Towards Data Science
Sharing concepts, ideas, and codes.

Overfitting vs. Underfitting: A Conceptual Explanation

Will Koehrsen Follow
Jan 27, 2018 · 6 min read

An example-based framework of a core data science concept

Say you want to learn English. You have no prior knowledge of the language but you've heard the greatest English writer is William Shakespeare. A natural course of action surely must be locking yourself in a library and memorizing his works. After a year of study, you emerge from your studies, travel to New York City, and greet the first person you see with "Good dawning to thee, friend!" In response, you get a look of disdain and a muttered 'crazy'. Unperturbed, you try again: "Dear gentlewoman, How fares our gracious lady?" Another failure and a hurried retreat. After a third unsuccessful attempt, you are distraught: "What shame what sorrow!". Shame indeed: you have just committed one of the most basic mistakes in modeling, overfitting on the training data.

In data science courses, an overfit model is explained as having high variance and low bias on the training set which leads to poor generalization on new testing data. Let's break that perplexing definition down in terms of our attempt to learn English. The model we want build is a representation of how to communicate using the English language. Our training data is the entire works of Shakespeare and our testing set is New York. If we measure performance in terms of social acceptance, then our model fails to generalize, or translate, to the testing data. That seems straightforward so far, but what about variance and bias?

You highlighted

Variance is how much a model changes in response to the training data. As

Top highlight

we are simply memorizing the training set, our model has high variance: it is highly dependent on the training data. If we read the entire works of J.K. Rowling rather than Shakespeare, the model will be completely different. When a model with high variance is applied on a new testing set, it cannot perform well because all it is lost without the training data. It's like a student that has memorized the problems in the textbook, only to be helpless when faced with real world faced problems.

You highlighted

Sometimes even grad students should go outside

Bias is the flip side of variance as it represents the strength of our assumptions we make about our data. In our attempt to learn English, we formed no initial model hypotheses and trusted the Bard's work to teach us everything about the language. This low bias may seem like a positive— why would we ever want to be biased towards our data? However, we should always be skeptical of data's ability to tell us the complete story. Any natural process generates noise, and we cannot be confident our training data captures all of that noise. Often, we should make some initial assumptions about our data and leave room in our model for fluctuations not seen on the training data. Before we started reading, we should have decided that Shakespeare's works could not literally teach us English on their own which would have led us to be cautious of memorizing the training data.

To summarize so far: bias refers to how much we ignore the data, and variance refers to how dependent our model is on the data. In any modeling, there will always be a tradeoff between bias and variance and when we build models, we try to achieve the best balance. Bias vs variance is applicable to any model, from the simplest to the most complex and is a critical concept to understand for data scientists!

You highlighted

. . .

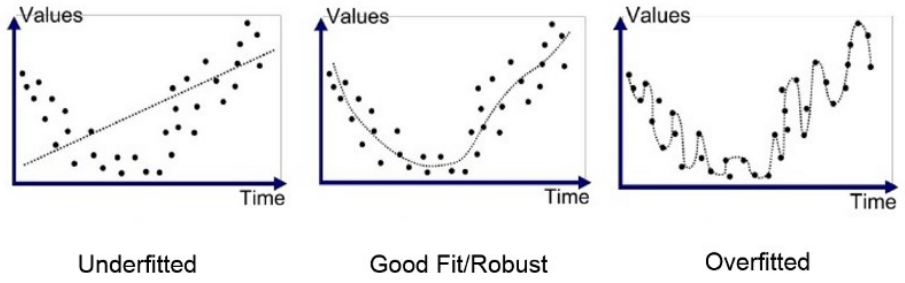
We saw that a model that overfits has high variance and low bias. What about the reverse: low variance and high bias? This is known as underfitting: instead of following the training data too closely, a model that underfits the ignores the lessons from the training data and fails to learn the underlying

You highlighted

relationship between inputs and outputs.

