MODEL OPTIMIZATION

Bias, Variance, and Overfitting Explained, Step by Step

You have likely heard about bias and variance before. They are two fundamental terms in machine learning and often used to explain overfitting and underfitting. If you're working with machine learning methods, it's crucial to understand these concepts well so that you can make optimal decisions in your own projects. In this article, you'll learn everything you need to know about bias, variance, overfitting, and the bias-variance tradeoff.

BORIS GIBA
17 APR 2021 • 20 MIN READ

Last modified: 17 Apr 2021

Background image by Sora Shimazaki (link)

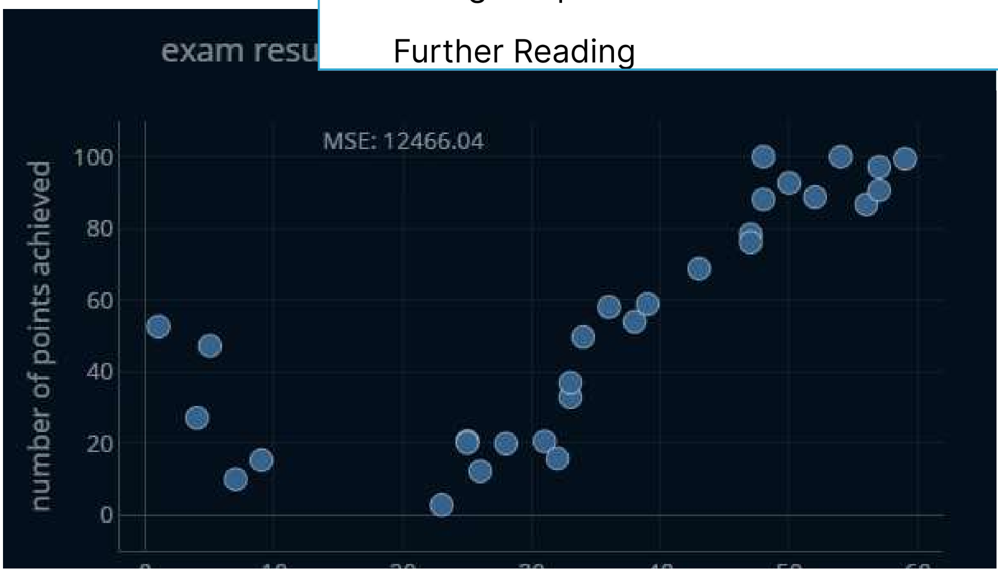
Outline

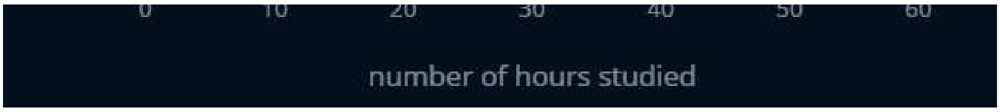
Bias and variance are very fundamental, and also very important concepts. Understanding bias and variance well will help you make more effective and more well-reasoned decisions in your own machine learning projects, whether you're working on your personal portfolio or at a large organization. In this article, you will learn what bias and variance are, what the so-called *bias-variance tradeoff* is, and how you can ma
machine learning pro
machine learning mo

Table of Contents

The Problem

Let's start by looking
teaching your friend
them a small dataset
dataset contains the
studying for the exa
(between 0 and 100)
You then tell your fri
points achieved base
dataset looks like thi





make interactive

Your friend is just getting started with machine learning and recently read an article about linear regression. Now they want to try out linear regression themselves! After some time, they come back to y
creating a model tha
model and make a se
predicted values alo
number of hours stu
easier to digest, the
function. The resulti

Table of Contents

make interactive

You look at the plot and notice a couple of things.

