SCIENCE ACADEMY
Excellence in Teaching & Learning

Linear Regression

Hi guys, welcome to the very first method in machine learning section.

In this lecture, we are going to talk about Linear Regression, which is a very simple approach and also considered the “work horse” for supervised machine learning. Linear regression has been around for a long time and is the topic of countless textbooks. It is very useful and widely used statistical or machine learning method. Moreover, it serves as a good jumping-off point for newer approaches.

✅ ***Optional Readings and References:***

Ch # 7 on Linear Regression in Machine Learning - A Probabilistic Perspective by Kevin Murphy

Ch # 3 on Linear Regression in An Introduction to Statistical Learning by Gareth et.al.

Original work by Sir Galton at http://www.galton.org

and off-course, http://scikit-learn.org

Dr. Junaid S. Qazi, PhD

General message: *Key concepts along with significant commentary / text is provided in the slides, so that they serve as a reference for the respective theory lecture. However, the suggested readings are recommended to explore more on the topic under discussion!*

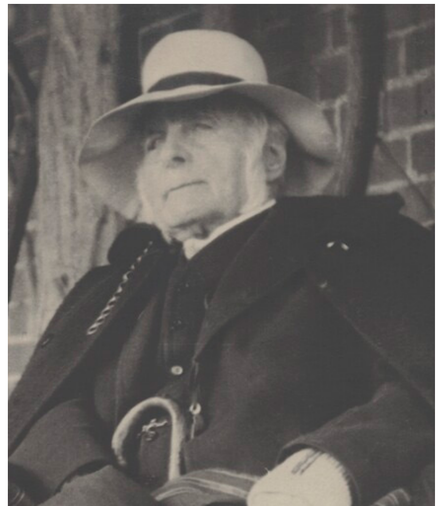
Good luck!

SCIENCE ACADEMY
Excellence in Teaching & Learning

Linear Regression

History:

The earliest form of regression was the method of least squares, which was published by Legendre in 1805, and later on by Gauss in 1809. However, the term "**regression**" was coined by **Sir Francis Galton** in his work, published in 1875 while he was describing the biological phenomenon for relating the heights of descendants to their tall ancestors. For Sir Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context.



✅ *Please note, all the information & images in the presentation are used under fair use policy. Direct links to the source are also provided for reference.*

SCIENCE ACADEMY
Excellence in Teaching & Learning

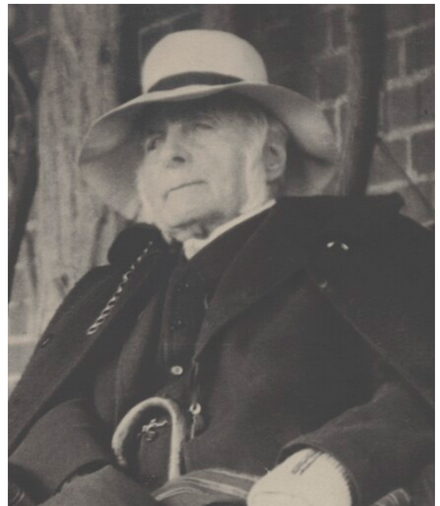
Linear Regression

History:

The earliest form of regression was the method of least squares, which was published by Legendre in 1805, and later on by Gauss in 1809. However, the term "**regression**" was coined by **Sir Francis Galton** in his work, published in 1875 while he was describing the biological phenomenon for relating the heights of descendants to their tall ancestors. For Sir Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context.

In his study, **Sir Galton**, discovered that:

- A man's son tends to be roughly as tall as his father but the **son's height tends to be closer** (regress or drift towards) **to the overall average heights**.



We will look at Sir Galton's data for Father's & their Son's heights in the coming slides!

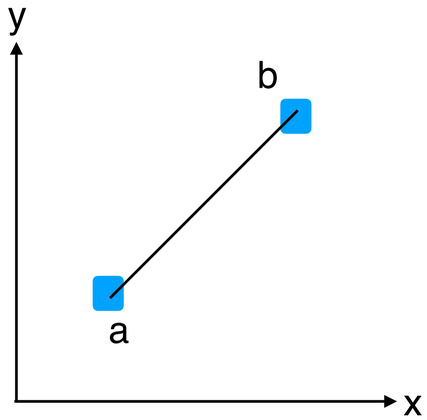
✅ *Please note, all the information & images in the presentation are used under fair use policy. Direct links to the source are also provided for reference.*

SCIENCE ACADEMY
Excellence in Teaching & Learning

Linear Regression

Let’s consider the simplest possible example:

We have 2 data points a & b, as shown in the diagram, if we draw a line, which is closer to each point as much as possible, the line will be exactly through the points. Simply, to fit a line using Least Square Method (classical linear regression and a standard approach) we only measure the closeness in the “up & down” directions.

