Overfitting and Underfitting in Machine Learning

basic term

- **Signal:** It refers to the true underlying pattern of the data that helps the machine learning model to learn from the data.

- **Noise:** Noise is unnecessary and irrelevant data that reduces the performance of the model.

- **Bias:** Bias is a prediction error that is introduced in the model due to oversimplifying the machine learning algorithms. Or it is the difference between the predicted values and the actual values.

- **Variance:** If the machine learning model performs well with the training dataset, but does not perform well with the test dataset, then variance occurs.

What is Overfitting?
When a model performs very well for training [data](#) but has poor performance with test data (new data), it is known as overfitting.

Reasons for Overfitting

- Data used for training is not cleaned and contains noise (garbage values) in it

- The model has a high variance

- The size of the training dataset used is not enough

- The model is too complex

Ways to Tackle Overfitting

- Using K-fold cross-validation

- Using Regularization techniques such as Lasso and Ridge

- Training model with sufficient data

- Adopting ensembling techniques

What is Underfitting?
When a model has not learned the patterns in the training data well and is unable to generalize well on the new data, it is known as underfitting. An underfit model has poor performance on the training data and will result in unreliable predictions. Underfitting occurs due to high bias and low variance.

Reasons for Underfitting

- Data used for training is not cleaned and contains noise (garbage values) in it

- The model has a high bias

- The size of the training dataset used is not enough

- The model is too simple

Ways to Tackle Underfitting

- Increase the number of features in the dataset

- Increase model complexity

- Reduce noise in the data

- Increase the duration of training the data