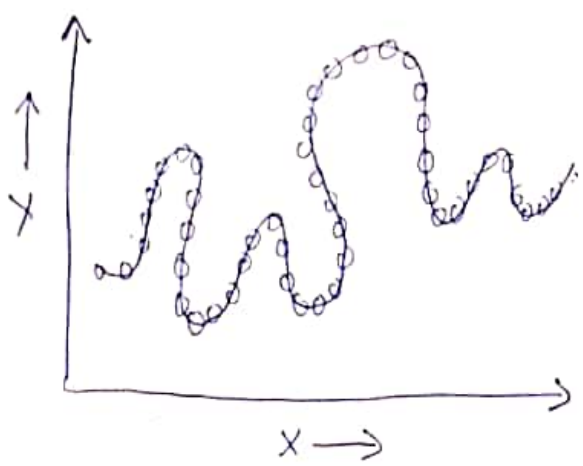
→ "Overfitting, Underfitting, Bias & Variance"



Overfitting

Line or the curve we are seeing here is basically the output of the training phase. It is basically the model that we have created. So, if our model covers almost all the points in our training dataset perfectly, then we say that the model is ['overfit'.]

'Underfitting'

if our line / model doesn't fit the points. Then it is known as an ['Underfitting condition'.]

Real life example: Let's say we want to construct a model that helps us to identify whether an object is a ball or not.

Let's first consider the Underfitting condition:

Let's say we have only one feature, which we have used to train our model. & the feature that we have selected is let's say the shape of the object. So, we are saying that, if the shape of the object is spherical, we want our model to say that it is a ball.

"Now let's say our model is ready to make predictions".

Now instead of a real ball, let us pass an orange or a test object to our trained model. Now since only shape has been used to train the model, & the shape of an orange is also spherical. So our model will say that it is a ball but infact it is an orange.

This is known as an underfitting condition because here we have used only one feature to train the model.

Note :- [Less knowledge means "Underfitting"]

Let's now understand the overfitting case :-

In this case we will overwhelm our model with lots of information (with lots of features)

Let's stick to the same example of identifying a ball

In the previous case we have used just one feature to train it

but now in this case we will use a lots of features to train our model with precise information. So here we give loads of knowledge to our model, due to the specificity of the knowledge or features our model will get confused & will give wrong answer.

Let the features be.

{ Sphere, Play, Eat & Radius = 5cm }

Here also our objective is to find out whether a given object is a ball or not.

Now let's see what does our model predict :>

Let's give a ball of radius = 10 cm as a test object to this model.

Sphere ✓
Play ✓
Eat ✓
Radius = 5cm ✗

So, our model will check each & every feature, we used to train it.

Since the test object is spherical, we have passed the first test

We can play with it, also it has passed the second test also.

Since we can eat it, it has passed the third test also.

But since the radius of our test object, ≠ 5cm our model will say that it is not a ball, but infact it is a ball.

So, this is an example of Overfitting condition.

→ Overfitting :> The situation where any any given model is performing too well on the training data but the performance drops significantly over the test data is called an "overfitting model".

→ Underfitting :> The situation, where the model is performing poorly over the test & the training set, then we call that an underfitting model.

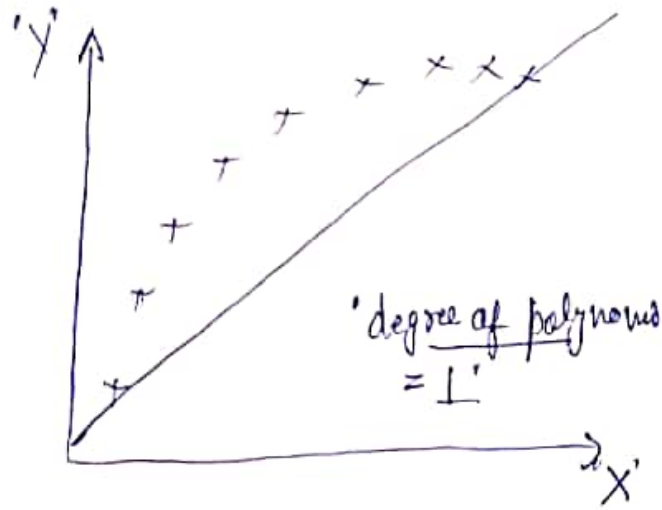
Let's now understand Overfitting & Underfitting for both Regression & Classification tasks.

