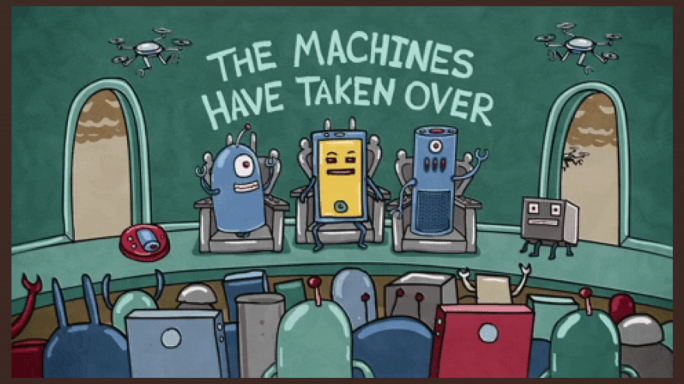
The Bias Variance Trade Off

By Mansi Goel

Contents

- Machine Learning
- Data
- Bias – Variance Trade Off
- Derivation of equation

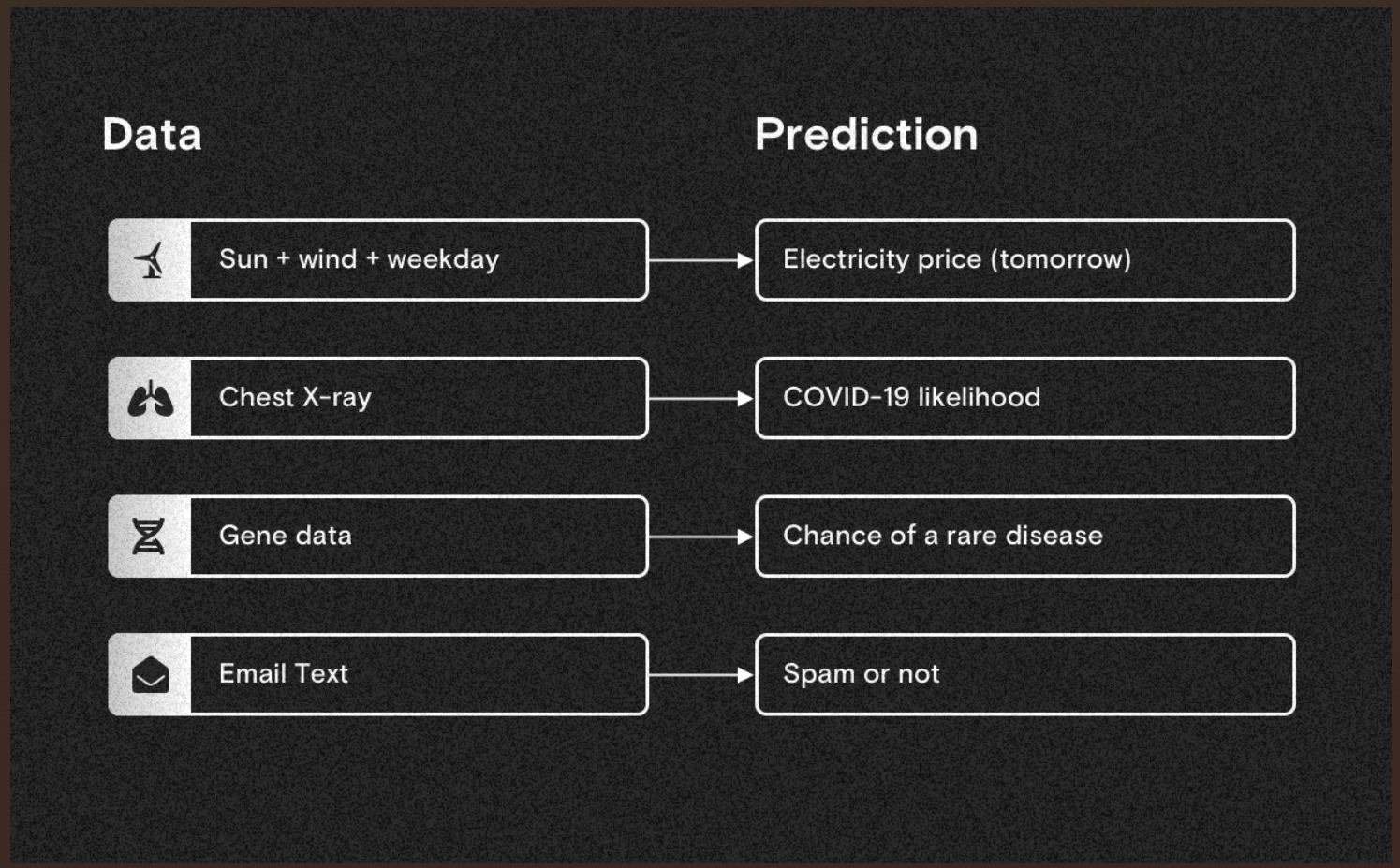
What is Machine Learning?



- Machine Learning is the art of providing systems
  - The ability to automatically learn from past data
  - Apply that learning without human intervention.

  This learning is done through various algorithms which require a huge amount of data to obtain accurate results .

- Machine learning is everywhere. Possibility is that you are using it in one way or the other and you don’t even know about it.