Let's think about this in terms of our example. Learning from our previous attempt to build a model of English, we decide to make a few assumptions about the model ahead of time. We also switch our training data and watch all episodes of the show *Friends* to teach ourselves English. To avoid repeating our mistakes from the first try, we make an assumption ahead of time that only sentences starting with the most common words in the language—the, be, to, of, and, a—are important. When we study, we do not pay attention to other sentences, confident we will build a better model.

After a long period of training, we again journey out onto the streets of New York. This time we fare slightly better, but again, our conversations go nowhere and we are forced to admit defeat. While we know some English and can comprehend a limited number of sentences, we failed to learn the fundamental structure of the language due to our bias about the training data. The model does not suffer from high variance but we overcorrected from our initial attempt and underfit!



Graphs of underfitting (high bias, low variance) vs overfitting (low bias, high variance) (Source)

What can we do? We paid strict attention to the data and we overfit. We ignored the data and we underfit. There has to be a way to find the optimal balance! Fortunately, there is a well-established solution in data science called validation. In our example, we used only a training set and a testing set. This meant we could not know ahead of time how our model would do in the real world. Ideally, we would have a "pre-test" set to evaluate our model and make improvements before the real test. This "pre-test" is known as a validation set and is a critical part of model development.

Our two failures to learn English have made us much wiser and we now decide to use a validation set. We use both Shakespeare's work and the *Friends* show because we have learned more data almost always improves a model. The difference this time is that after training and before we hit the streets, we evaluate our model on a group of friends that get together every week to discuss current events in English. The first week, we are nearly kicked out of the conversation because our model of the language is so bad. However, this is only the validation set, and each time we make mistakes we are able to adjust our model. Eventually, we can hold our own in conversation with the group and declare we are ready for the testing set. Venturing out in the real world once more, we are finally successful! Our model is now well suited for communication because we have a crucial element, a validation set for model development and optimization.

This example is necessarily simplified. In data science models, we use numerous validation sets because otherwise we end up overfitting to the validation set! This is addressed by means of cross-validation, where we split the training data into different subsets, or we can use multiple validation sets if we have lots of data. This conceptual example stills covers all aspects of the problem. Now when you hear about overfitting vs. underfitting and bias vs. variance, you have a conceptual framework to understand the problem and how to fix it!

. . .

Data science may seem complex but it is really built out of a series of basic building blocks. A few of those covered in this article are:

- **Overfitting:** too much reliance on the training data
- **Underfitting:** a failure to learn the relationships in the training data
- **High Variance:** model changes significantly based on training data
- **High Bias**: assumptions about model lead to ignoring training data
- Overfitting and underfitting cause poor **generalization** on the test set
- A **validation set** for model tuning can prevent under and overfitting

Data science and other technical fields should not be divorced from our everyday lives. By explaining concepts with real-world examples, we can put them into context. If we understand the framework, then we can fill in the details by using the techniques on problems. The next post will provide an example using graphs and metrics, so if you want a more solid backing, check it out. Until then, fare you well dear readers!

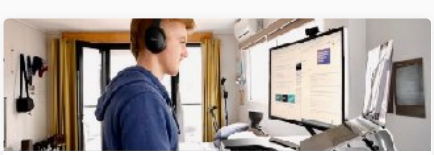
I welcome feedback and constructive criticism. I can be reached on Twitter @koehrsen_will.

I would like to thank Taylor Koehrsen (PharmD by the way) for helping me to sound less like an engineer in my writing!

Data Science | Machine Learning | Model | Education | Towards Data Science

Will Koehrsen
Data Scientist at Cortex Intel, Data Science Communicator

Towards Data Science
Sharing concepts, ideas, and codes.



More from Towards Data Science
12 Things I Learned During My First Year as a Machine Learning...

More from Towards Data Science
How a simple mix of object-oriented programming can sharpe...

More from Towards Data Science
What Separates Good from Great Data Scientists?

Daniel Bourke
Jul 6 · 11 min read ★ 2.7K

Tirthajyoti Sarkar
Jul 5 · 11 min read ★ 1.4K

Amadeus Magrabi
Jun 30 · 6 min read 1.3K