Problem Statement

1. What are the three stages to build the hypotheses or model in machine learning? ¶

The three stages to build the hypotheses or model in machine learning are:

1. Model building
2. Model testing
3. Applying the model

2. What is the standard approach to supervised learning?

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

There are two ways to split the data into train and test.

1) Split the data into ratio of 20:80, i.e. 20% test data and 80% train data or 30:70 or 35:65 etc

2) Crossvalidation where the whole test train data divided into n fold validation. This will help us to resolve the overfitting.

3. What is Training set and Test set?

What is a 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.

What is a 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, we'll need to make some adjustments to our model and try again.

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

General Principle :-

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. 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)

Bagging :-

Bootstrap aggregating, also called bagging, is a machine learning ensemble meta-algorithm designed to improve the stability and accuracy of machine learning algorithms used in statistical classification and regression. It also reduces variance and helps to avoid overfitting.

Boosting :

Boosting is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms that convert weak learners to strong ones. ... Algorithms that achieve hypothesis boosting quickly became simply known as "boosting".

5. How can you avoid overfitting ?

There are multiple waye to avoid the overfitting.

1) By using a lot of data overfitting can be avoided, overfitting happens relatively as you have a small dataset, and you try to learn from it.

2) But if you have a small database and you are forced to come with a model based on that. In such situation, you 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.

3) The number of independent parameters in your fit is much smaller than the number of data points you have.

4)