Model Evaluation – Overfitting

-

Overview

- Overfitting vs Underfitting
- Underfitting to Overfitting in Prediction(Regression)
- Underfitting to Overfitting in Classification
- Impact of Overfitting on Performance
- Approaches to reduce Overfitting
- Summary

What is Overfitting

- ML models can generalize and produce highly complex explanations of relationships (models) between predictor variables and response variable where the ‘fit’ maybe excellent.

- However, when used with new unseen data, models of great complexity may not do so well!

- Overfitting happens when a model learns the details and also noise in the training data to the extent that it negatively impacts the performance of the model on new data. (excellent performance on training, but poor generalization on new data).

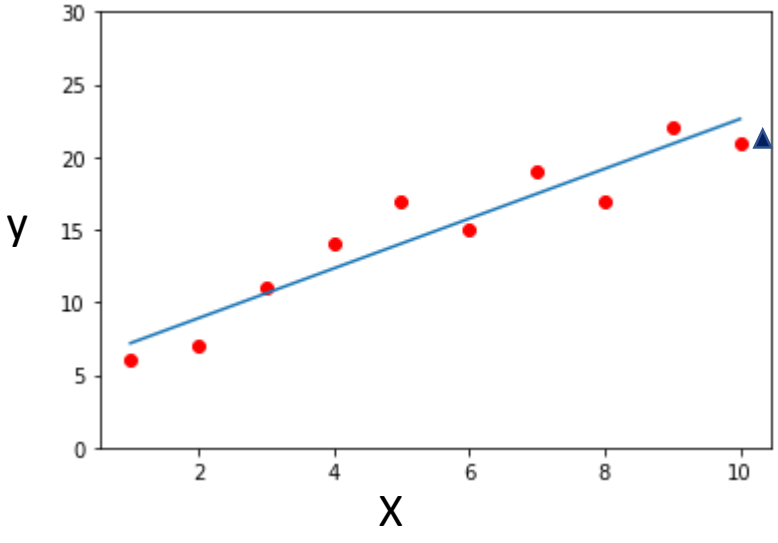
- Overfitting produces poor predictive performance – past a certain point, the error rate on new data starts to increase

- Conversely, Underfitting occurs when the model fails to capture or generalize the underlying pattern in the data, mostly due to over-simplicity of the model. (poor performance on training as well as poor generalization on new data)

Underfitting to Overfitting – Prediction (Regression)

Given a dataset, where

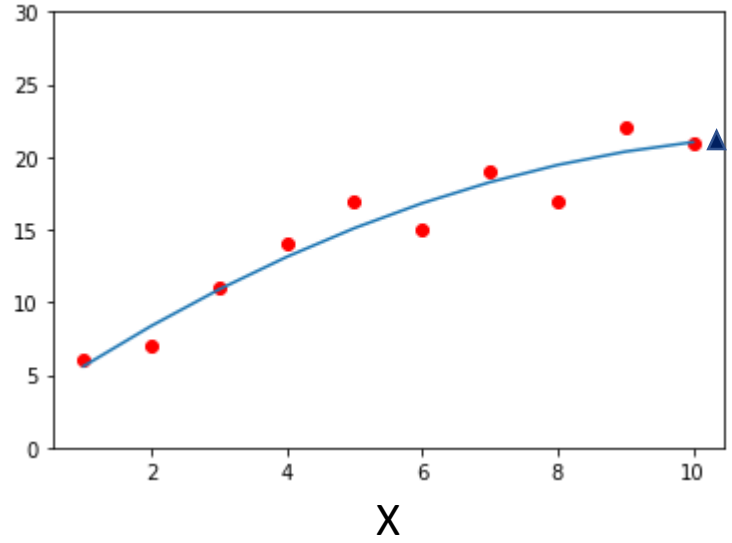
X = [1, 2, 3, 4, 5, 6, 7, 8, 9, 10] and
y = [6, 7, 11, 14, 17, 15, 19, 17, 22, 21]