1. There definitely seems to be a relationship between the number of hours studied and the number of points achieved.
2. This relationship does seem somewhat linear, especially when considering the segment between 40 and 60 hours studied.
3. However, the linear relationship is not that apparent throughout the entire graph. When you take a look at the interval between 0 and 30 hours studied, it does seem to have mor
   the interval betwe

With linear regressio
linear function) to m
features (number of
points achieved). Bu
same trend through
off one half of the da
case, we need to use
regression model ca
we can perform a sm
regression instead. V
relationships in our d
of just linear ones. Y
they go ahead and t

Table of Contents

Imagine you are in th
have to find out the
model, meaning the maximum power which you will apply to your features. Is a degree of 2 or 3 already enough? Do you need a degree of 10 or higher? Pause for a second and think about this question. Below you can find an interactive visualization where you can take a look at all the possible polynomial regression plots for the degrees 1-20. After you've come up with a number in your head, take a look at the visualization and see if your intuition was right.*

Table of Contents

If you estimated the
congratulations! You
eyes, can tell that th
best fits the data. Bu
the degree of our po
always gets closer a
model with degree 2
the middle than the
degree is, the "wiggl
model with a higher
can take on more complicated function shapes. So maybe the model with degree=20 is truly the best one? Ok, let's take a step back.

We have all of these different functions and some can capture the relationship between our features and our target better than others. If we were to compare the models with degree=1 and degree=20 respectively, we would say that the one with degree=20 "captures the relationship better". But in a practical scenario, there has to be a better way of expressing this without having to say “captures this relationship better” every time. Wouldn’t it be more satisfying if we could somehow say that the model with degree=20 has a higher/lower *something* than the model with degree=1? We can do exactly that by introducing the notion of **bias**.

The Bias

We want the bias to express how well a certain machine learning model fits a particular dataset. So let’s try and come up with a formal defi
examples. Then, we
in statistics and whe
from statistics. Our c

> *The bias of a spe*
> *on a specific data*
> *learning model ca*
> *the features and*
> *bias of any one m*
> *particular model*
> *for any number o*

That seems like a rea

Practical Examp

Let’s take a look at th
them using our new
polynomial regression models with degrees 1,4, and 15 respectively. Below is the graph showing our dataset and the predictions of the model with a degree of 1.

Table of Contents

As you see, this mod
predicted values are
therefor we can say
because it does not
the number of points
studied.

In practice, we proba
displaying our predic
bias is based on how
want to compare two
concrete way of mea
we often use the (tra
make deductions ab
based task in front o
squared error to measure the performance of a specific model. To make the results a bit easier to interpret, let's also add a square root on top so we can immediately tell by how much the average point is approximately off. If you are wondering why we don't use a different metric like the mean absolute error or why we don't use the root mean squared error by default, all of this is explained in more detail in this segment of the article about linear regression.

Table of Contents

If we compute the root mean squared error (RMSE) for the predictions of this model, we get a value of 19.8. This means that one single point can be off by up to ~19.8, i.e. if a student in truth achieved 80 points, our model might only give them 60, which would probably not make the student very happy. With this we can capture the following behavior:

- large training error → large bias

Let's now look at the model with degree 4:

Table of Contents

This model predicts our y a lot better than the first one. It still makes some errors, but it does its job a lot better than the first model. If we compute the RMSE for the predictions of this model, we get 7.68. It's about 2.5 times better than the previous model! If a student in truth achieved 80 points, our model might only give them 72 (or up to 88), which still is not ideal, but it's considerably better than the previous model. We achieved this decrease in error by just moderately increasing our degree. We've only increased the degree from 1 to 4, but managed to reduce our error by a factor of around 2.5! So this increase in complexity was certainly worth it. Because our model has a rather small error, we can say that it has a small bias since it does its task relatively well. With this we can capture the following behavior:

- small training error → small bias

Let's now take a closer look at the model with degree 15:

Table of Contents

This model does look a bit weirder than the previous two, but it does in fact predict our data even better than the second one. If we compute the RMSE for the predictions of this model, we get 4.93. It's about 1.5 times lower than the previous error. However, we did increase our degree from 4 to 15, which is quite a lot. So this increase in complexity certainly is not as good of a deal as the previous one. But a lower is a lower error, right? So technically, this model is the best of the bunch.