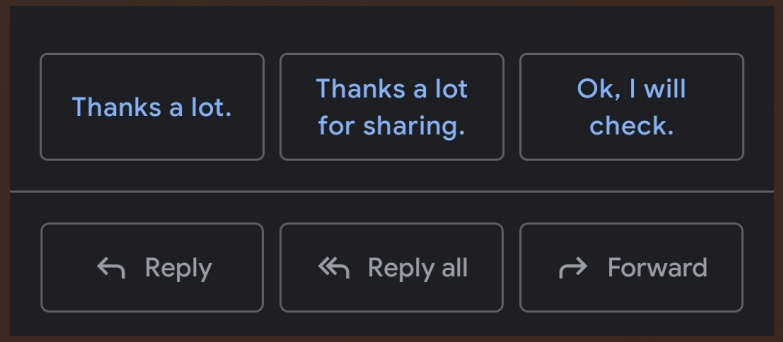


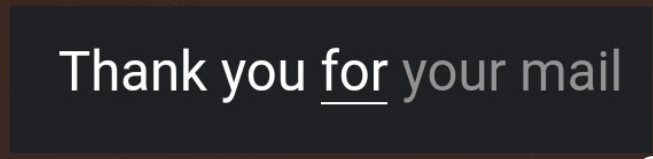
- Same Riverdale Movie, but two different artistic image thumbnails, based on user's past preference for romance (sweet smiles) or thriller (serious, dramatic looks) movie genres.

A few more examples…

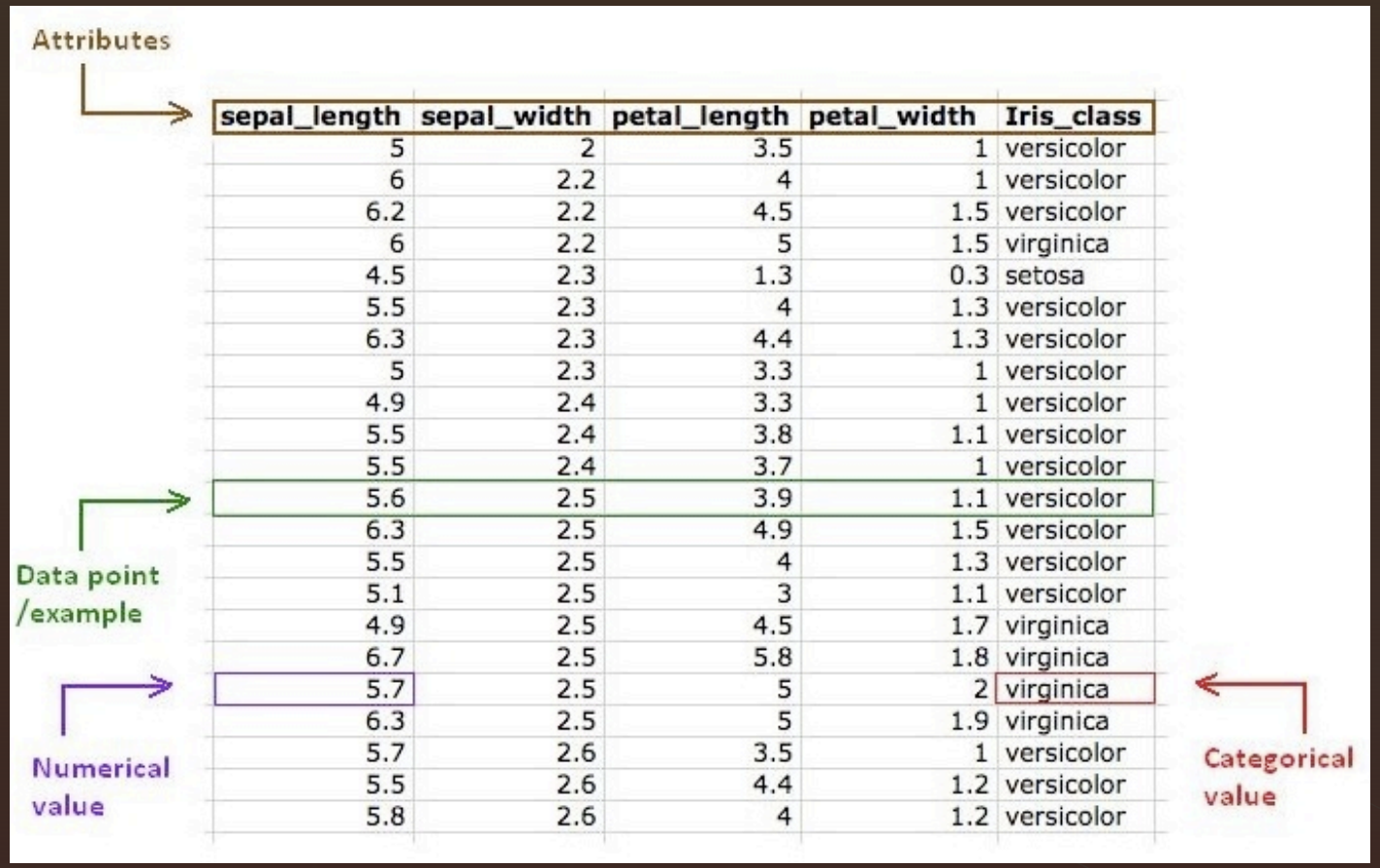
- **Gmail Inbox's Smart Reply –** It uses machine learning to automatically generate replies to emails, saving mobile users the hassle of tapping out answers on those tiny keyboards.
- **Amazon -** Uses machine learning to provide customised recommendations to its customers. According to a report, Amazon's recommendation engine is driving 35% of its total sales.