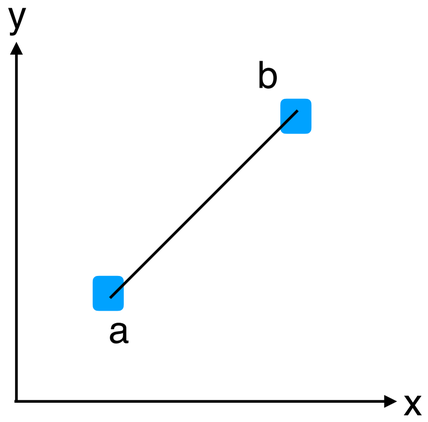


SCIENCE ACADEMY
Excellence in Teaching & Learning

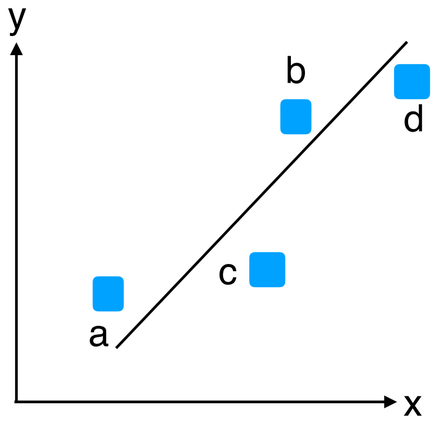
Linear Regression

Let’s consider the simplest possible example:

We have 2 data points a & b, as shown in the diagram, if we draw a line, which is closer to each point as much as possible, the line will be exactly through the points. Simply, to fit a line using Least Square Method (classical linear regression and a standard approach) we only measure the closeness in the “up & down” directions.



Let’s add 2 more data point, c & d:
Now if we draw a line following the same rule, Least Square Method, the line will be different for the given points. It may not pass through the data points!

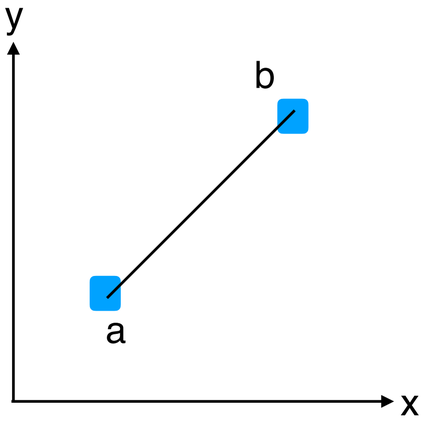


SCIENCE ACADEMY
Excellence in Teaching & Learning

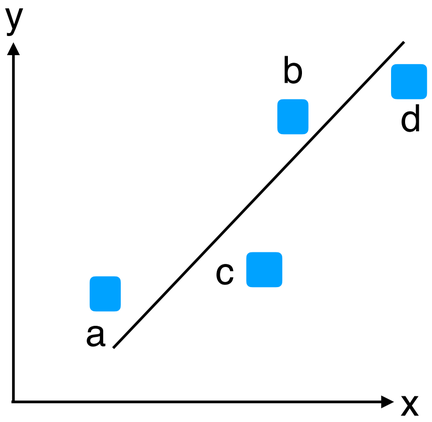
Linear Regression

Let's consider the simplest possible example:

We have 2 data points a & b, as shown in the diagram, if we draw a line, which is closer to each point as much as possible, the line will be exactly through the points. Simply, to fit a line using Least Square Method (classical linear regression and a standard approach) we only measure the closeness in the "up & down" directions.



Let's add 2 more data point, c & d:
Now if we draw a line following the same rule, Least Square Method, the line will be different for the given points. It may not pass through the data points!



Now the questions are:

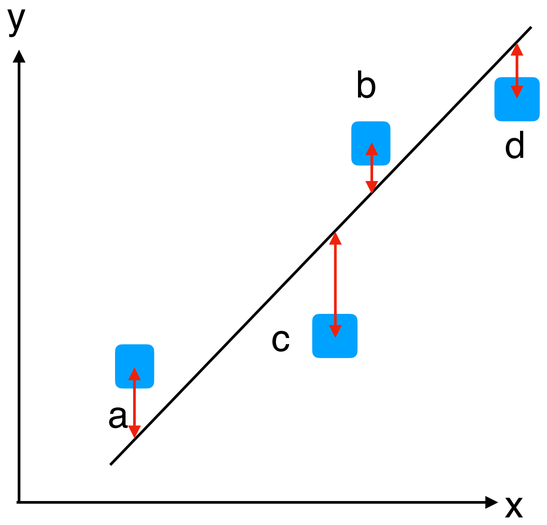
- **Is the black line best fitted line for the given data points?**
- **If yes, why?**

SCIENCE ACADEMY
Excellence in Teaching & Learning

Linear Regression

Let's explore little more to learn about the best fitted line:

Following the standard approach, the best fit in the least-squares sense minimizes the sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a model).