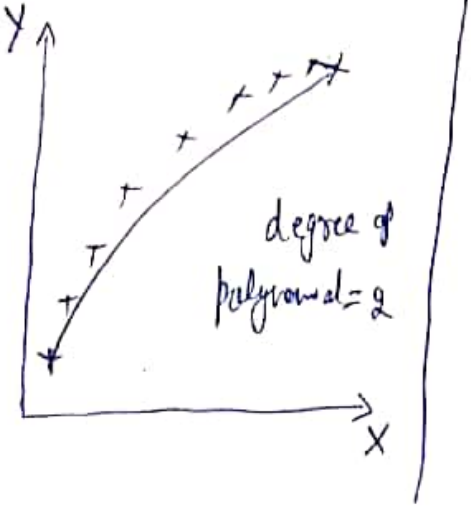
Let's first understand 'Overfitting' & 'Underfitting' using Regression.

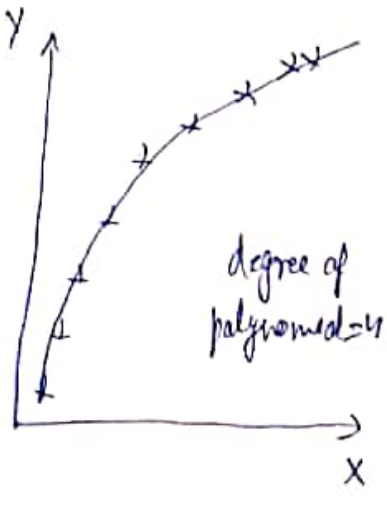
Let's say we have a problem statement wrt 'X' & 'Y' & we have some points or observations in the 2d space & our aim is to create a best fit line with the help of a linear regression.

& there are various different kinds of linear regression, like multiple linear regression, polynomial linear regression.

Here I will be using a polynomial linear regression







When deg. of polynomial = 1, then the polynomial linear regression will be acting like a simple linear regression.

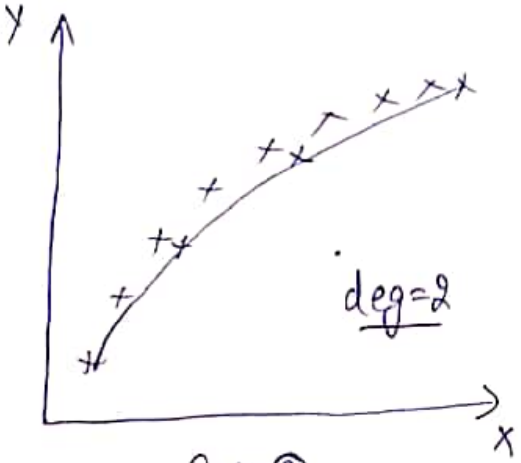
And linear regression, we know will create a best fit line on the training data. & it is not suitable for data points which are not linearly scattered.

Case①

Now when we compute the mean squared error for this scenario it is definately going to be a large value.

Error will be on the higher side.

Now suppose we increase the degree of polynomial to 2... Now the best fit line will look like a curve. which is little bend

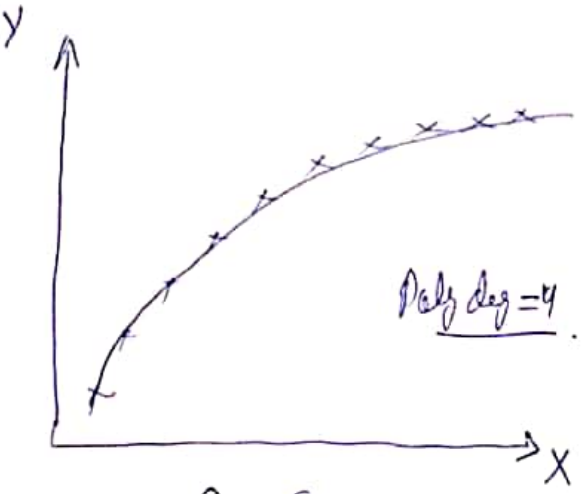


Case②

In this scenario we can see that it is satisfying a sufficient no of points & when we find the squared error, it is going to be less as compared to the previous one

Now let's go one step ahead. Let's suppose we increase the degree of polynomial to 4. Now Here we can see that this is the condition where every point is exactly fitted by the curve.

P.T.O



Case ③

In Case ① , for the training dataset our model is giving a very high error.

So this scenario is called ['Underfitting'.]