Linear Model

$y = aX + c$

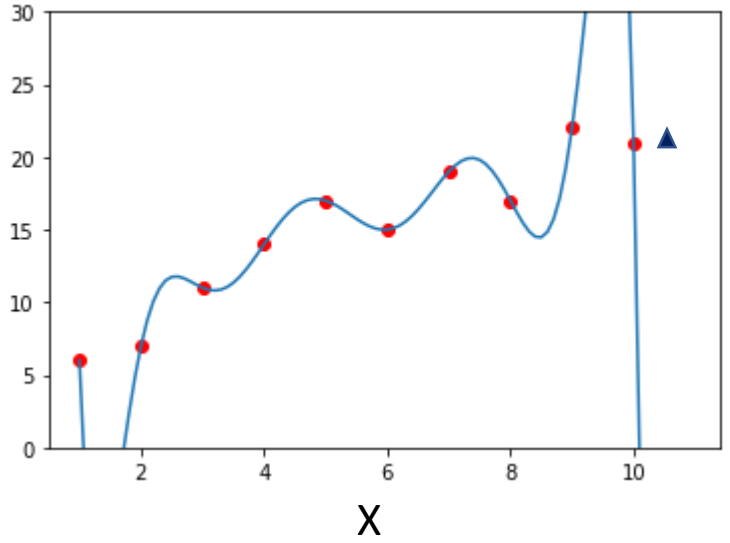
Accuracy: 89.5 %



Quadratic Model

$y = a_1X + a_2X^2 + c$

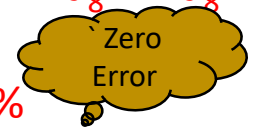
Accuracy: 93%



Polynomial Model

$y = a_1X+a_2X^2+a_3X^3+a_4X^4+a_5X^5+a_6X^6+a_7X^7+a_8X^8+ a_8X^8+c$

Accuracy: 99.99%



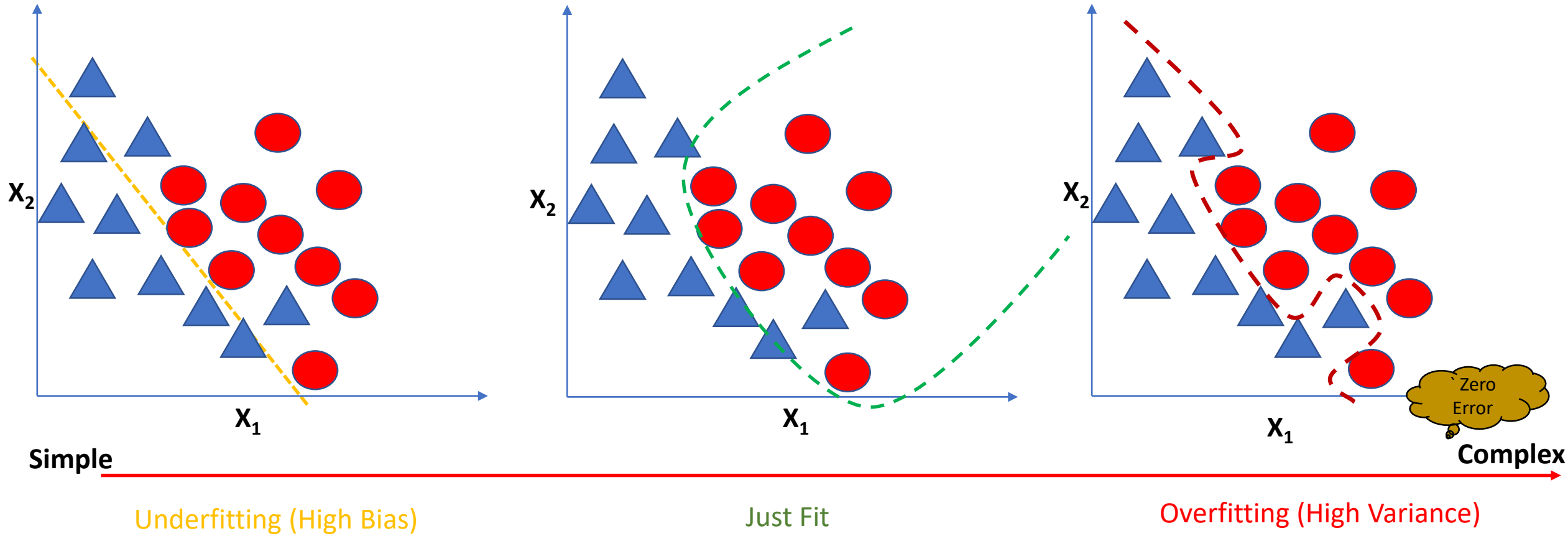
Simple → **Complex**

Underfitting (High Bias)