Because our model has a very low error, we can say that it has a very low bias since it does its task very well. With this we can capture the following behavior:

- very small training error → very small bias

Revisiting the Definition of Bias

Ok, so we have defir
looked at a couple o
error is correlated wi
the lower the bias. T
the bias.

When talking about t
talk about one mode
always talked about
model and one partic

Let's now look at the
to see if our previous
Wikipedia:

> *Statistical bias is*
> *of its results whe*
> *differs from the t*
> *being estimated.*

Table of Contents

If we look back on how we introduced the bias, this makes sense. We compared the predictions of our model with the actual values present in our dataset. But the real definition is worded just a tiny bit differently, and contains two very important words:

> *Statistical bias is a feature of a statistical technique or of its results whereby the expected value of the results differs from the* ***true underlying*** *quantitative parameter being estimated.*

This means that the bias is a way of describing the difference between the actual, *true* relationship in our data, and the one our model learned. In our examples, we've looked at the error between our predictions and the data points. Sure, that is a very sensible way to measure the bias of our machine learning models. But
accurately represent
features and the targ
contain noise, meani
true relationship of c
on it. Since I've creat
function*, I know the
and the target. Belov
particular function. I'
button with which yo
dataset.

Table of Contents

make interactive

If the dataset contains no noise at all, it means that the dataset represents the *exact* underlying relationship. But almost every dataset out there will have some noise in it. This might be because the dataset was not properly cleaned, or maybe because there is some intrinsic noise to the underlying problem itself. Often times it is simply not possible to get perfect, noiseless data. For this reason, there might be an *irreducible error*, an error, that no machine learning model (or even human) can un statistics compares t underlying relationsh Because oftentimes,

I can show you how the noise because I g mathematical functio sampled from actual shape of some math it's exactly the oppos usually have a seemi create a mathematic relationship in that d created a function an using that function. I would have no way o what part of the data

Table of Contents

This is the reason why we compare our dataset to our predictions. Ideally, we want our function to learn this true, underlying relationship, by only looking at the noisy dataset. Almost always we won't be able to create a model that perfectly matches this relationship (and we also don't have a way of checking if it did), but oftentimes we can generate a function that is *good enough*, assuming the dataset itself is *good enough*. So what does this mean? If our model creates a relationship that we think is not as sensible, 4/5 times it will be the dataset that is to blame, and not the model.

But how *do* we tell if our model is good?

Naturally, we are interested in keeping the bias as low as possible, because we want our model to make the best predictions possible. But if you take a look at the second and third model, would you really choose the third model over the second one? It does have a lower bias, so this is certainly an argument. But there is just something *off* about the third
model. Maybe it's the
the second one. May
too good at its task.
There has to be a wa
not so great, or whe
to do this, we are no
variance.

The Variance

With the bias, we no
machine learning mo
between feature and
always only conside
we bring in values ou
happens when we br
won't consider a dat
was trained to predi
dataset that is very *similar* to our current one? This could be the data of a different exam and different students from the same school. Alternatively, we can split our existing dataset into a *training*-set and a *testing*-set. We won't go into detail about why this makes sense in this particular article, but if you want to read more about this, I'd recommend you take a look at the article Training and Testing Datasets Explained, Step by Step.

Table of Contents

We want the variance to express how *consistent* a certain machine learning model is in its predictions when compared across *similar* datasets. So let's apply the same approach that we used for the bias and try to first come up with a formal definition of variance ourselves. It might look like the definition below.