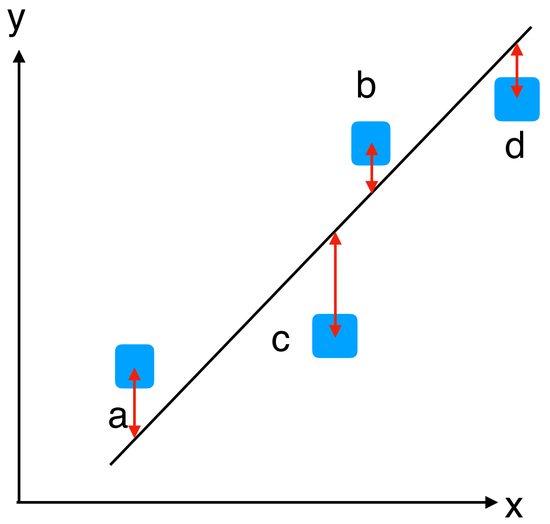


SCIENCE ACADEMY
Excellence in Teaching & Learning

Linear Regression

Let's explore little more to learn about the best fitted line:

Following the standard approach, the best fit in the least-squares sense minimizes the sum of squared residuals (a residual being: the difference between an observed value, and the fitted value provided by a model).



In very simple words, to get the best fitted line in linear regression, we attempt to minimize the vertical distance between all the data points and their distance to the fitted line.

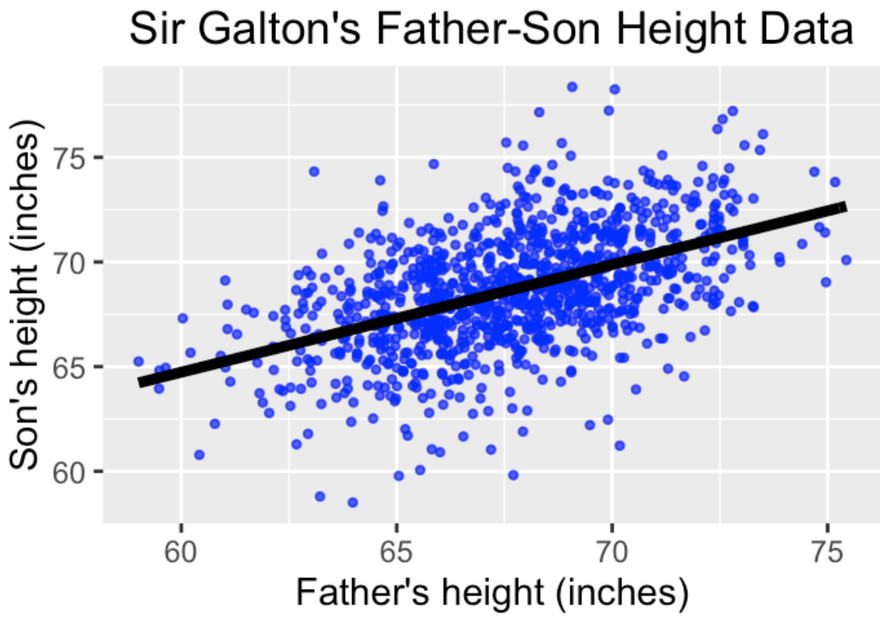
Good to know (optional):
Linear regression models are often fitted using the least squares approach, but they may also be fitted in other ways, such as by minimizing the "lack of fit" in some other norm (as with least absolute deviations regression), or by minimizing a penalized version of the least squares loss function as in ridge regression (L^2-norm penalty) and lasso (L^1-norm penalty). Conversely, the least squares approach can be used to fit models that are not linear models. Thus, although the terms "least squares" and "linear model" are closely linked, they are not synonymous.

SCIENCE ACADEMY
Excellence in Teaching & Learning

Linear Regression

Let's look at **Sir Galton's Father-Son heights data**:

- Each blue data points are the height of a Father along "x" in inches and height of his Son along "y".
- Black line is fitted regression line using least-square model.
- With this data from multiple men and their son's heights, if we know the heigh of a man, we can predict his son's height even the son is not born yet!



