Bias and Variance Tradeoff

Week 3 | Lesson 2.1

LEARNING OBJECTIVES

After this lesson, you will be able to:

- Define Bias and Variance
- Explain models error in terms of bias and variance
- Interpret output of models we have learned so far

Where are we in the Data Science Workflow?

http://bit.ly/GA_DS_Workflow

Review: Linear and Multilinear Regression + Scikit-learn

- Define,
- Explain, or
- Recall

Past, Present, Future with Linear Regression?

You can now:

- find linear models
- fit linear models
- tune linear models
- review mean squared errors (MSE)

We will now continue to learn:

- Analyze how well models fit the data
- How we can make good choices and better models

Usually we attempt to quantify the *error* in our models, the difference between predictions and true values, and we'll learn multiple ways to do so, i.e. MSE.

Why Do We care about Bias and Variance?

Understanding how different sources of error

Helps us improve the data fitting process resulting in potentially more accurate models.

We define bias and variance in three ways:

- conceptually,
- graphically and
- mathematically.

Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Introduction: Bias-Variance Tradeoff

Conceptual Approach

Error Due to Bias: difference between expected (average) prediction and actual value

Error Due to Variance: variability of model prediction for a given point

Graphically - Bias and Variance comparison

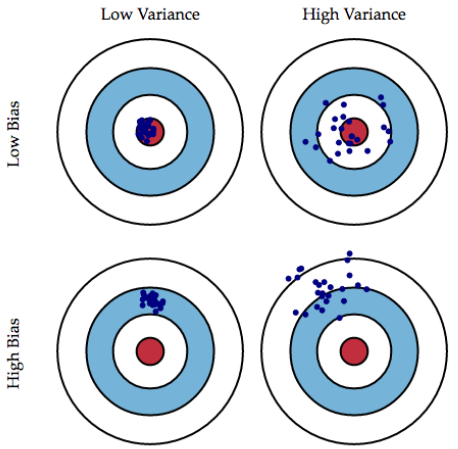


Fig. 1 Graphical illustration of bias and variance.

Center of the target is a model, where the bullseye is a perfect prediction.

Each hit is synonmous to a run of your model.

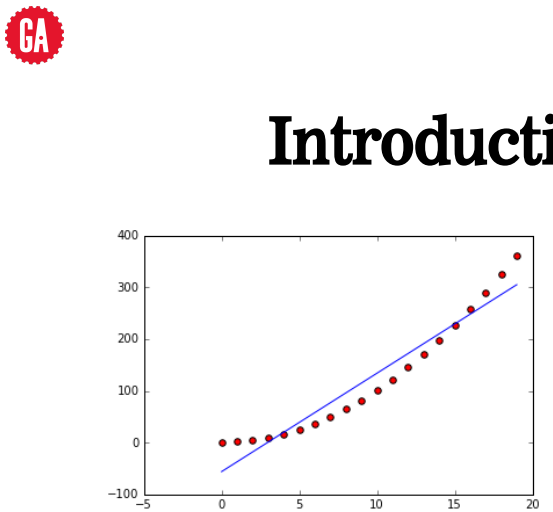
What happens as different scenarios are repeated?

Mathematical Definition of Bias and Variance

Generally, we describe our modeling relationships as: `Y = f(x) + e`, where e is the error term or noise.

This error or noise are decomposed to: `Err(x) = Bias^2 + Variance + Irreducible Error`

Now, Let's review Bias and Variance Individually



Introduction -- What is Bias?

Example, a line is *not* a good global approximation for a parabola:

In this case our model has an opinion about the shape of the data that's not quite right, so we say it is *biased*.

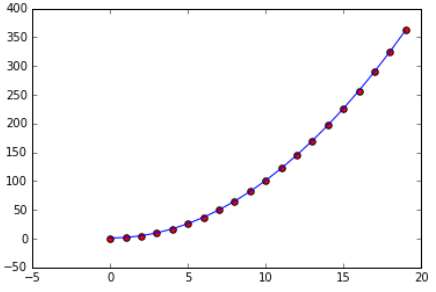
More data isn't going to help, and a different sampling of data won't either. Our model is simply *biased* and won't ever be a good fit to the data.

Let's see how to calculate the bias so that we can make this idea precise.

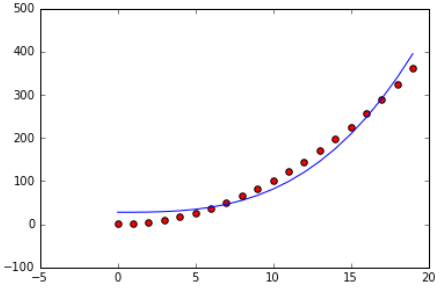
As modelers, our goal is to find the function `F` that minimizes our error.

