Overfitting and Underfitting:

Suppose we have a dataset with 1000 data points irrespective of the number of features we have and we have to train a model based on it which will predict something.

We should split the entire datapoints into two important parts:

1. Training dataset (For example 70%): For training

2. Test dataset (For example 30%): For testing the model. The model will not know about these datapoints while training.

We split the training dataset into two:

a) Train – To train the model

b) Validation – Hyperparameter tuning of the model

We can combine train and validation data to improve the performance of our model.

Whenever we get good accuracy wrt training data, we use the term low bias.

Whenever we get bad accuracy wrt training data, we use the term high bias.

Whenever we get bad accuracy wrt test data, we use the term high variance.

Whenever we get good accuracy wrt test data, we use the term low variance.

Scenarios:

1.

Training data: Very good accuracy [Low bias]

Test data: Very Good accuracy [Low variance]

This is what we want, i.e, Generalized model.

2.

Training data: Very good accuracy (Eg: 90%) [Low bias]

Test data: Bad accuracy (Eg: 50%) [High variance]

Overfitting.

3.

Training data: Low accuracy [High bias]  
Test data: Low accuracy [ High variance]

Underfitting.

