Machine learning

Machine learning models

Machine learning models are algorithms that can find patterns or make predictions on **unseen data**.

Errors in Machine Learning

In any Machine learning algorithm, we need to mainly focus on 3 important concept

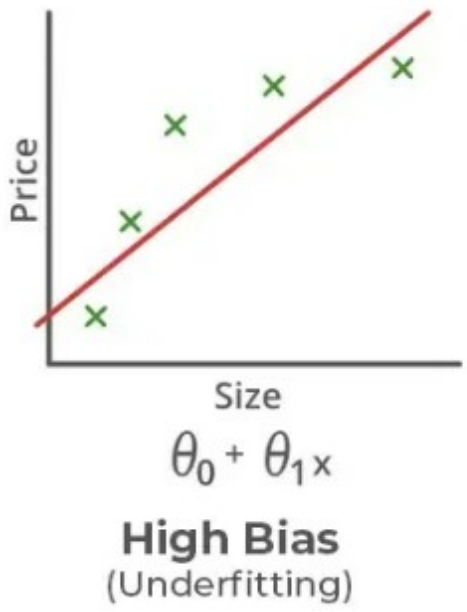
- **predicted values** - values resulted from the model
- **actual values** - original values
- **error** - difference between actual and predicted output

Errors in Machine Learning

There are two main types of errors present in any machine learning model

- **irreducible errors** - are errors which will always be present in a machine learning model, because of unknown variables, and whose values cannot be reduced. It is caused by unusual variables that have a direct influence on the output.
- **reducible errors** are those errors whose values can be further reduced to improve a model. They are caused because our model's output function does not match the desired output function and can be optimized.
  - **bias** - is the difference between our actual and predicted values. (bias is related to training error)
  - **variance** - is our model's sensitivity to any fluctuations in the data. (variance is related to testing error)

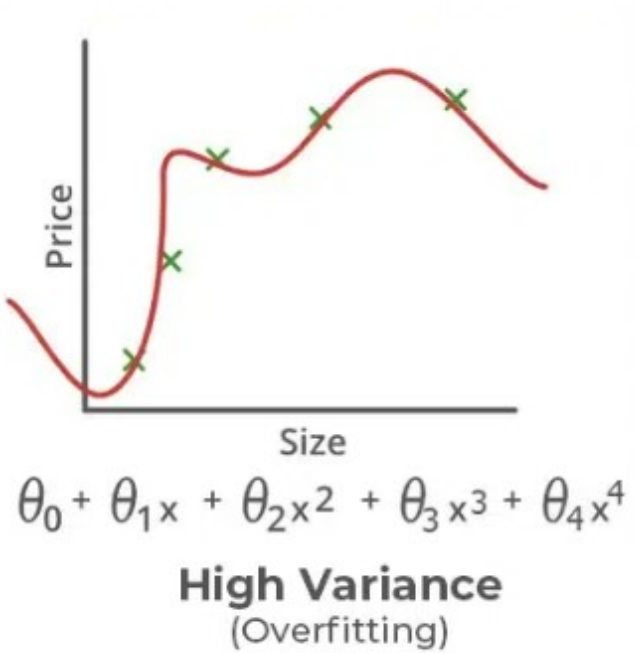
Bias



- **Bias** is the difference between our actual and predicted values.
- **Bias** is the simple assumptions that our model makes our data to be able to predict new data.

When the **bias is high,** assumptions made by our model are too basic, the model can't capture the important features of our data. This means that our model hasn't captured patterns in the training data and hence cannot perform well on the testing data too. **If this is the case, our model cannot perform on new data.**

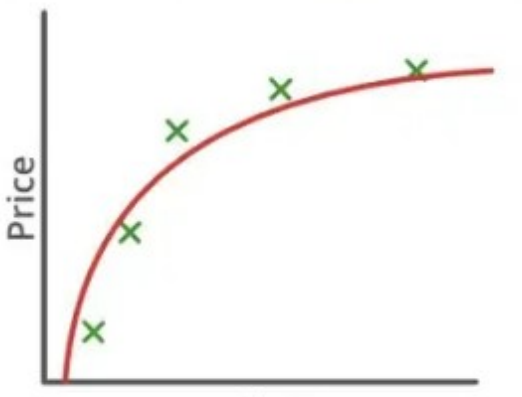
Variance



- **Variance** is our model's sensitivity to any fluctuations in the data.
- **Variance** - shows how our model may learn from noise. This will cause our model to consider trivial features as important.

High variance occurs when a model learns the training data's noise and random fluctuations rather than the underlying pattern. As a result, the model performs well on the training data but poorly on the testing data.

Bias-Variance Tradeoff



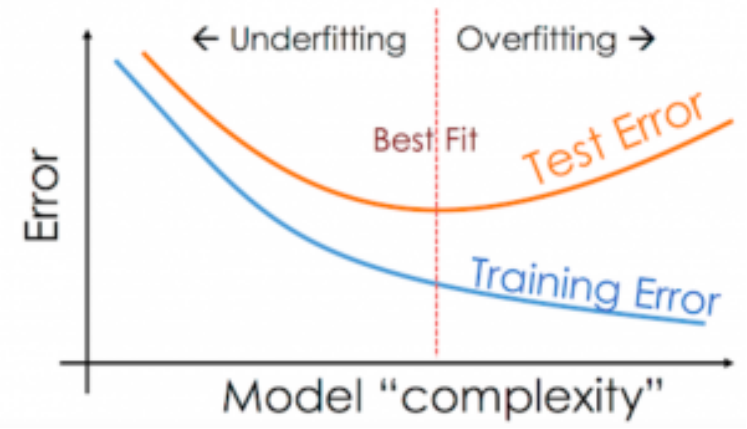
$\theta_0 + \theta_1 x + \theta_2 x^2$

Low Bias, Low Variance
(Goodfitting)

An optimized model will be sensitive to the patterns in our data, but at the same time will be able to generalize to new data. In this, both the bias and variance should be low so as to prevent **overfitting and underfitting**.

