**Question 1.**

What are the three stages to build the hypothesis or model in machine learning?

**Ans.**

A learning algorithm comes with a hypothesis space, the set of possible hypotheses it can come up with in order to model the unknown target function by formulating the final hypothesis. Classifier: A classifier is a special case of a hypothesis

* Model Building
* Model Testing
* Applying the model

Model: A machine learning model can be a mathematical representation of a real-world process. To generate a machine learning model you will need to provide training data to a machine learning algorithm to learn from.

Algorithm: Machine Learning algorithm is the hypothesis set that is taken at the beginning before the training starts with real-world data. When we say Linear Regression algorithm, it means a set of functions that define similar characteristics as defined by Linear Regression and from those set of functions we will choose one function that fits the most by the training data.

Training: While training for machine learning, you pass an algorithm with training data. The learning algorithm finds patterns in the training data such that the input parameters correspond to the target. The output of the training process is a machine learning model which you can then use to make predictions. This process is also called “learning”.

**Question 2:** What is the standard approach to supervised learning?

**Ans.**

The standard approach to supervised learning is to split the set of examples into the training set and the test set.

**Question 3:** What is training set and test set?

**Ans.**

* Training Set –

In Machine Learning, a training set is a dataset used to train a model. In training the model, specific features are picked out from the training set. These features are then incorporated into the model. Thereby, if the training set is labeled correctly, the model should be able to learn something from these features.

* Test Set –

The test set is a dataset used to measure how well the model performs at making predictions on that test set. If the prediction scores for the test set are unreasonable, then we have to make some adjustments to our model and retry again.

**Question 4.**

What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?

**Ans.**

Ensemble methods are learning algorithms that construct a. set of classifiers and then classify new data points by taking a (weighted) vote of their predictions. The original ensemble method is Bayesian aver- aging, but more recent algorithms include error-correcting output coding, Bagging, and boosting.

The general principle of an ensemble method is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model. bagging is a method in ensemble for improving unstable estimation or classification schemes.

Ensemble Learning — Bagging and Boosting. Bagging and Boosting are similar in that they are both ensemble techniques, where a set of weak learners are combined to create a strong learner that obtains better performance than a single one.

Ensemble methods are meta-algorithms that combine several machine learning techniques into one predictive model in order to decrease variance (bagging), bias (boosting), or improve predictions (stacking). In parallel ensemble methods where the base learners are generated in parallel (e.g. Random Forest).

Bagging is a way to decrease the variance in the prediction by generating additional data for training from dataset using combinations with repetitions to produce multi-sets of the original data. Boosting is an iterative technique which adjusts the weight of an observation based on the last classification

The term 'Boosting' refers to a family of algorithms which converts weak learner to strong learners. Boosting is an ensemble method for improving the model predictions of any given learning algorithm. The idea of boosting is to train weak learners sequentially, each trying to correct its predecessor

**Question 5.:** How can you avoid overfitting ?

**Ans.**

What is ‘Overfitting’ in Machine learning?

In machine learning, when a statistical model describes random error or noise instead of underlying relationship ‘overfitting’ occurs. When a model is excessively complex, overfitting is normally observed, because of having too many parameters with respect to the number of training data types. The model exhibits poor performance which has been overfit.

why overfitting happens?

The possibility of overfitting exists as the criteria used for training the model is not the same as the criteria used to judge the efficacy of a model.

By using a **lot of data** overfitting can be avoided, overfitting happens relatively as we have a small dataset, and we try to learn from it. But if we have a small database and we are forced to come with a model based on that. In such situation, we can use a technique known as cross validation. In this method the dataset splits into two section, testing and training datasets, the testing dataset will only test the model while, in training dataset, the datapoints will come up with the model.

In this technique, a model is usually given a dataset of a known data on which training (training data set) is run and a dataset of unknown data against which the model is tested. The idea of cross validation is to define a dataset to “test” the model in the training phase.

* Cross-Validation
* Pre-pruning
* Post-pruning
* Regularization
* Feature Selection
* Take more data