of the course website

Section 2

Bias-Variance Trade-off

Training vs Test MSE

Suppose we consider a variety of model shapes to predict Y, with each model of increasing complexity. What happens to the training MSE and the test MSE as model complexity increases?

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

Average MSE can be decomposed as the sum of 3 quantities:

$$\text{AveMSE}(\widehat{\text{f}}) = \text{Var}(\hat{f}) + \left[\text{Bias}(\hat{f})\right]^2 + \text{Var}(\epsilon)$$

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

Average MSE can be decomposed as the sum of 3 quantities:

$$\text{AveMSE}(\widehat{\text{f}}) = \text{Var}(\hat{f}) + \left[\text{Bias}(\hat{f})\right]^2 + \text{Var}(\epsilon)$$

- A proof is given in *The Elements of Statistical Learning*

MSE Decomposition

The U-curve for test MSE is a result of competition between two sources of error in a model

Average MSE can be decomposed as the sum of 3 quantities:

$$\mathrm{AveMSE}(\widehat{\mathrm{f}}) = \mathrm{Var}(\hat{f}) + \left[\mathrm{Bias}(\hat{f})\right]^2 + \mathrm{Var}(\epsilon)$$

- A proof is given in *The Elements of Statistical Learning*

To minimize MSE, we need to *simultaneously* minimize both variance and bias.

Bias and Variance

Variance refers to size of fluctuations in $\hat{f}$ if different training sets were used

Bias and Variance

Variance refers to size of fluctuations in $\hat{f}$ if different training sets were used

- What type of models tend to have low/high variance?

Bias and Variance

Variance refers to size of fluctuations in $\hat{f}$ if different training sets were used

- What type of models tend to have low/high variance?

Bias refers to the size of fluctuations produced by the difference between model shape assumptions and model shape reality

Bias and Variance

Variance refers to size of fluctuations in $\hat{f}$ if different training sets were used

- What type of models tend to have low/high variance?

Bias refers to the size of fluctuations produced by the difference between model shape assumptions and model shape reality

- What type of models tend to have low/high bias?

The Trade-off

What is the problem?

The Trade-off

What is the problem?

How do we solve it?