✅ *Sir Galton's data is available to use for free for learning, figure presented here is generated in R using Sir Galton's data for Father's & their Son's heights.*

*** more the data points we have, better the model we get!**

SCIENCE ACADEMY
Excellence in Teaching & Learning

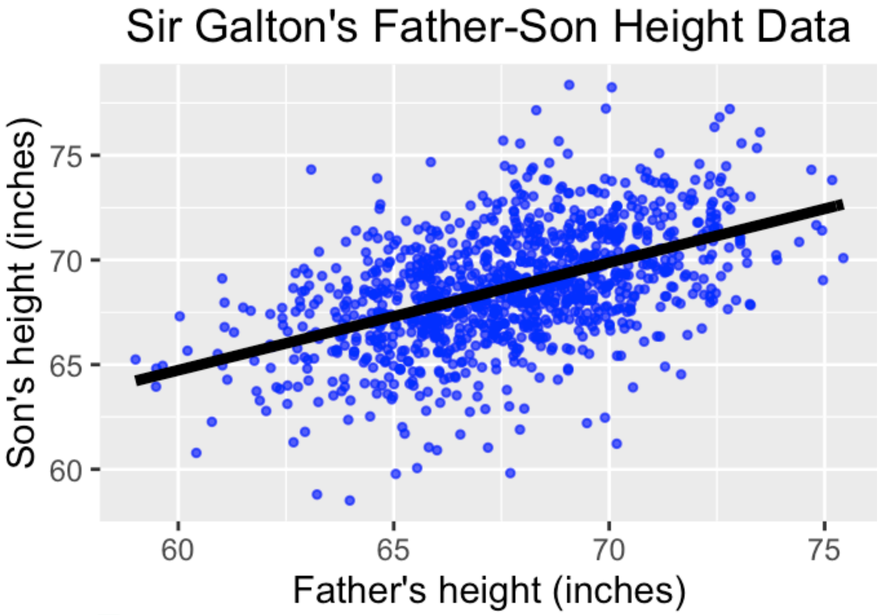
Linear Regression

Let's look at **Sir Galton's Father-Son heights data**:

- Each blue data points are the height of a Father along "x" in inches and height of his Son along "y".
- Black line is fitted regression line using least-square model.
- With this data from multiple men and their son's heights, if we know the heigh of a man, we can predict his son's height even the son is not born yet!

This is the idea behind supervised learning. We have a bunch of labeled data points* to create a model.
Once we have a model, we can deploy that on unlabeled data for future predictions!

*** more the data points we have, better the model we get!**



✅ *Sir Galton's data is available to use for free for learning, figure presented here is generated in R using Sir Galton's data for Father's & their Son's heights.*

SCIENCE ACADEMY
Excellence in Teaching & Learning

Before we move on to the notebook for hands-on training, lets discuss few very important and fundamental concepts in Machine Learning

SCIENCE ACADEMY
Excellence in Teaching & Learning

No Free Lunch!

Much of machine learning is concerned with devising different models, and different algorithms to fit them. For a particular dataset, one specific method may work best, but some other method may work better on a similar but different dataset. *There is no universally best model* — this is sometimes called the **no free lunch theorem** (Wolpert 1996). The reason for this is that a set of assumptions that works well in one domain may work poorly in another.

✅ *Optional Readings and References:*
Machine Learning - A Probabilistic Perspective - section 1.4 & 6.4
Introduction to Statistical Learning - section 2.2

SCIENCE ACADEMY
Excellence in Teaching & Learning

No Free Lunch!

Much of machine learning is concerned with devising different models, and different algorithms to fit them. For a particular dataset, one specific method may work best, but some other method may work better on a similar but different dataset. *There is no universally best model* — this is sometimes called the **no free lunch theorem** (Wolpert 1996). The reason for this is that a set of assumptions that works well in one domain may work poorly in another.

As a consequence of the no free lunch theorem, we need to develop many different types of models, to cover the wide variety of data that occurs in the real world.

✅ *Optional Readings and References:*
Machine Learning - A Probabilistic Perspective - section 1.4 & 6.4
Introduction to Statistical Learning - section 2.2

SCIENCE ACADEMY
Excellence in Teaching & Learning

No Free Lunch!

Much of machine learning is concerned with devising different models, and different algorithms to fit them. For a particular dataset, one specific method may work best, but some other method may work better on a similar but different dataset. *There is no universally best model* — this is sometimes called the **no free lunch theorem** (Wolpert 1996). The reason for this is that a set of assumptions that works well in one domain may work poorly in another.

As a consequence of the no free lunch theorem, we need to develop many different types of models, to cover the wide variety of data that occurs in the real world.

For each model, there may be many different algorithms we can use to train the model, which make different speed-accuracy-complexity tradeoffs.

✅ *Optional Readings and References:*
Machine Learning - A Probabilistic Perspective - section 1.4 & 6.4
Introduction to Statistical Learning - section 2.2

SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias and Variance

As we move on in Machine Learning section, we will explore further to widen our concepts on types of models, their selection and evaluation etc. Let's discuss one of the fundamental concept in the model performance at this stage which is "bias-variance tradeoff". This has great importance while working with real world data!

