Credit: Pixabay

The curse of bias and variance explained intuitively

Published on April 10, 2018

Manish Barnwal | + Follow
Data Scientist at Zeta Global
3 articles

Do you enjoy beers? How about data science? Does the title of this post excite you? If the answer to any of these questions is yes, you should go ahead and read this.

Statistics is the field of study where we try to draw conclusions about the population from a sample. Why do we talk about sample? Why can't we get the conclusions about the population directly from the population? Let me illustrate this by an example.

Let us say we want to understand which brand of beer do the people of Bangalore prefer? An interesting question. If I ask you this question, how would you approach this problem?

You can't go around asking each and every person their favorite beer. Or can we? No, we can't cover each and every individual because the 'population' is huge. One thing you can do is you may ask among your circle of friends their preference of beer and get an overall idea of the population. But we have a problem with this analysis. Do you see the problem? Your estimation is suffering from bias or we say your sample is biased. A biased sample is when it is not random. There is some form of personal preference in the choice of picking data.

In your case, chances are most of your friends will be of same age as you. Intuitively, I feel that your age also decides what kind of beer you like. Say, when I was in college I'd never heard of Corona and Kingfisher was the best beer I had tasted. So we can't estimate the best brand of beer that Bangalore prefers from the sample of your friends.

So population is a broader set of data that covers all the data points in the entire universe. A sample is a subset of that data. We try to infer the characteristics of the population from the sample or try to answer questions about the population from the sample. Getting or collecting population data is tough as explained in the above example (The favorite beer example).

I hope the concept of population and sample is clear now. It's normally not feasible to get the data for the complete population so we try to estimate parameters or findings of the population from a sample.

Another example. Not the beer but a rather completely different one - marriages.

Say, now we want to understand what are the factors that affect the age at which one gets married. Some of the factors I can think of without much thought are:

1. Gender
2. Love or arranged marriage
3. Plan for higher studies
4. Company type - Government of Private
5. Salary
6. Region
7. Religion

Now we don't know the true relationship between marital age and the variables listed above. There should be a true function that maps response variable to the predictors but we don't know what that function is.

Let's say that true function is f

f: (gender, love-marriage, higher-studies, ...)
--> marital age

We don't know the true **f()** but we can estimate the true function using the data we may have from the past. It's easy to collect the data for the respective variables and the marital age. We are trying to come up with a function say *fcap* that resembles closely to the true function f.

Whenever we try to estimate true f from the data in hand, we will obviously get some error. This error can be categorized into two types:

1. **Reducible error**

As the name suggests, this is something that the analyst has some control over. This can be reduced based on the kind of data you collect and the models (not all models are the same, right?) one uses to estimate the true f. This error can arise from a combination of 'bias' and 'variance'. We will talk about bias and variance in next few paragraphs.

2. **Irreducible error**

As the name suggests, this error is something that the analyst has no control over. There is always some information that is difficult to capture in data. There is always some randomness in the data and that is difficult to explain. This error can't be reduced using any model whatsoever.

Let us dig deeper into reducible error. We will talk about bias and variance here.

Bias

When we talk about estimating the true f, there are various models that can be used. Now, not all models are the same. Each has some characteristics of its own. A linear regres

the simplest model is different from say a random forest model. Bias is the error that captures how far is the predicted value (say predicted marital age given the variables like gender, love or arranged marriage, etc.) from the true value (actual marital age).

Now, you may ask is there a relationship between the bias and the type of models used? Yes, there is one. Bias tends to decrease as the complexity of model increases i.e it is expected that the model error will decrease if we use a more complex model instead of a simple model. Now, this is intuitive, isn't it?

You may ask what is meant by the complexity of a model or an example of it? Linear regression is a very simple model whereas random forest is a more complex model. Linear regression tries to fit just a line to the actual data and it assumes a linear relationship. A random forest model is more complex is the sense that it uses an ensemble of decision trees and is able to explain non-linear relationship as well.

Variance

When you estimate the true f, you use some data to train the model, that data is called training data. The data that the machine uses to train on, to find the patterns. Now once your model is trained, you want to use the model to predict on unseen data (the data that the model has not been trained on) that data is called test data.

There is always some variability in the training data and the test data. You can't expect both these data to be exactly same. It is important to note that when you train your model, the model doesn't learn the exact values but instead try to find patterns so that when this pattern is seen on test data, the model is able to predict correctly.

Many complex models overfit the data i.e models that perform well on training data but fails drastically on test data. An example of an overfitted model could be this - the accuracy on training data is high (say 90%) but the same model's accuracy on test data is extremely low! (say 60%).

How is variance related to the complexity of models? As the complexity increases, the chances of overfitting increases i.e the variance increases. Coming to the random forest and linear regression example, a random forest's variance is expected to be higher than that from a linear regression model.

So if we talk about complexity, bias, and variance together, this is the relationship between them.

As the complexity of model increases, the bias decreases but the variance increases.

So there is a trade-off between bias and variance. You can't get both low bias and low variance at the same time. You will have to accept a trade-off. Do you now understand why we call it the curse of bias and variance? I hope you do.

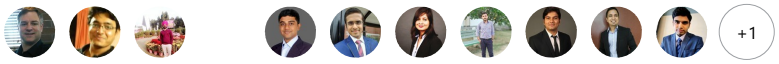
Did you find the article useful? If you did, share your thoughts in the comments. Share this post with people whom you think would enjoy reading this. Let's talk more about data-science!

This article was originally posted on Manish Barnwal's blog. If you enjoyed reading this and would like to read more on data-scien

and machine learning, please come by.

Report this

12 Likes



2 Comments

Shubham Sirothia 1d
Data Scientist (Clinical Care) at Option Care

Interesting example of marriage age prediction, I wonder what would be my age :P

Like Reply | 3 Likes · 1 Reply

Manish Barnwal 16h
Data Scientist at Zeta Global

You are just a few steps away from the answer - start with collecting the data. :D

Like Reply

Add a comment...

Manish Barnwal
Data Scientist at Zeta Global

+ Follow

More from Manish Barnwal

Random Forest explained intuitively
Manish Barnwal on LinkedIn

The Big Data Problem
Manish Barnwal on LinkedIn

ked in

unity Guidelines
y & Terms
feedback

lIn Corporation © 2018

Questions?
Visit our Help Center.

Manage your account and privacy.
Go to your Settings.

Select Language
English (English)

Messaging