29.16 How to determine overfitting and underfitting?

Either by using Simple cross-validation (or) k'-fold cross-validation, we get the best value of 'K'(here hyperparameter in K-NN) for the model, which neither leads the model to overfit nor to underfit.

Accuracy = (Number of points correctly classified)/(Total number of points)

Error = 1 - Accuracy

Our main aim is always to maximize the accuracy and minimize the error.

For now, we shall consider the case of simple cross-validation, for that we have to split the dataset 'D_n' into 'D_{Train}', 'D_{cv}' and 'D_{Test}'.

In simple cross-validation, we use the 'D_{Train}' to find the nearest neighbors, 'D_{cv}' to find the optimal 'K' value and 'D_{Test}' to find out the model performance at prediction on the unseen data.

Training Error

We have to fit the model on 'D_{Train}' and make predictions on the same 'D_{Train}'. Here we come across a few misclassifications while predicting the class labels of 'D_{Train}'. This error obtained is called the **Training Error**.

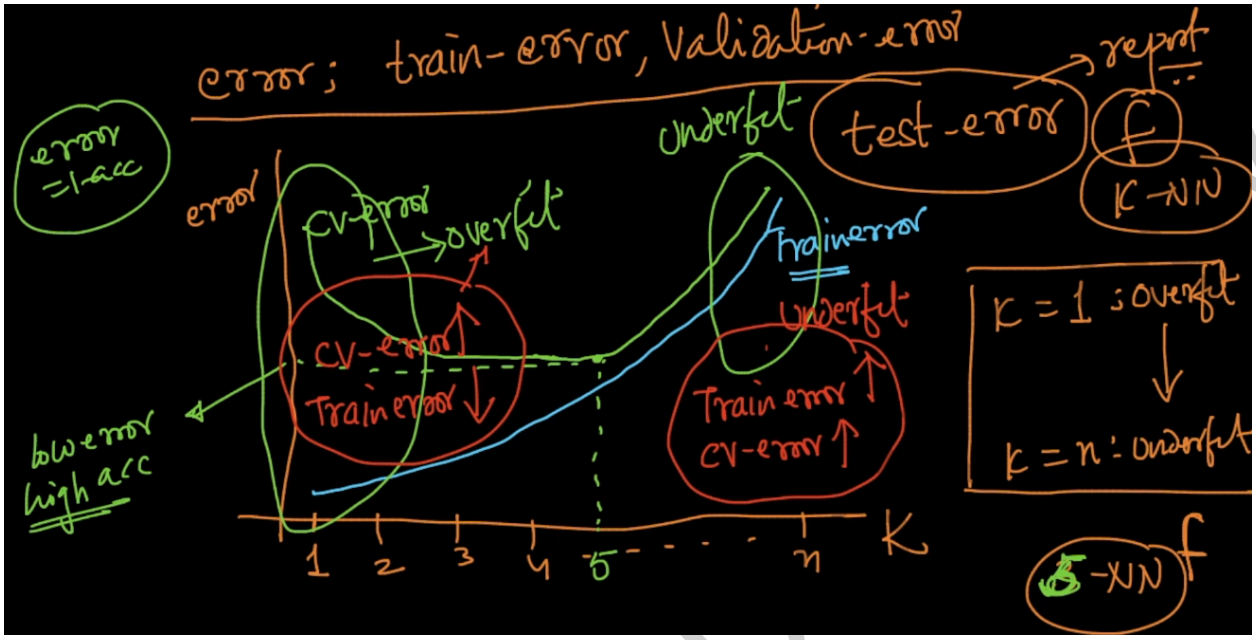
Cross-Validation Error

We have to fit the model on 'D_{Train}' and make predictions on 'D_{cv}'. Here we come across a few misclassifications while predicting the class labels of 'D_{cv}'. This error obtained is called the **Cross-Validation Error**.

Test Error

We have to fit the model on 'D_{Train}' and make predictions on the same 'D_{Test}'. Here we come across a few misclassifications while predicting the class labels of 'D_{Test}'. This error obtained is called the **Test Error**.

Overfitting (vs) Underfitting



If the **Training Error is high** and the **Cross-Validation Error is high**, then we call it **Underfitting**.
If the **Training Error is low**, but the **Cross-Validation Error is high**, then we call it Overfitting.
If the **Training Error is moderate** and the **Cross-Validation Error is moderate**, then we call it the **Best Fit**. (Here both the errors are close enough)