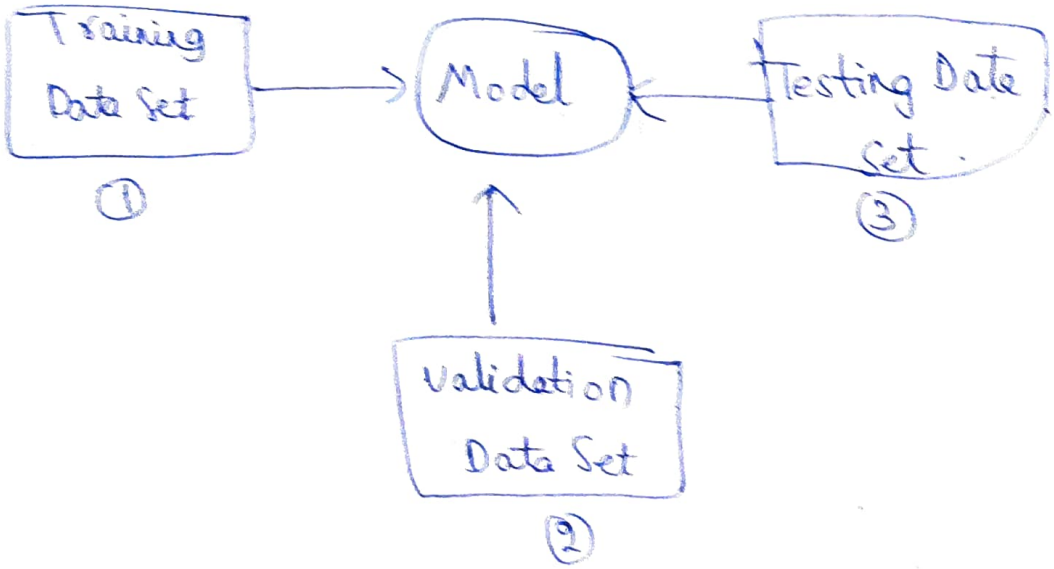
Model Selection and generalization :-

There are 3 steps to determine a model with lowest error.

1. (Train) the system with Training Data Set.
2. (Validate) the output of model using validation data set.
   - perform model selection.
3. (Test) the model using testing dataset - Evaluate the model with test data set



In Constructing a Model

1. The set of assumptions, used to learn the algorithm is called the "inductive Bias".

2. We introduce "Inductive Bias is when we assume a hypothesis class "H".

3. In learning the class of Family cars, there are infinitely many ways of seperating the (+ve) examples from the (-ve) examples.

4. Assuming the shape of a rectangle is one inductive Bias, then the rectangle with the largest margin, is another Inductive Bias

5. The class of functions, that can be learned and extended by using a hypothesis class with larger capacity, Containing more complex hypothesis.

6. In Regression, as we increase the order of the polynomial, the capacity and complexity increases

Underfitting:-

1. Underfitting means the training errors and testing errors are more and the system will be simple.
2. Error rate will be very high.

Overfitting:-

1. If we perfectly match the data points in our training dataset, our model probably won't generalize very well, because the data isn't perfect (there always a bit of noise.

Model selection

1. A Model to generate the right output for input instances (the given training set).
2. A model trained on the training set, predicts the right output for new instances is called Generalization.
3. Match the Complexity of the hypothesis class H with the Complexity of the function underlying the data.

4 Triple Trade off :-

In all learning algorithms that are trained from example data, there is a trade-off b/w 3 factors

a The complexity of hypothesis that fit to data i.e, the Capacity of the hypothesis class.

b The amount of training data.

c The generalization error on new examples.

5 amount of training data increases, the generalization error decreases.

6 As the complexity of the model class H increases, the generalization error decreases first and then starts to increase.

Training Set and Validation Set

1 Dividing the data set to 2 parts.

2 one part is for (training), and remaining part is called (validation set) used to test generalization ability

3 If large training and validation sets, then the hypothesis is the most accurate on the validation set. is the (best inductive bias).

4 This process is called "(cross-validation)"

Generalization:- that is, how well our hypothesis will correctly classify future examples that are not part of the training set.

Most general hypothesis:- G, is the largest rectangle, that includes all the positive examples and none of the negative examples.

Most specific hypothesis:- (S), that is the hypothesis (tighest rectangle) that includes all the positive examples and none of the negative examples.

C is always larger than S..