... and the list goes on

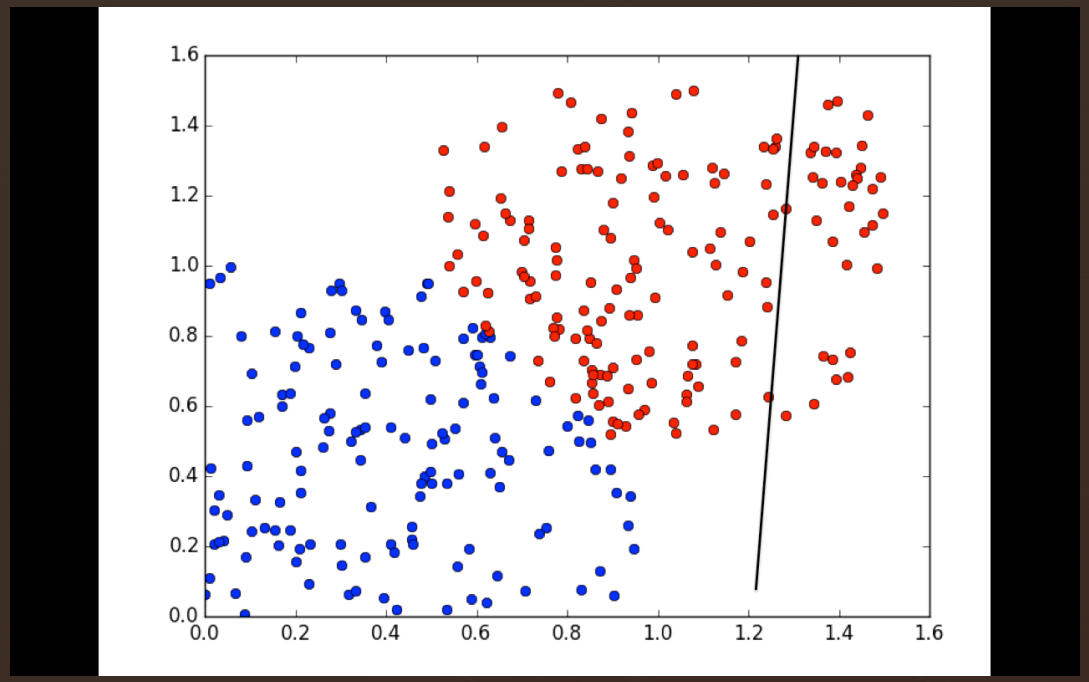


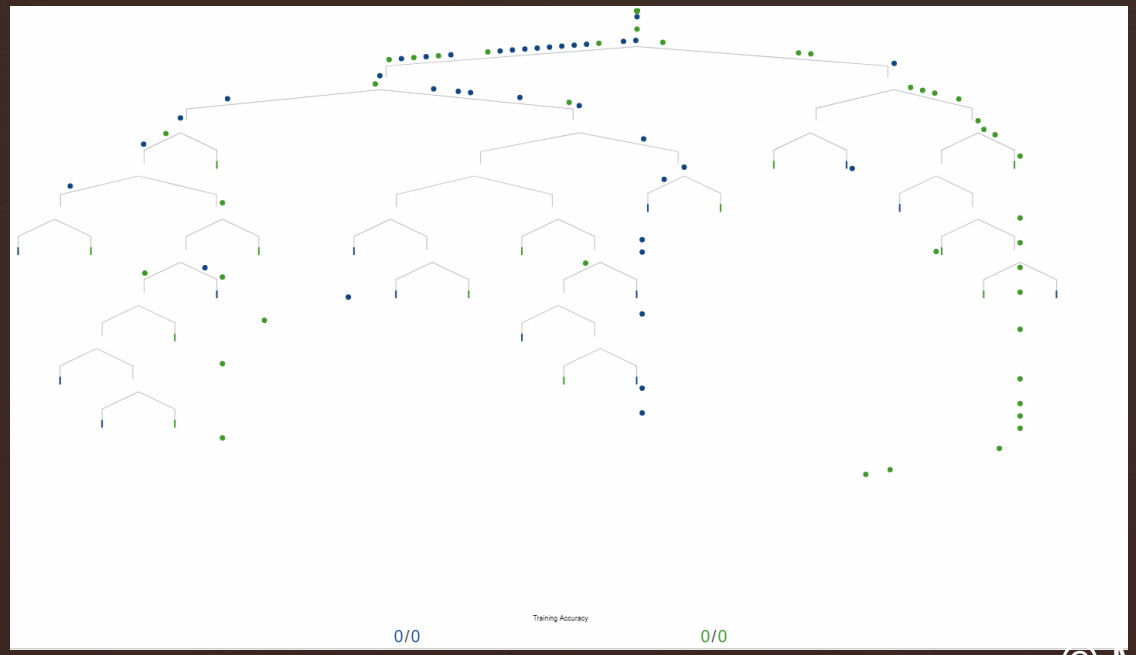


- The data we obtain is divided into two parts – **training data** and **testing data**. The training data is a key input to any algorithm which then comprehends from such data and memorizes the information for future prediction.
- After a model has been processed using the training set, you test the model by making predictions against the test set. Because the data in the testing set already contains known values for the attribute that you want to predict, it is easy to determine whether the model's guesses are correct.
- Typically, when you separate a data set into a training set and testing set, most of the data is used for training, and a smaller portion of the data is used for testing.
- So, a machine learning algorithm accomplishes its task when the model has been adjusted with respect to the data. We say that we have **"fit the model on the data"** or that **"the model has been trained on the data."**



sepal_length	sepal_width	petal_length	petal_width	Iris_class
5	2	3.5	1	versicolor
6	2.2	4	1	versicolor
6.2	2.2	4.5	1.5	versicolor
6	2.2	5	1.5	virginica
4.5	2.3	1.3	0.3	setosa
5.5	2.3	4	1.3	versicolor
6.3	2.3	4.4	1.3	versicolor
5	2.3	3.3	1	versicolor
4.9	2.4	3.3	1	versicolor
5.5	2.4	3.8	1.1	versicolor
5.5	2.4	3.7	1	versicolor
5.6	2.5	3.9	1.1	versicolor
6.3	2.5	4.9	1.5	versicolor
5.5	2.5	4	1.3	versicolor
5.1	2.5	3	1.1	versicolor
4.9	2.5	4.5	1.7	virginica
6.7	2.5	5.8	1.8	virginica
5.7	2.5	5	2	virginica
6.3	2.5	5	1.9	virginica
5.7	2.6	3.5	1	versicolor
5.5	2.6	4.4	1.2	versicolor
5.8	2.6	4	1.2	versicolor


1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
0.0
0.2
0.4
0.6
0.8
1.0
1.2
1.4
1.6


Training Accuracy
0/0
0/0

The Bias Variance Trade Off

Notations & Assumptions

- x : vector of independent variables/predictor variables
- y : dependent variable/target variable
- f(x) : represents the true underlying relationship between *x* and *y*. f remains fixed.
- ε : random variable representing noise with $\mathbb{E}[\varepsilon] = 0$, Var(ε)= σ^2=$\mathbb{E}[\varepsilon^2]$
- $y = f(x) + \varepsilon$: relationship between x and y
- $\hat{f}(x)$: function that can accurately predict the true relationship *f*. Our aim is to bring predictions as close as possible to their observed values: $y \approx \hat{f}(x)$

Errors

- While achieving the goal of approximating f by $\hat{f}$, we encounter some errors which have to be minimized for more accurate approximations.
- These errors are of two types:
  1. Reducible Errors
  2. Irreducible Errors
- Bias and variance are reducible errors.
- The irreducible error refers to the error that can not be reduced, often known as noise. So, even the best models will always have an error which can not be removed. Hence there is a trade-off between bias and variance to decrease the reducible error which in turn would minimise the total error.
- The total error of a model can we expressed as - $Bias^2 + Variance + \sigma^2$ where σ^2 is the irreducible error.

Bias

- It is the difference of the average prediction(*over different realizations of training data*) to the true function *f(x),* for a given unseen (test) point *x*

$$Bias[\hat{f}(x)] = \mathbb{E}\left[\hat{f}(x)\right] - f(x)$$

- Or, in simpler words, it is the inability of a model to accurately capture the true relationship.
- Models with **high bias** are oversimplified and fail to capture the complexity of the data. Such models lead to a higher training and testing error.
- **Low bias** corresponds to a good fit to the training dataset. Generally, more flexible models result in lower bias.

Variance

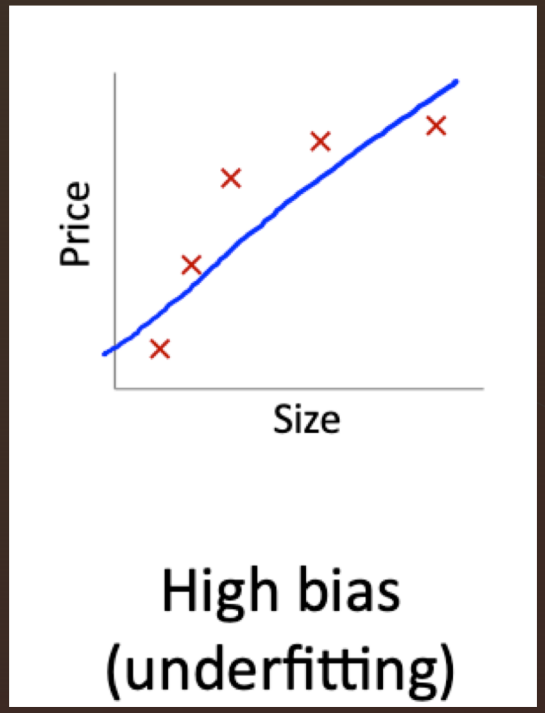
- *Variance* is defined as the mean squared deviation of $\hat{f}(x)$ from its expected value $\mathbb{E}[\hat{f}(x)]$ over different realizations of training data.
- Using different training sets will result in different estimations but ideally the estimate should not vary much over different training sets

$$Var[\hat{f}(x)] = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$$

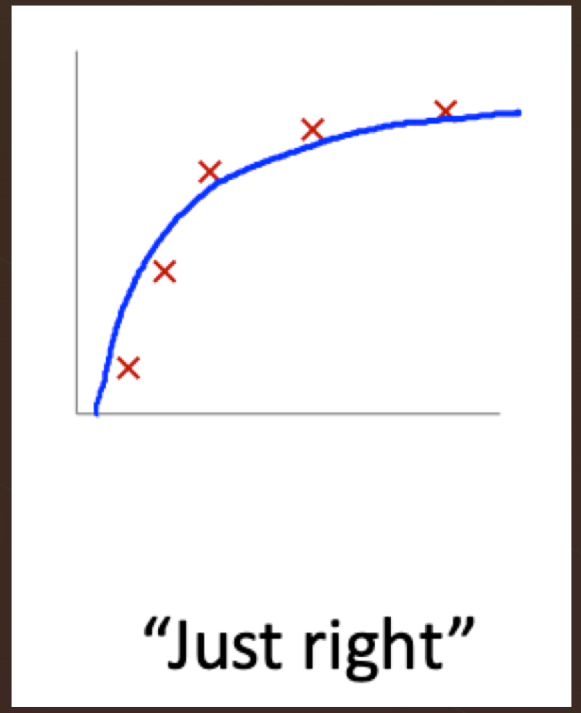
- **High variance** implies that the model does not perform well on previously unseen data (testing data) even if it fits the training data well. Generally, more flexible models have higher variance.