> *The variance of a specific machine learning model trained on a specific dataset describes how much the performance of the machine learning model differs when evaluated c*
> *origin. So for our*
> *model would tell*
> *particular model*
> *change the value*
> *our performance*

Ok, now let's take a l
regression models th

Practical Exam

This time, I'll add a fu
the dataset when cli
information. First, we
model and the (pote
also track the relativ
MSE and the initial M
degree=1:

Table of Contents

make interactive

This model predicts
is also maintained if
the button a couple
and the R.DIFF, you'l
much. To be more ex
difference of 1000 R
you will receive a val
change our dataset,
which makes the mo
model performs poo
makes bad predictio
the performance of t
multiple similar datas
variance.

Table of Contents

With this we can cap

- small fluctuation c

In practice, we usually don't want to randomly alter our dataset and keep track of the relative difference to find out whether our variance is low or high. So how can we estimate the variance of a machine learning model? We look at the difference in performance between the training-set and the testing-set (or for that matter, the training set and any other dataset of similar origin). If this difference is high, so is the variance. If it is low, so is the variance.

Because the model with degree=1 has a high bias but a low variance, we say that it is *underfitting*, meaning it is not "fit enough" to accurately model the relationship between features and targets.

Let's now look at the model with degree 4:

Table of Contents

This model predicts good performance is dataset. If you again keep an eye on the RMSE and the R.DIFF, you'll notice that the values change a little bit, but not a lot. To be more exact, if you compute the average relative difference of 1000 RMSE values for slightly altered datasets, you will receive a value of ~13.6. This means that if we slightly change our dataset, our MSE will fluctuate by around 13.6%, which makes the model *somewhat* consistent. It's certainly not as consistent as the first model, but it's not terrible as well. In other terms, our model performs well, and does so pretty consistently. It always makes at least decent predictions for a dataset of this origin. Because the performance of this model is somewhat consistent across multiple similar datasets, we can say that it has a somewhat low variance.

With this we can capture the following behavior:

- medium fluctuation of the error → medium variance

Because the model v
variance, we say tha
balance between bia

Let's now look at the

Table of Contents

make interactive

This model predicts our data very well. However, this outstanding performance is not at all maintained when we change up our dataset a bit. If you press the button just a few

times and look at the RMSE and R.DIFF, chances are you'll see some larger numbers than before. These numbers also change quite a lot between clicks. To be more exact, if you compute the average relative difference of 1000 RMSE values for slightly altered datasets, you will receive a value of ~169.5. 169.5! This means that if we slightly change our dataset, our RMSE will fluctuate by around **169.5%**, which makes the predictions of this model as reliable as your office or home printer in the moment you need it most. This means that our model performs well sometimes, and catastrophically other times. Becaus
inconsistent across
high variance.

With this we can cap

- high fluctuation o

Because this model
say that it is *overfitti*
this very exact datas
relationship that is tr

Revisiting the D

Ok, so we have defin
we know that the err
The lower the error f
higher the error fluct

Table of Contents

When talking about the variance of a particular model, we always talk about one model, but *multiple* datasets. In practice, you would compare the model error on the training dataset and the error on the testing (or validation) dataset.

Let's now also look at the definition used in statistics. In the words of Wikipedia:

> *In probability theory and statistics, variance is the expectation of the squared deviation of a random variable from its mean. In other words, it measures how far a set of numbers is spread out from their average value.*

The important part is "*spread out from their average value*". This means that the variance is a way of describing the difference between the expected (or average) value and the predicted value of our model. This definition is similar to the one we described ea

Naturally, we are inte possible, because w predictions across al second (degree=4) a that the third model one, but it also has a the second model. T best model out of th minimizing both bias first model (degree= third model minimize

The Tradeoff

Let's think about hov way to achieve a low model or to train our we want to extract *more insights* from our dataset, we want our model to *learn more* from our data. The more our model learns from our data, the better it will be at solving its task.

Table of Contents

And how might we decrease variance? The easiest way to decrease variance would be to pick a more simple model or to train our existing model for a shorter amount of time. In other words, we want to extract *fewer insights* from our dataset, we want our model to *learn less* from our data. The less our model learns from our data, the more general it will be, because it has only learned so much to base its predictions on.