Parabola example again - with much less bias

If we tried to fit a model like `y=a x + b x^2 + e` to our parabolic model, we could find a model with much less bias (on average):



On the other hand if we had tried a cubic model without a quadratic or linear term, i.e. `y = a x^3 + b1`, we'd still have a lot of bias (on average):



Find the Sweet Spot for Bias

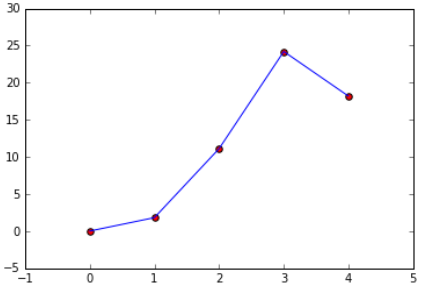
The sweet spot is where the bias is low and the model is the right complexity.

Data will not always comply since there are:

- inherent errors,
- small fluctuations in the data can introduce artifacts that our model will attempt to fit.
- too many parameters will end up fitting a model too close to the data from the same source (we'll share techniques avoid overfitting in later lessons)

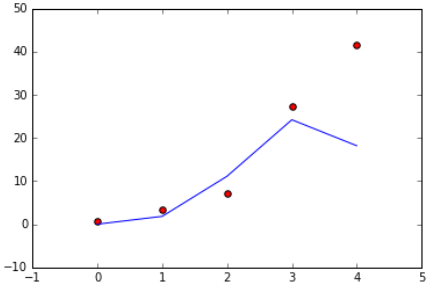
Variance -- Fitting Data Points

(1) For example, with the following data points we could try to fit a line or a higher dimensional polynomial:



The polynomial fits nearly perfectly but the underlying data may not actually come from such a shape.

(2) If we take data from the same source and repeat the fit:



In this case the model we fit on the first sample overfit the data and doesn't fit the second sample well even though they are from the same source. We call this source of error the error due to *variance*.

Check: We've described two sources of error:

- Bias
- Variance

What is the difference between them?

The Tradeoff of Errors

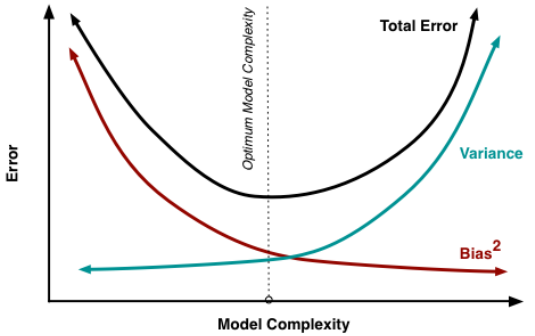
The third source of error is error inherent in our model. This is the error term `e` which has variance `\sigma^2`.

Combined sources of error in one equation gives us the squared error:

$$\mathrm{E}\Big[\big(y - \hat{f}(x)\big)^2\Big] = \mathrm{Bias}\big[\hat{f}(x)\big]^2 + \mathrm{Var}\big[\hat{f}(x)\big] + \sigma^2$$

Trade-off Visualized

This equation is why bias and variance are often described as a tradeoff. This relationship is more accurate to describe the relationship with an plot:



- There is usually a sweet spot in "model space" where bias and variance are both small.
- Note too that there's nothing we can do to reduce the inherent noise in the data.
- The tradeoff is in the rates of bias and variance as a function of model complexity.

Error measures

- MSE, Mean of Squared Errors
- SSE, Sum of Squared Errors

$$s^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\mathbf{b})^2$$

What is sum of squared errors (SSE)?

- also known as, residual sum of squares and sum of squared residuals
- SSE is commonly decomposed into *bias* and *variance*.
- Generally,

```
Total Sum of Squares = Explained Sum of Squares + Residual Sum of Squares
```

- Conceptually, a model is biased if it makes assumptions about what the data *should* look like and misses the mark.

What is mean square errors (MSE)?

- aka out-of-sample mean squared error
- Measures the average of the squared errors or deviations between predicted and actual values
- Analogy to standard deviation

$$s^2 = \mathrm{MSE} = \frac{\mathrm{SSE}}{n-k} = \frac{1}{n-k}\sum_{i=1}^{n}(y_i - \mathbf{x}_i\mathbf{b})^2$$

Demo: Examples of Bias and Variance (15 mins)

Open the demo notebook with series of models and explore both the bias and the variance.

Bias and variance are properties that arise over many data samples and model fits.

Guided Practice: Explore Bias and Variance (20-30 mins)

We've seen that as the complexity in our models change so too do the bias and variance. Let's investigate from another angle using linear regression. Rather than change the model, we'll change the underlying data to be drawn from a higher dimensional model.

Your tasks are:

- Fill in the code to fit a linear regression to the data
- investigate the bias and variance as the data source changes in complexity

Get started with the Starter Code!

Independent Practice: Explore Bias and Variance (20-25 minutes)

Now let's look at some situations where adding more polynomial terms decreases both bias and variance.

Get started with the Starter Code!

Conclusion

There are three fundamental sources of error that arise when fitting a model to data:

- Bias
- Variance
- Inherent noise

We can't do much about the inherent noise in the data, but we can often reduce both bias and variance with good choices of models.

ADDITIONAL RESOURCES

- A nice exposition on Bias and Variance
- Bias-Variance Tradeoff on Wikipedia