✅ *Optional Readings and References:*
Machine Learning - A Probabilistic Perspective - section 1.4 & 6.4
Introduction to Statistical Learning - section 2.2

SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias and Variance

As we move on in Machine Learning section, we will explore further to widen our concepts on types of models, their selection and evaluation etc. Let's discuss one of the fundamental concept in the model performance at this stage which is "bias-variance tradeoff". This has great importance while working with real world data!

- **Bias** is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

✅ *Optional Readings and References:*
Machine Learning - A Probabilistic Perspective - section 1.4 & 6.4
Introduction to Statistical Learning - section 2.2

SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias and Variance

As we move on in Machine Learning section, we will explore further to widen our concepts on types of models, their selection and evaluation etc. Let's discuss one of the fundamental concept in the model performance at this stage which is "bias-variance tradeoff". This has great importance while working with real world data!

- **Bias** is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).
- **Variance** is an error from sensitivity to small fluctuations in the training set. High variance can cause an algorithm to model the random noise in the training data, rather than the intended outputs (overfitting).

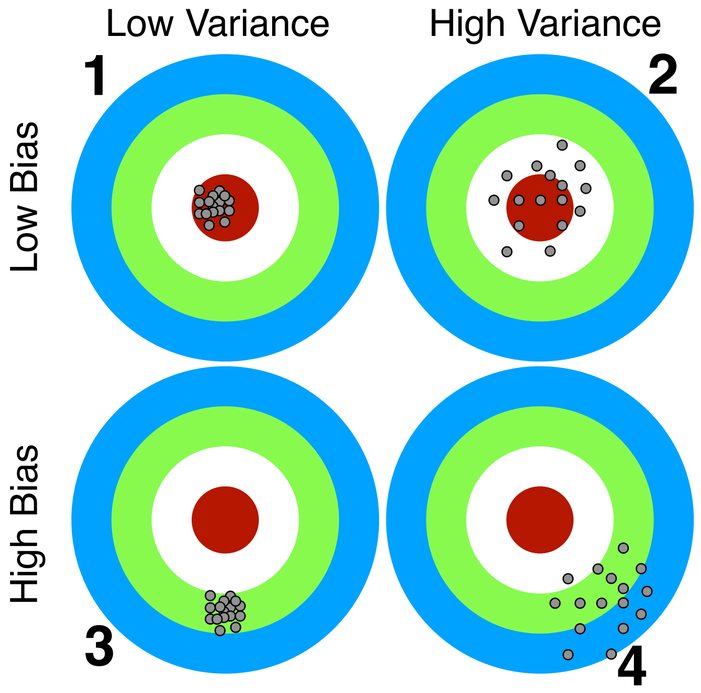
To get better understanding, let's create a graphical visualization of bias and variance using a bulls-eye diagram.

✅ *Optional Readings and References:*
Machine Learning - A Probabilistic Perspective - section 1.4 & 6.4
Introduction to Statistical Learning - section 2.2

SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias and Variance

Graphical visualization could be helpful, let's try to understand the concept bias and variance using a bulls-eye diagram.

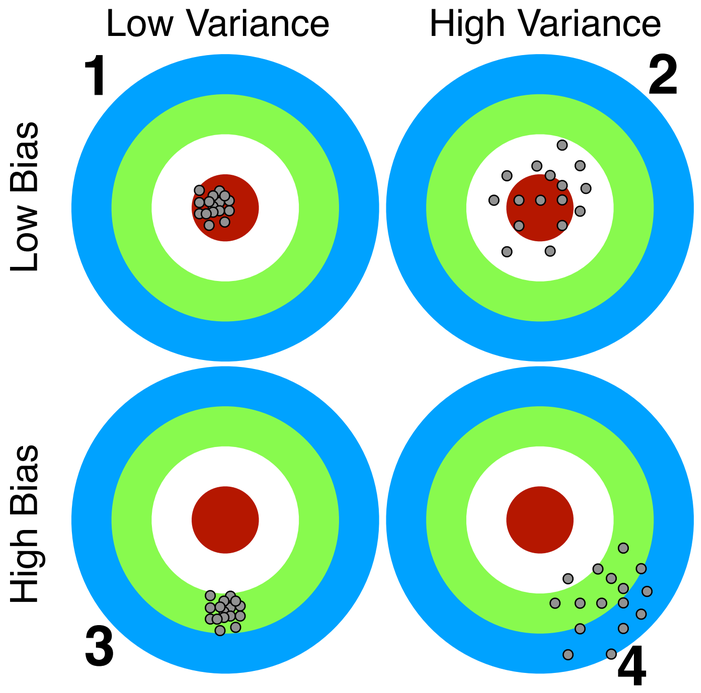


SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias and Variance

Graphical visualization could be helpful, let's try to understand the concept bias and variance using a bulls-eye diagram.