A good analogy would be to think of a student preparing for their mathematics and physics exams. If the student does not prepare at all, they will do poorly in both exams (high bias). If they put a lot of time into their preparation, they will probably do pretty well in both exams (low bias, low variance).
However, if the stud
preparation, and doe
over again (prolonge
train themselves not
exams, but to be goo
they have solved ove
on the exam that the
without any error. Ho
have not practiced, t
answering it (low bia

So there seems to b
variance. You can on
variance starts to inc
graph like this one:

Table of Contents

make interactive

The Tradeoff?

If you have a rather simple model like polynomial regression, there certainly is a tradeoff between bias and variance. As
you saw, increasing
decreased bias cons
variance. After that s
increase dramatically
meaningless.

However, if you have
deep neural network
Oftentimes it seems
really such "tradeoff
get into too much de
enough to fill a sepa
you to some externa
reading more about
Neal which explains
you are looking for a
the paper A Modern
Neural Networks, wh
wrote the aforementioned blog post. The paper covers the topic a bit more in-depth.

Table of Contents

An All-Purpose Remedy?

No matter what machine learning model you are training and which problem you are trying to solve, there is one thing you can do that will almost always result in a better, more stable machine learning model. What is it? Simply said, "getting more (and better) data". There have been numerous papers written and studies performed* about whether increasing data capacity (and quality) yields a greater positive effect on model performance than trying to come up with a more elaborate machine learning algorithm. While it's impossible to generalize the results to each and every application, the general notion seems to be that increasing data capacity yields greater results than choosing a more complicated model.

So what does this m
the performance of y
have already tried ou
and alternative mode
model performance i
machine learning mo
issue is frequently pr
model.

Table of Contents

Diving Deepe

In this article, I briefl
irreducible error, an
itself, and that no ma
can undergo. This is
might be amazing in
Therefor it makes se
are analyzing the bia
models have been st
any further no matter what you do, it may just be the case that you've reached the ideal bias for this particular problem and that it is simply not possible to improve any further with this particular dataset that you have.

Oftentimes you won't know the irreducible error for a problem. It is pretty difficult to compute and it's even more difficult to prove that it really is irreducible. In practice, you can't compute the irreducible error for every problem you try to solve with machine learning. What you can do instead, is to train multiple different machine learning models*, calculate the bias for each of those models and compare them to each other. If you have multiple models that are all optimized for that particular problem and they all possess very similar biases, you might not get any better results by optimizing your models.

Further Reading

Bias and variance ar
describe overfitting
what exactly over-ar
actually prevent it in
can stop overfitting
read the post How T
where you will learn
make your model mc
as underfitting.

You can only improv
until it becomes very
any further. Howeve
results when you op
models are trained o
can improve your da
you can do so, you s
Optimization, where
exactly how you can do so.

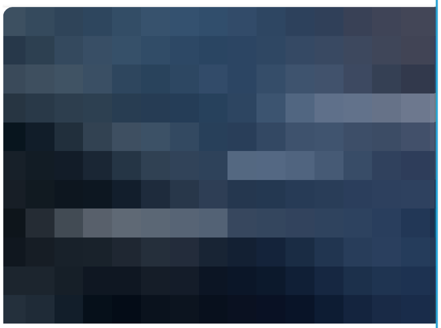
Table of Contents

0 comments Log In

Comments powered by Talkyard.

These posts may int

Table of Contents



MACHINE LEARNING MATH

Gradient Descent f Linear Regression Explained, Step by Step

Gradient descent is one of most famous techniques in machine learning and used for...

BORIS GIBA
03 APR 2021 • 23 MIN READ

BORIS GIBA
01 MAY 2021 • 15 MIN READ

RSS

Imprint & Privacy Policy (EN)

Reset Cookie Preferences

Impressum & Datenschutz (DE)

Table of Contents