

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

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.

3. What is Training set and Test set?

In various areas of information science like machine learning, a set of data is used to discover the potentially predictive relationship known as 'Training Set'. Training set is an examples given to the learner, while Test set is used to test the accuracy of the hypotheses generated by the learner, and it is the set of example held back from the learner. Training set are distinct from Test set.

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

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 In ensemble: bagging is a method in ensemble for improving unstable estimation or classification schemes. Bagging both can reduce errors by reducing the variance term. Boosting in ensemble: boosting method are used sequentially to reduce the bias of the combined model. Boosting can reduce errors by reducing the variance term.

5. How can you avoid overfitting ?

Following are the commonly used methodologies :

Cross-Validation : Cross Validation in its simplest form is a one round validation, where we leave one sample as in-time validation and rest for training the model. But for keeping lower variance a higher fold cross validation is preferred.

Early Stopping : Early stopping rules provide guidance as to how many iterations can be run before the learner begins to over-fit.

Pruning : Pruning is used extensively while building CART models. It simply removes the nodes which add little predictive power for the problem in hand.

Regularization : This is the technique we are going to discuss in more details. Simply put, it introduces a cost term for bringing in more features with the objective function. Hence, it tries to push the coefficients for many variables to zero and hence reduce cost term.

Steps for reducing overfitting:

Add more data.

Use data augmentation.

Use architectures that generalize well.

Add regularization (mostly dropout, L1/L2 regularization are also possible)

Reduce architecture complexity.

In []:

