Dataset Splitting in ML, Overfitting and Underfitting

Dr. Manju Venugopalan
Asst Professor (Sr. Gr)
Dept of CSE
Amrita School of Computing, Bengaluru

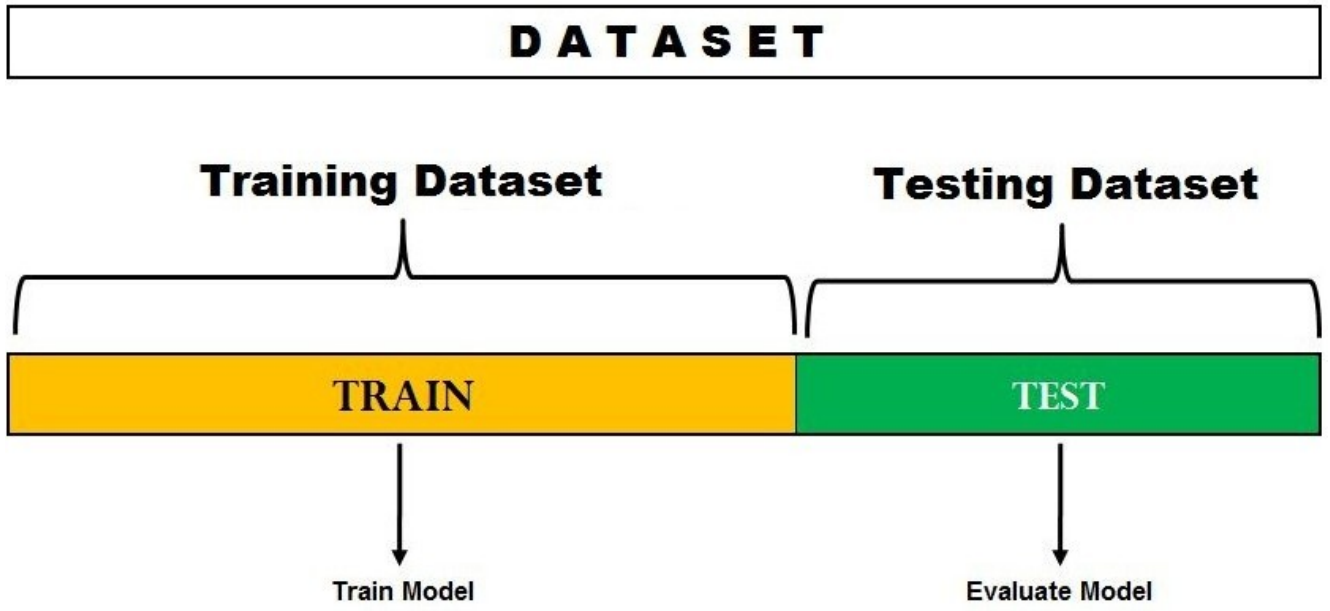
What is Data Splitting?

- Data splitting comes into the picture when the given data is divided into two or more subsets so that a ML model can get trained, tested and evaluated.

- Usually, we create two parts of the main dataset.
  - ✓ **Train-Test Split:** The regular cut up is 70-80% for training and 20-30% for testing, but this may vary depending on the scale of the dataset and the precise use case.

Train-Test Split: Hold Out method

- The data can be divided into 70-30 or 80-20 based on the use case.
- As a rule, the proportion of training data has to be larger than the test data.



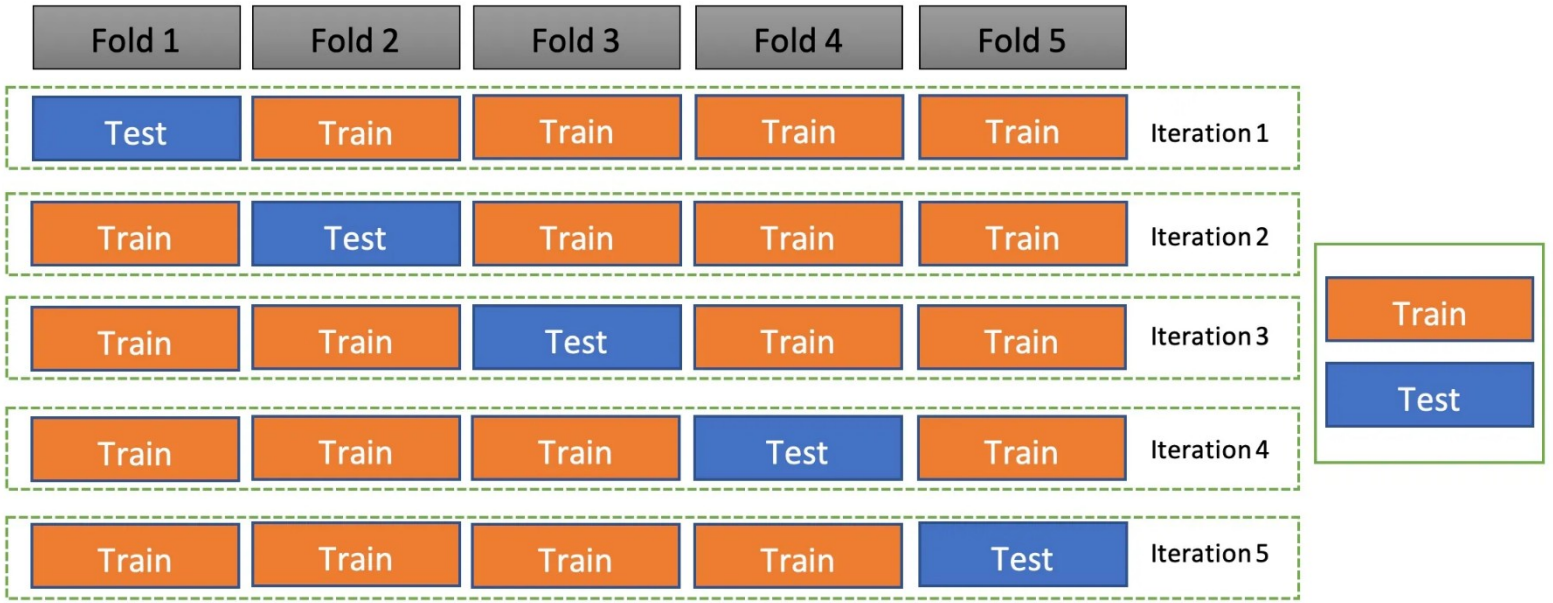
Here is another common process for splitting data:

- **K-fold Cross Validation:** The dataset is divided into k equally sized folds, and the version is educated and evaluated okay instances. Each time, k-1 folds are used for training, and 1-fold is used for validation/testing. This allows in acquiring greater strong overall performance estimates and reduces the variance in model evaluation.

K-Fold Cross-Validation

- In this **resampling technique**, the whole data is divided into k sets of almost equal sizes. The first set is selected as the test set and the model is trained on the remaining k-1 sets. The test error rate is then calculated after fitting the model to the test data.

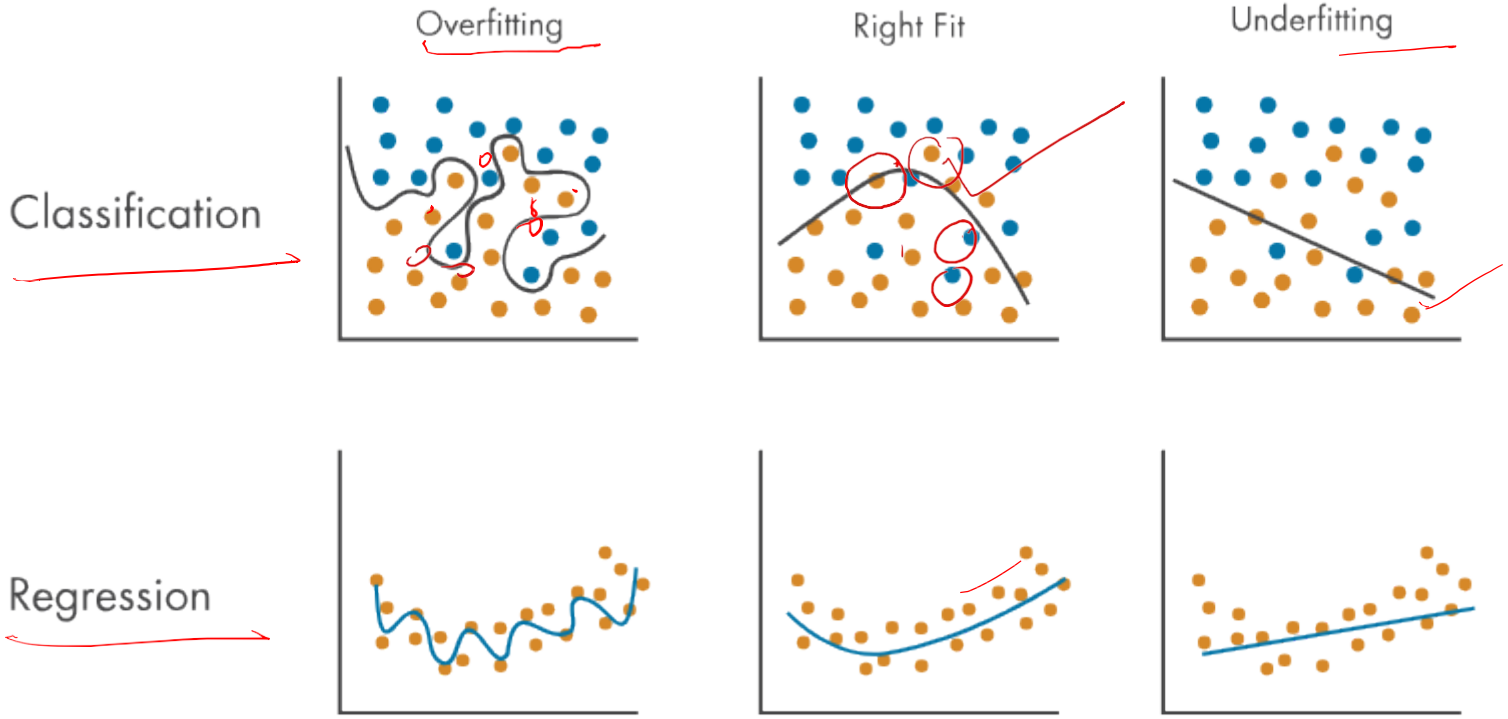
- In the second iteration, the 2nd set is selected as a test set and the remaining k-1 sets are used to train the data and the error is calculated. This process continues for all the k sets.