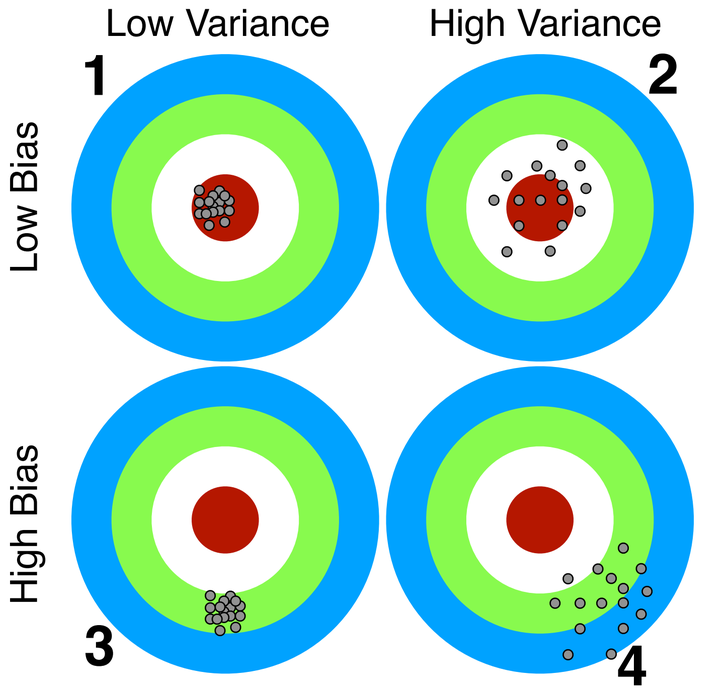
- Consider, the centre of the target is a model that perfectly predicts the correct values.
- Moving away from the bulls-eye, the predictions get worse and worse.
- We can repeat our entire model building process to get a number of separate hits on the target.
- Each hit represents an individual realization of our model, given the chance variability in the training data we gather.



SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias and Variance

- If we have good distribution in our training data, the model predicts very well and close to the bulls-eye
- If our training data is full of outliers or non-standard values, this results in poorer predictions.
- These different realizations result in a scatter of hits on the target.



Here comes Bias-variance tradeoff

SCIENCE ACADEMY
Excellence in Teaching & Learning

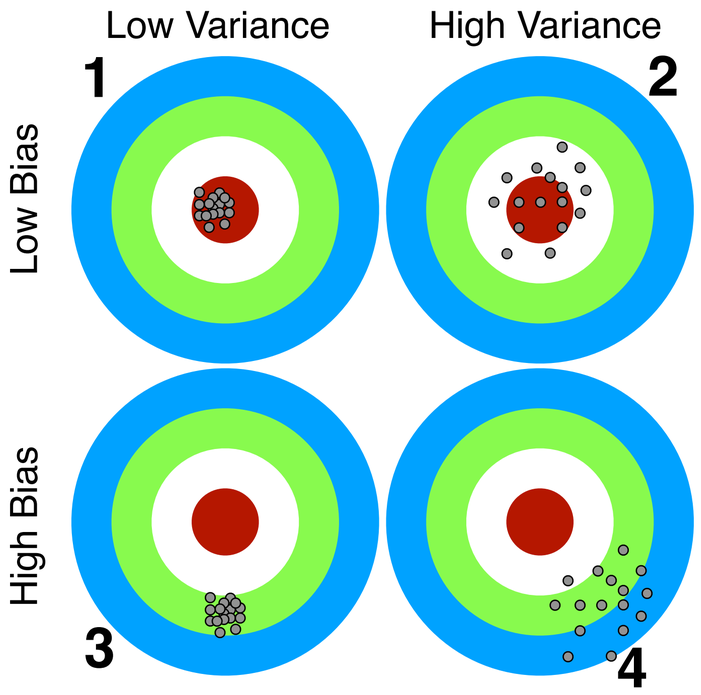
Bias Variance Tradeoff

1. Low Bias - Low Variance:
- Predicts correct values on the bulls-eye

2. Low Bias - High Variance:
- Predicts values around the bulls-eye with high degree of variance

3. High Bias - Low Variance:
- Predictions would be high bias at a certain location with low variance.

4. High Bias - High Variance:
- Predictions are all over the places



SCIENCE ACADEMY
Excellence in Teaching & Learning

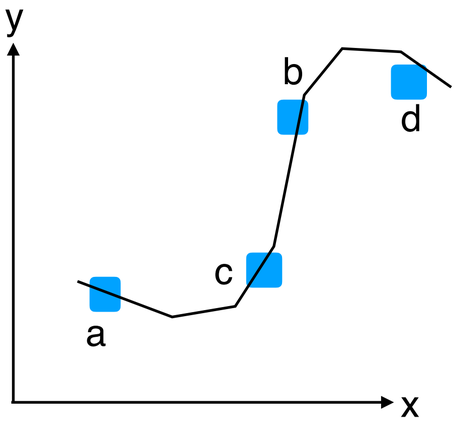
Bias Variance Tradeoff

In the beginning, to fit the training set very well, its very common for the learners to add more and more complexity in the model so that the line / fit pass through almost all the data points. This will lead to the wrong predictions of the test data which is unseen by the model - overfitting of the training data.