Overfitting & Underfitting

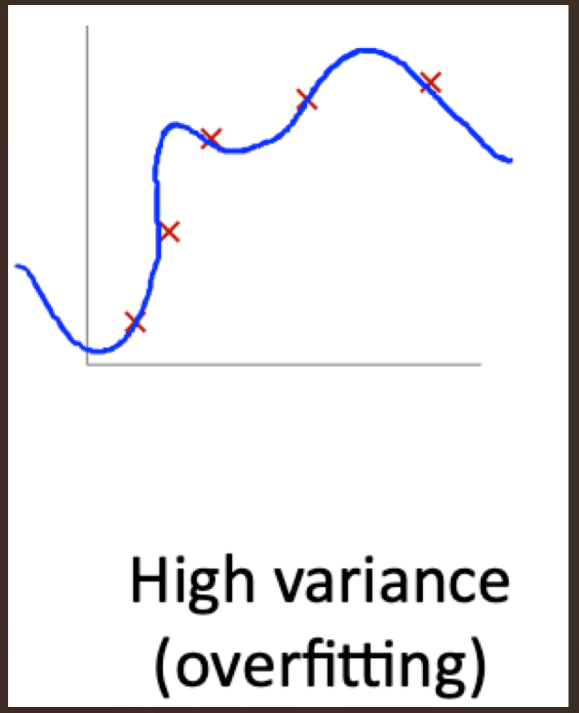
- When our model suffers from high variance, it is unable to generalize well beyond the training data and we call this *overfitting.*
- When our model suffers from high bias, the average response of the model is far from the true value and we call this *underfitting.*



High bias
(underfitting)



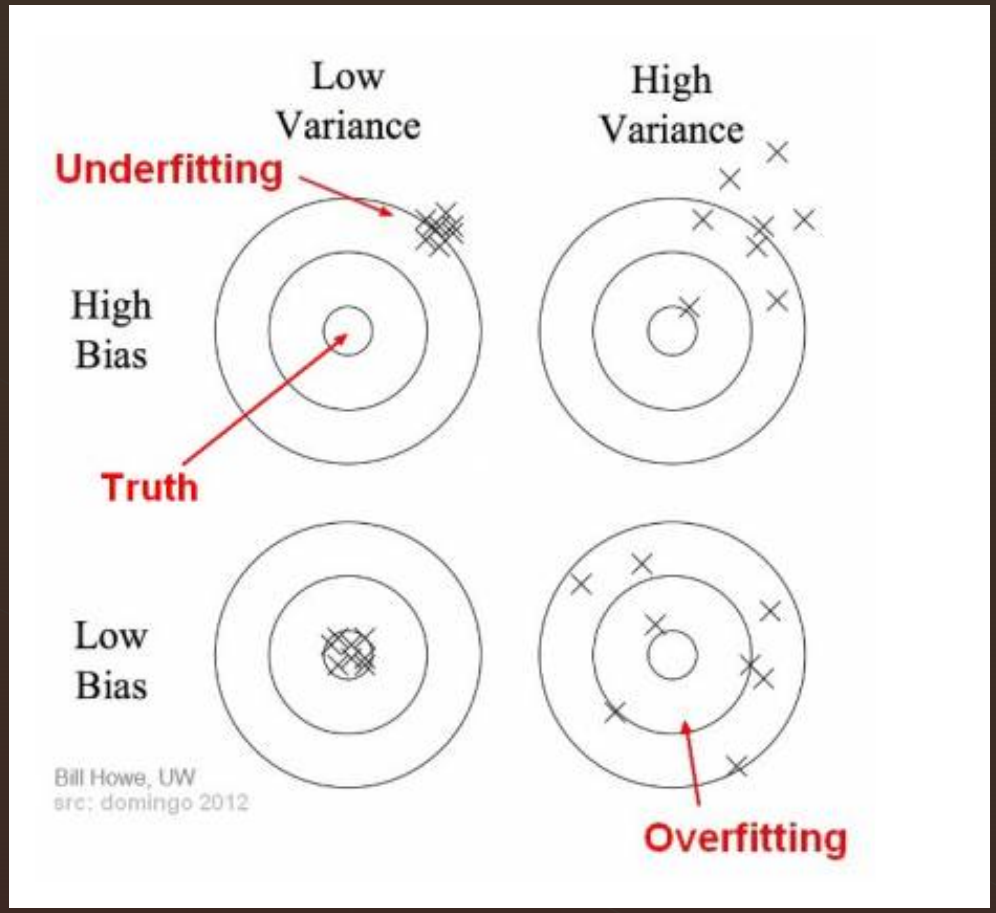
"Just right"



High variance
(overfitting)

The Trade-Off

- If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model is too complex then it's going to have high variance and low bias.
- Ideally, a model should have **low bias** so that it can accurately model the true relationship and **low variance** so that it can produce consistent results and perform well on testing data.
- So we need to find a good balance without overfitting and underfitting the data so that the combined error of these two competing forces is minimum.