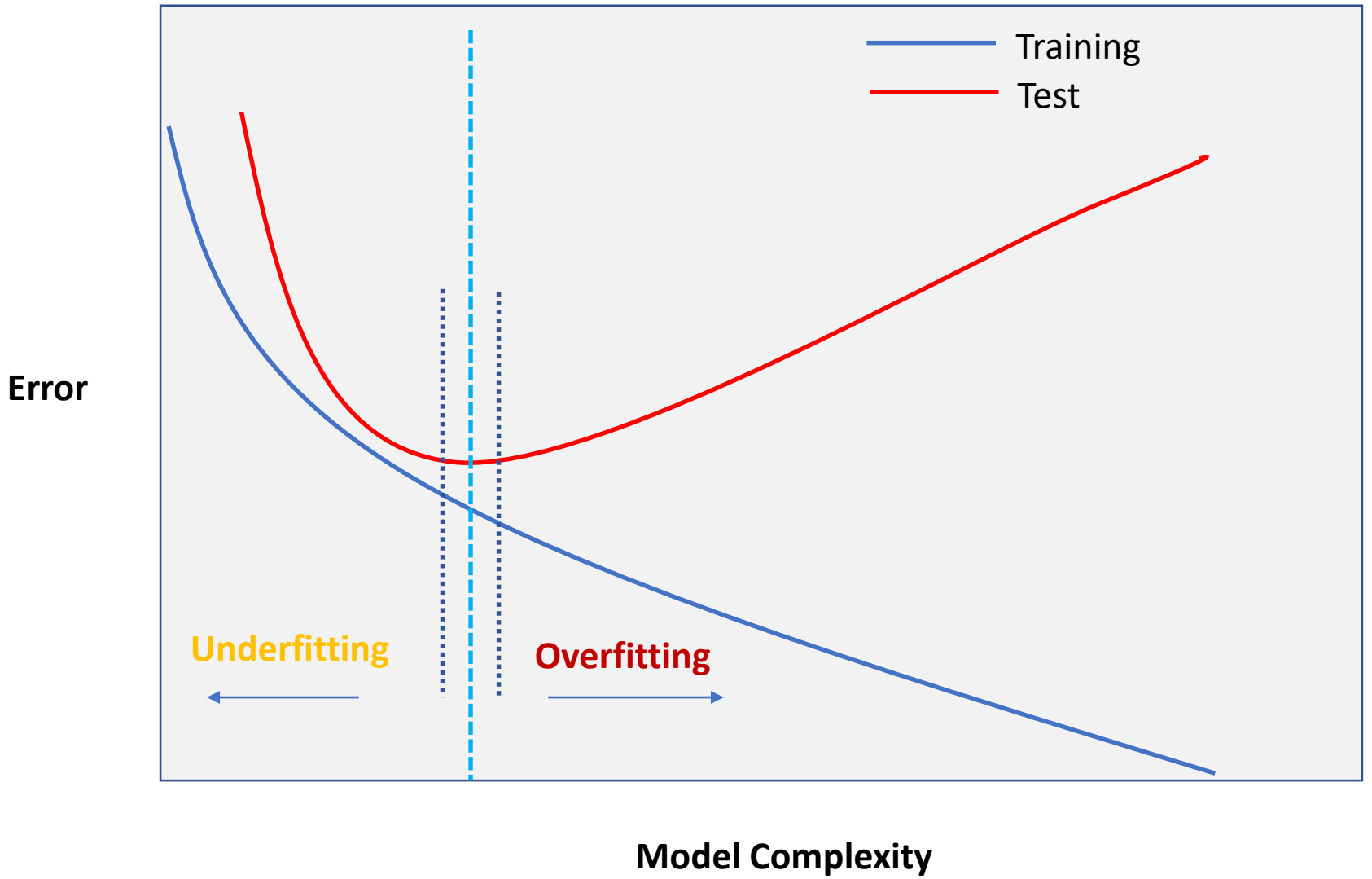
Just Fit

Overfitting (High Variance)

Underfitting to Overfitting – Classification



Impact of Overfitting on Performance



Approaches to reduce Overfitting

- Partitioning the available data into Training – Validation – Test partitions and performing Cross Validation
- Reducing the number of features used in building the model
- Regularization

Summary

- Overfitting : excellent performance on training data, but poor generalization on new data. Also known as High Variance
- Underfitting : Poor performance on training as well as poor generalization on new data. Also known as High Bias
- Complex models with zero error typically end up overfitting
- Partitioning of data into Training-Validation-Test partitions for cross validation, reducing number of features, regularization, etc., are some of the techniques to reduce the impact of overfitting

References

- https://towardsdatascience.com/overfitting-vs-underfitting-a-complete-example-d05dd7e19765
- https://chunml.github.io/ChunML.github.io/tutorial/Underfit-Overfit/
- https://machinelearningmastery.com/overfitting-and-underfitting-with-machine-learning-algorithms/
- https://keeeto.github.io/blog/bias_variance/