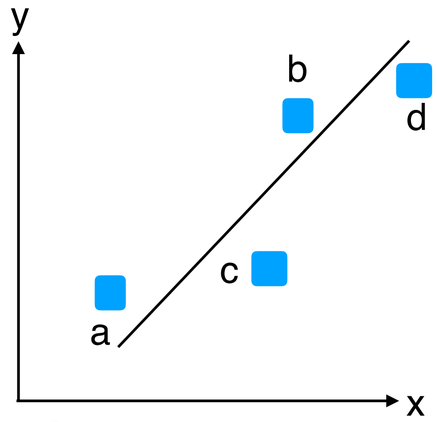
SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias Variance Tradeoff

In the beginning, to fit the training set very well, its very common for the learners to add more and more complexity in the model so that the line / fit pass through almost all the data points. This will lead to the wrong predictions of the test data which is unseen by the model - overfitting of the training data.



Complex and wrong



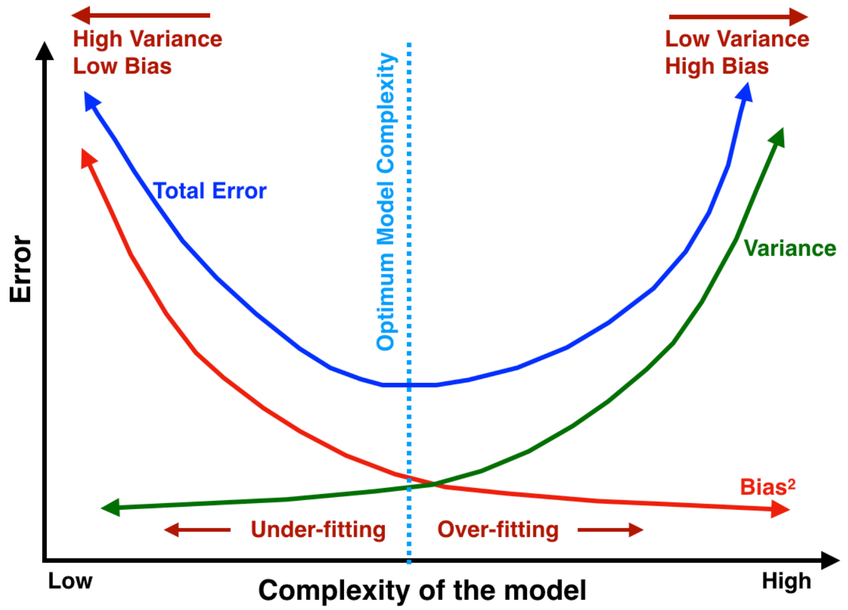
Simple and good

For example, in the model in left plot, the blue data point might be fitted very well as the line passes through each points but if we give a new dataset to this model, its may fail to predict. Whereas, the plot on the right is simple and can predict the unseen dataset easily.

SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias Variance Tradeoff

Understanding bias and variance is critical for understanding the behaviour of prediction models, but in general what you really care about is overall error, not the specific decomposition. The **sweet spot** for any model is the level of complexity at which the increase in bias is equivalent to the reduction in variance.



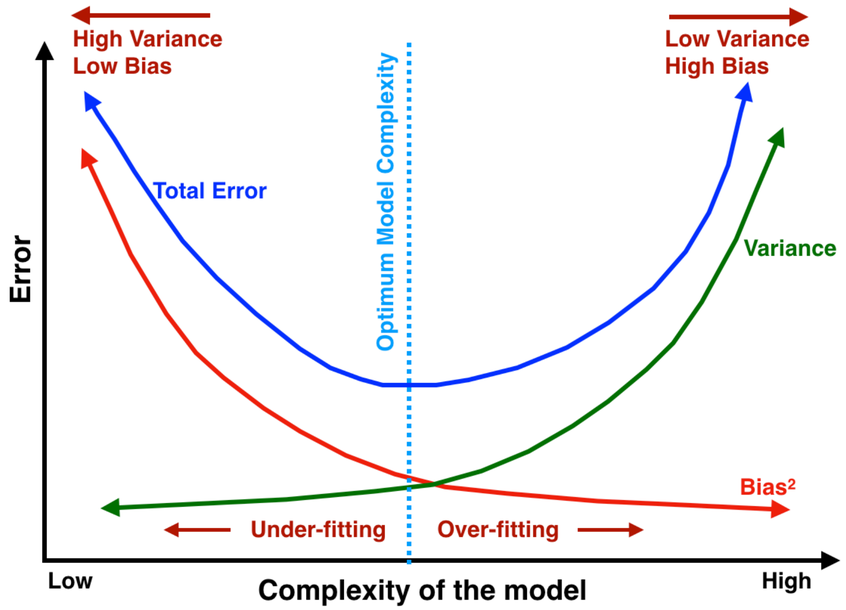
Bias and Variance contributing to the total error

SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias Variance Tradeoff

Understanding bias and variance is critical for understanding the behaviour of prediction models, but in general what you really care about is overall error, not the specific decomposition. The **sweet spot** for any model is the level of complexity at which the increase in bias is equivalent to the reduction in variance.

At its root, dealing with bias and variance is really about dealing with over- and under-fitting. Bias is reduced and variance is increased in relation to model complexity. As more and more parameters are added to a model, the complexity of the model rises and variance becomes our primary concern while bias steadily falls.



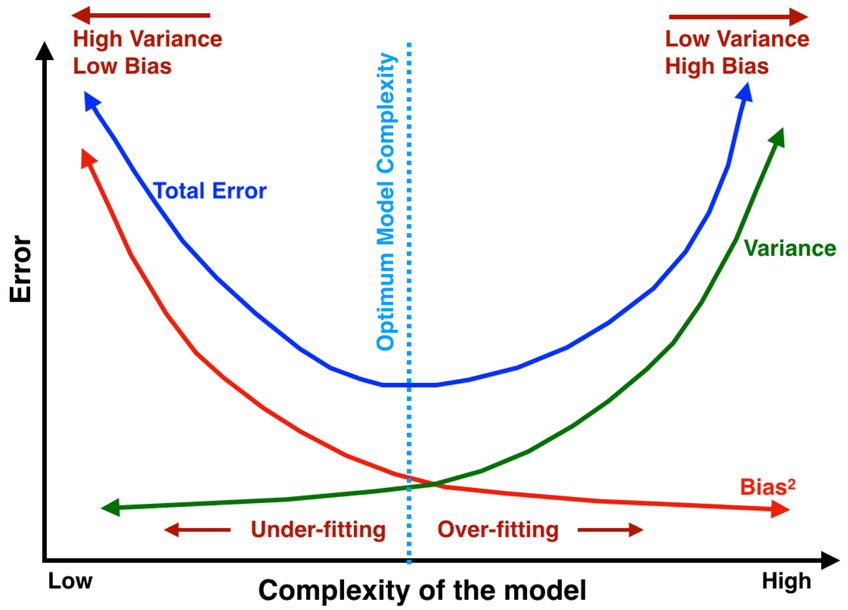
Bias and Variance contributing to the total error

SCIENCE ACADEMY
Excellence in Teaching & Learning

Bias Variance Tradeoff

Understanding bias and variance is critical for understanding the behaviour of prediction models, but in general what you really care about is overall error, not the specific decomposition. The **sweet spot** for any model is the level of complexity at which the increase in bias is equivalent to the reduction in variance.

At its root, dealing with bias and variance is really about dealing with over- and under-fitting. Bias is reduced and variance is increased in relation to model complexity. As more and more parameters are added to a model, the complexity of the model rises and variance becomes our primary concern while bias steadily falls.



Bias and Variance contributing to the total error

If our model complexity exceeds this sweet spot, we are in effect over-fitting our model; while if our complexity falls short of the sweet spot, we are under-fitting the model. ***In practice, there is not an analytical way to find this location. Instead we must use an accurate measure of prediction error and explore differing levels of model complexity and then choose the complexity level that minimizes the overall error.*** A key to this process is the selection of an *accurate* error measure as often grossly inaccurate measures are used which can be deceptive.

SCIENCE ACADEMY
Excellence in Teaching & Learning

Excellent!

We have covered the basics on linear regression and some very important concepts. Time to learn this widely used technique with more practical example using real data.

In the next lecture, we will create our very first linear regression model using Python's Machine Learning library Scikit-Learn. I can't wait to do this!

Let's move on to work on our first machine learning project!

✅ *Optional Readings and References:*

Ch # 7 on Linear Regression in Machine Learning - A Probabilistic Perspective by Kevin Murphy

Ch # 3 on Linear Regression in An Introduction to Statistical Learning by Gareth et.al.

Original work by Sir Galton at http://www.galton.org

and off-course, http://scikit-learn.org