In the above diagram, centre of the target is a model that perfectly predicts correct values. As we move away from the bulls-eye our predictions become get worse and worse. We can repeat our process of model building to get separate hits on the target.

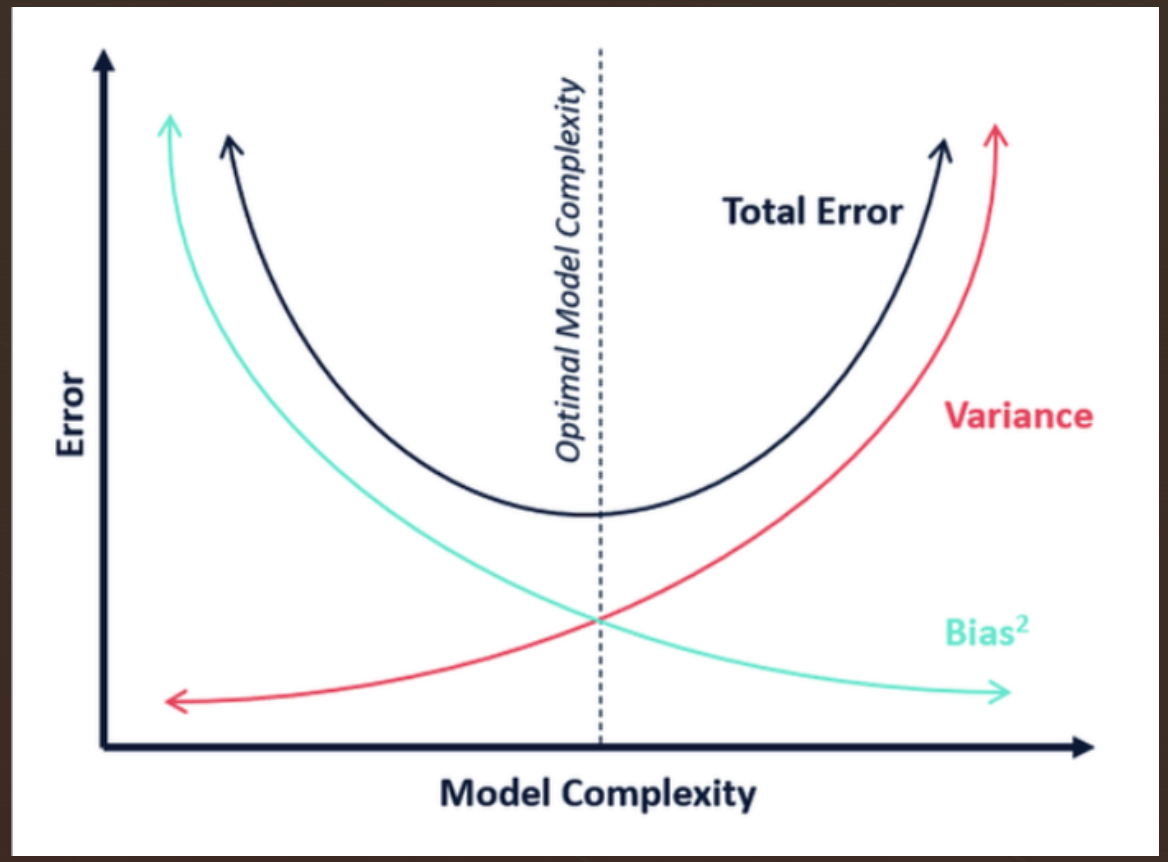
Mathematical View

- The total error of the model can be represented by the following equation:
- $Total\ Error = Bias^2 + \text{Variance} + \text{Irreducible Error}$ where:
- $Total\ Error = MSE = \mathbb{E}[(y - \hat{f}(x))^2]$
- $Bias[\hat{f}(x)] = \mathbb{E}[\hat{f}(x)] - \hat{f}(x)$
- $Var[\hat{f}(x)] = \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x)])^2]$
- $Irreducible\ Error = \sigma^2$

Derivation

- $\mathbb{E}\left[\left(y-\hat{f}(x)\right)^2\right]=(\mathbb{E}[\hat{f}(x)]-\hat{f}(x))^2+\mathbb{E}\left[\left(\hat{f}(x)-\mathbb{E}[\hat{f}(x)]\right)^2\right]+\sigma^2$
- Starting from LHS;
- $\mathbb{E}\left[\left(y-\hat{f}(x)\right)^2\right]=\mathbb{E}\left[\left(f(x)+\varepsilon-\hat{f}(x)\right)^2\right]$
- $=\mathbb{E}\left[\left(f(x)-\hat{f}(x)\right)^2\right]+\mathbb{E}[\varepsilon^2]+2\mathbb{E}\left[\left(f(x)-\hat{f}(x)\right)\varepsilon\right]$ $[\because \mathbb{E}[X+Y]=\mathbb{E}[X]+\mathbb{E}[Y], \mathbb{E}[aX]=a\mathbb{E}[X]]$
- $=\mathbb{E}\left[\left(f(x)-\hat{f}(x)\right)^2\right]+\sigma^2+2\mathbb{E}\left[\left(f(x)-\hat{f}(x)\right)\right]\mathbb{E}[\varepsilon]$ $[\because \mathbb{E}[\varepsilon^2]=\sigma^2, \mathbb{E}[\varepsilon]=0]$
- $=\mathbb{E}\left[\left(f(x)-\hat{f}(x)\right)^2\right]+\sigma^2$