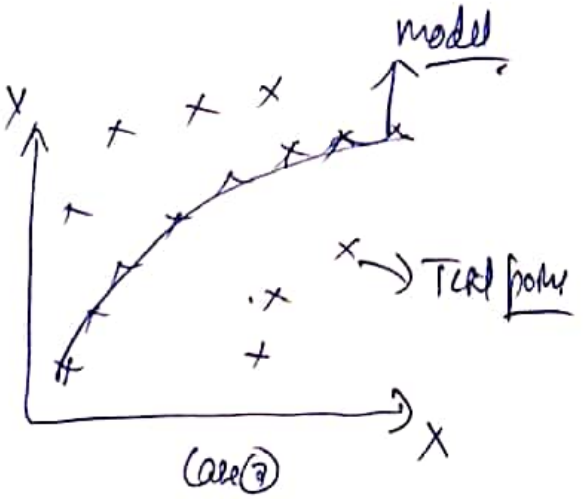
In Case ② for the training dataset our model is giving a very very less error,

So, this is the scenario which we can call as 'Overfitting'

Overfitting means, with respect to the training data our line fit perfectly all the points.

But if we have some new points. from the test set as shown we can see that the best fit line will not satisfy these test points

∴ Test accuracy is going to very small



In Overfitting, Even if the accuracy for training data is very high, but for the 'test-data' it is going to be very less.

In overfitting :-
- For Training data :- Accuracy ↑
- For Testing data :- Accuracy ↓

In Underfitting :-
- For Training data :- Accuracy ↓
- For Testing data :- Accuracy ↓

[Our Objective should be, for both 'Training' & 'Testing' data our accuracy should be high.]

∴ This Objective is achieved by Case no. (2).

Out of these three models, we will be selecting the middle one, in order to solve a problem.

This case (2) model is giving us ['Low Bias'] & ['Low Variance']

In Underfitting we have 'High Variance' & 'High Bias'

In Overfitting, we always have 'Low Bias' & 'High Variance'

['Bias' basically means the error of the training data.
'Variance' basically means the error of the testing data.]

In Overfitting,

We have 'Low Bias' & 'High Variance'

Now let's go to Classification Problem Statement :->

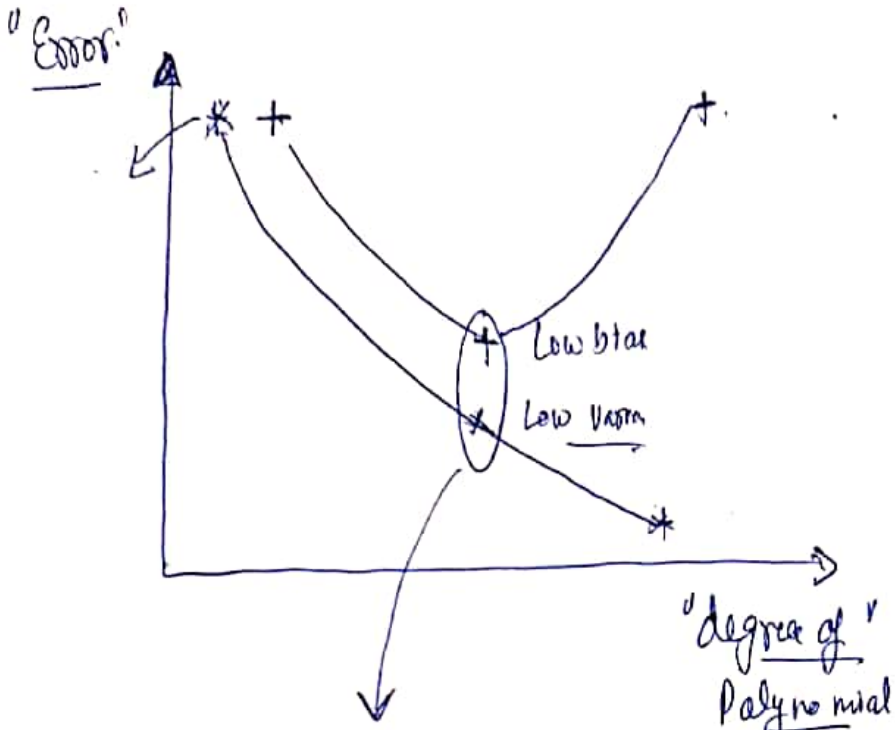
Suppose we have used three models with different hyperparameter tunings.

"Classification"

Model 1	Model 2	Model 3
Training Error = 1%	Training Error = 25%	Training Error < 10%
Test Error = 20%	Testing Error = 26%	Testing Error < 10%
Low Bias.	High Bias	Low Bias
High Variance.	High Variance	Low Variance
'Overfitting'.	"Underfitting"	"Most generalized" model

P.T.O

Graphical Representation of Bias & Variance. with respect to all these three cases.



We are interested in the generalized model. [where we have low bias & low variance]