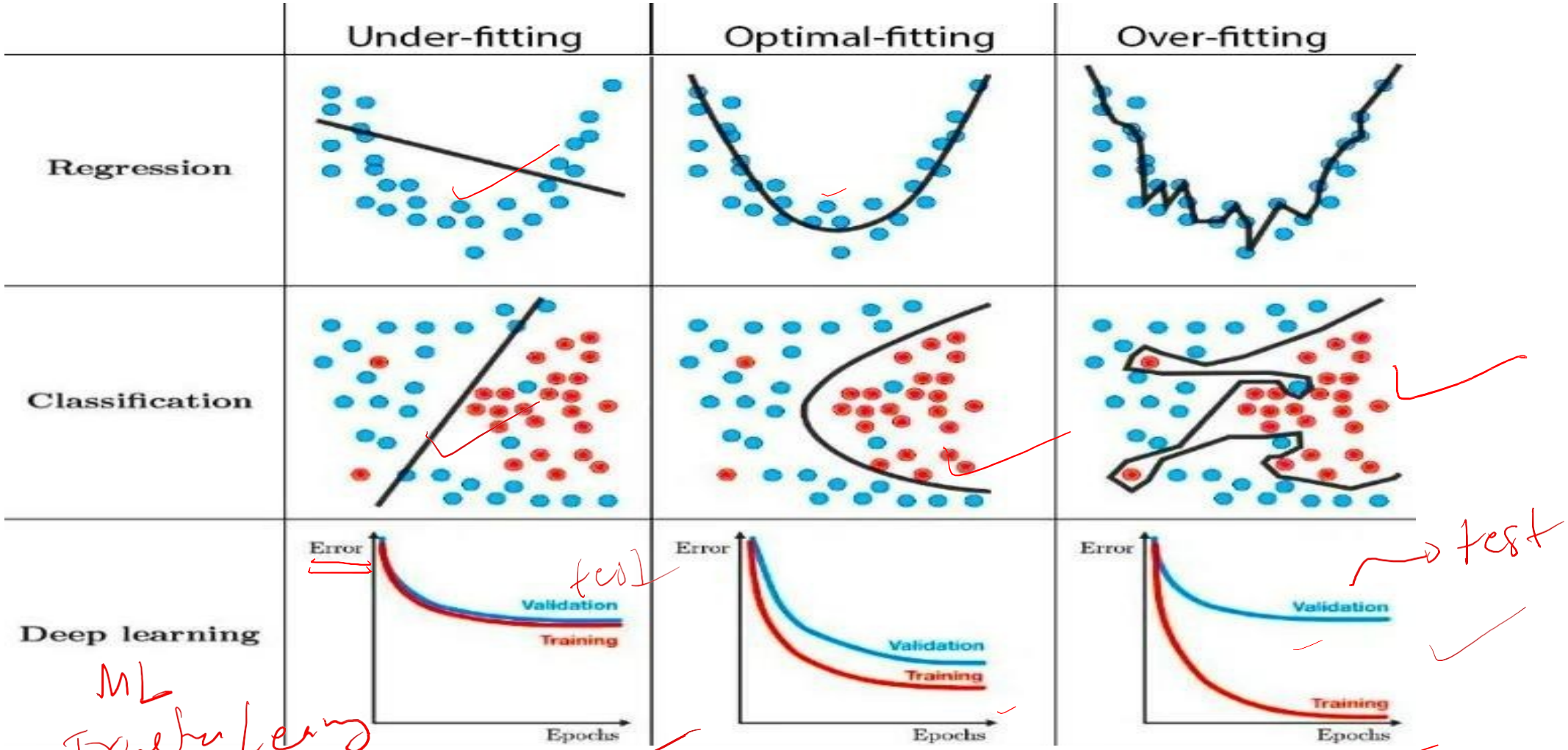
Underfitting and Overfitting Machine Learning Models

- **Overfitting:** Good performance on the training data, poor generalization to other data(test data).

- **Underfitting:** Poor performance on the training data and poor generalization to other data (test data).



Model Behavior in under-fitting, optimal-fitting, Over-fitting



How to Limit Underfitting and Overfitting

- Both overfitting and underfitting can lead to poor model performance. But by far the most common problem in applied machine learning is overfitting.
- **How To Limit underfitting:** The remedy is to add more features, more quantitative and qualitative training data, and try alternate machine learning algorithms.
- **How To Limit Overfitting**:
  - Cross-validation
  - Train with more data (Data augmentation)
  - Feature selection
  - Adopting ensemble techniques

Thank you !!!!!