- $= \mathbb{E}\left[\left(\left(f(x) - \mathbb{E}[\hat{f}(x)]\right) - \left(\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right)\right)^2\right] + \sigma^2$
- $= \mathbb{E}[\left(\mathbb{E}[\hat{f}(x)] - f(x)\right)^2] + \mathbb{E}\left[\left(\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right)^2\right]$
  $-2\mathbb{E}[\left(f(x) - \mathbb{E}[\hat{f}(x)]\right)\left(\hat{f}(x) - \mathbb{E}[\hat{f}(x)]\right)] + \sigma^2$
- $= (\mathbb{E}[\hat{f}(x)] - f(x))^2 + Var[\hat{f}(x)] - 2\left(f(x) - \mathbb{E}[\hat{f}(x)]\right)\left(\mathbb{E}[\hat{f}(x)] - \mathbb{E}[\hat{f}(x)]\right) + \sigma^2$
- $= Bias[\hat{f}(x)]^2 + Var[\hat{f}(x)] + \sigma^2$



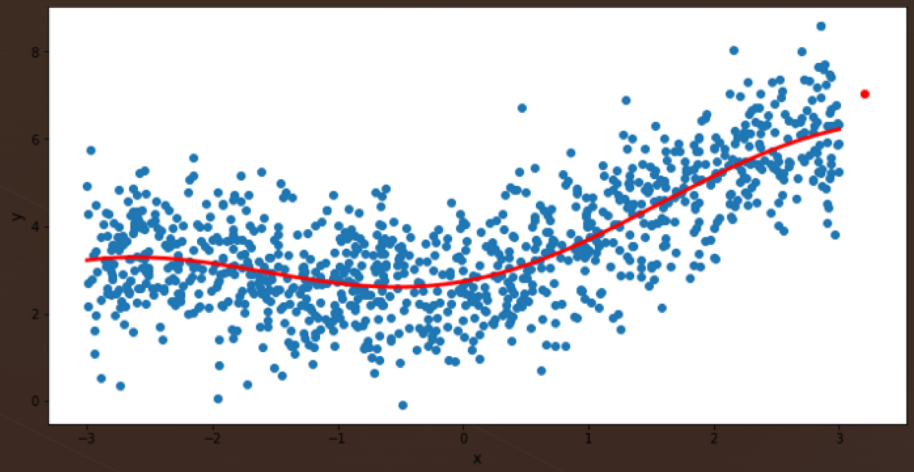
$$Total\ Error = Bias^2 + \text{Variance} + \text{Irreducible Error}$$

The total error can never go below the irreducible error and hence it is necessary to minimize $Bias^2 + \text{Variance}$ to achieve accuracy.

Bias – Variance Trade Off in Practice

- Assume, the underlying true function *f* that dictates the relationship between *x* and *y* is:
- $f(x) = \frac{1}{2}x + \sqrt{\max(x, 0} - cosx + 2$
- The noise ε is modeled by a standard normal distribution i.e $\varepsilon \sim N(0,1)$
- As a reminder, $y = f(x) + \varepsilon$

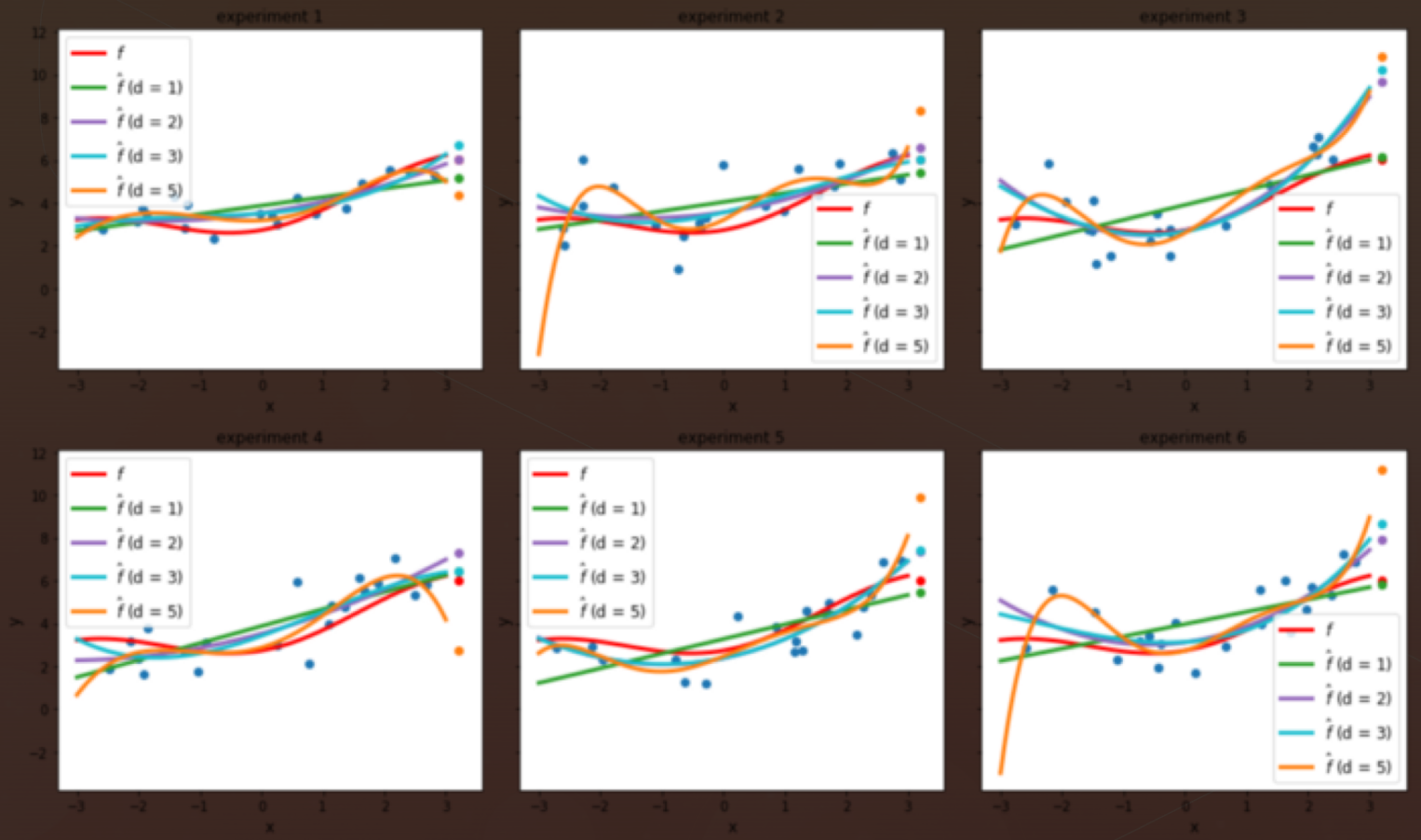
- We will randomly generate 1,000 points from this process and get the following plot:



- Blue dots represent (*x*, *y*) pairs and red line is the underlying true function *f(x).*
- Red dot is the unseen (test) point we want to predict.
- We see that *f* follows a non-linear pattern due to the addition of square root and cosine in the function's definition.
- For our purposes, these 1,000 points represent the whole underlying population.

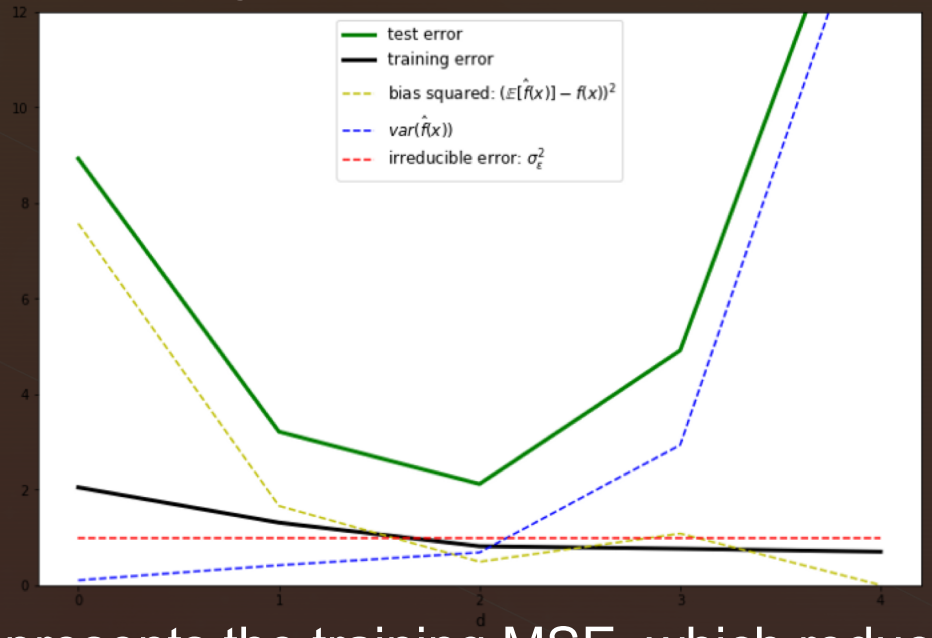
- The model we will use here is a ‘polynomial regression’ with varying degrees of complexity.
- In this model, we try to approximate *y* with $\hat{f}(x)$ as described by the equation :$\hat{f}(x) = w_0 + w_1x + w_2x^2 + \ldots + w_nx^d$
- Now, let’s assume we could only use 20 points (out of the 1,000) to train our polynomial regression model.
- We consider four different regression models, degree *d*=1, *d*=2, *d*=3 and *d*=5.

- If we randomly sample 20 points from the underlying population and we repeat this experiment 6 times, this is a possible outcome we get:



- Blue dots represent the 20 training data ,the red line is the underlying true function *f* and the other lines represent the fitting of the four different models.
- The green, purple, blue, orange dots represent the prediction $\hat{f}(x)$ of test point *x* under each model.
- There's less variation in lines with smaller degrees of complexity, say for d=1. The slope of the line does not change all that much between different experiments.
- On other other hand, a more complex model (*d*=5) is much more sensitive to small fluctuations in the training data. This is the *variance* problem we mentioned in previous sections.
- A simplistic model is very robust to changes in training data, but a more complex is not. On other hand, the deviation of $\hat{f}(x)$ from $f(x)$ on average (the *bias*), is larger for more simplistic models, since our assumptions are not as representative of the underlying true relationship *f*.

- Now let's consider some test points and compute the test MSE, squared bias and variance. If we do this for the earlier 4 models, we get the following plot:



- Black line represents the training MSE, which reduces with model complexity, as more complex models tend to fit training data better.
- In this particular example, we see that the best model for the underlying problem is a quadratic one (d=2), since it achieves minimum test error.

Conclusion

- Defined two reducible error terms : Bias & Variance
- Stated and proved the equation representing the composition of the total error
- Demonstrated that model choice has to battle two competing forces: bias and variance
- A good model should strike a balance between the two but can never achieve zero test error due to the presence of irreducible error
- Our model should not be overly simplistic, but not too complex either so that it can generalize well to previously unseen data.

Thank You! 😀