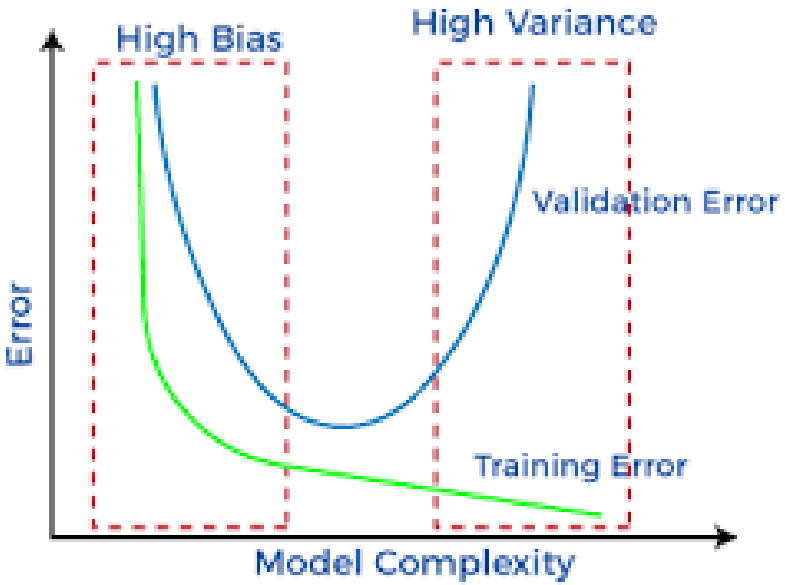
Bias-Variance Tradeoff

- Bias-variance tradeoff is the balance between bias and variance.
- In this case we can capture the essential patterns in our model while ignoring the noise present it in.
- Bias-variance tradeoff helps to optimize the error in our model and keeps it as low as possible.

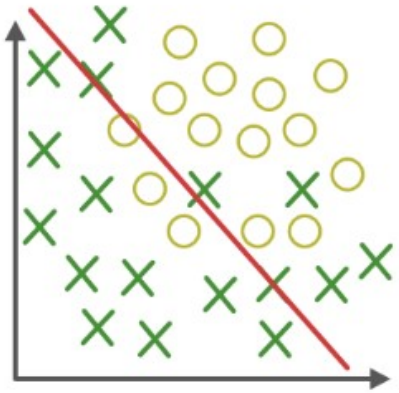


Underfitting and Overfitting

- **Underfitting and overfitting** both introduce error and reduce the generalizability of the model (the ability of the model to generalize to future, unseen data).
- They are also opposed to each other: somewhere between a model that underfits and has bias, and a model that overfits and has variance, is an optimal model that balances the bias variance trade-off.



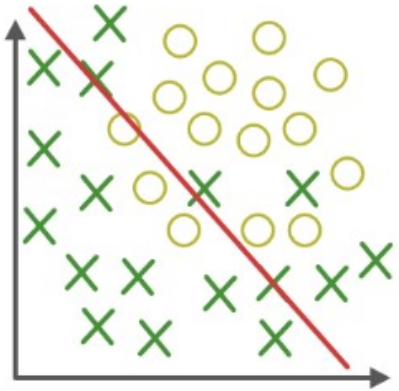
Underfitting



Reasons for **underfitting**

- the model is too simple, so it may be not capable to represent the complexities in the data.
- the input variables which are used to train the model are not the appropriate representations of the underlying factors influencing the target variable.
- the size of the training dataset used is not enough.
- incorrectly selected model parameters.
- variables are not scaled.

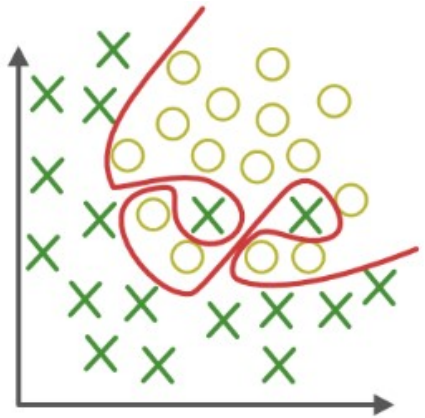
Underfitting



Techniques to reduce **underfitting**

- increase model complexity.
- increase the number of features, performing feature engineering.
- remove noise from the data.
- increase the number of epochs or increase the duration of training to get better results.
- add data to training set.

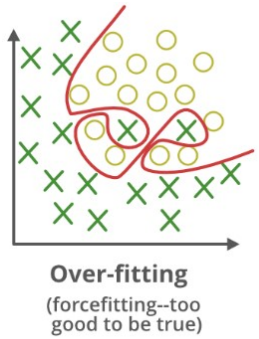
Overfitting



Reasons for **overfitting**

- model is too complex.
- the size of the training data.

Overfitting



Techniques to reduce **overfitting**

- increase the training data can improve the model's ability to generalize to unseen data and reduce the likelihood of overfitting.
- improving the quality of training data reduces overfitting by focusing on meaningful patterns, mitigate the risk of fitting the noise or irrelevant features.
- reduce model complexity.
- early stopping during the training phase (have an eye over the loss over the training period as soon as loss begins to increase stop training).
- use dropout(removal of redundant neurons) for neural networks.

Bias-Variance Tradeoff

Thank you for your attention!!!