MACHINE LEARNING - 1 - Assignment

By Prakash Ghosh ¶

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 as follows:

- Model building
- Model testing
- Applying the model

2. What is the standard approach to supervised learning?

Supervised

Supervised learning algorithms are trained using labeled examples, such as an input where the desired output is known. For example, a piece of equipment could have data points labeled either “F” (failed) or “R” (runs).

In order to solve a given problem of supervised learning, one has to perform the following steps:

- Determine the type of training examples. Before doing anything else, the user should decide what kind of data is to be used as a training set. In case of handwriting analysis, for example, this might be a single handwritten character, an entire handwritten word, or an entire line of handwriting.

- Gather a training set. The training set needs to be representative of the real-world use of the function. Thus, a set of input objects is gathered and corresponding outputs are also gathered, either from human experts or from measurements.

- Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented. Typically, the input object is transformed into a feature vector, which contains a number of features that are descriptive of the object. The number of features should not be too large, because of the curse of dimensionality; but should contain enough information to accurately predict the output.

- Determine the structure of the learned function and corresponding learning algorithm. For example, the engineer may choose to use support vector machines or decision trees.

- Complete the design. Run the learning algorithm on the gathered training set. Some supervised learning algorithms require the user to determine certain control parameters. These parameters may be adjusted by optimizing performance on a subset (called a validation set) of the training set, or via cross-validation.

- Evaluate the accuracy of the learned function. After parameter adjustment and learning, the performance of the resulting function should be measured on a test set that is separate from the training set.

3. What is Training set and Test set?

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 that is independent of the training 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 of an 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:
Bagging is a method in ensemble for improving unstable estimation or classification schemes. Bagging both can reduce errors by reducing the variance term.

Boosting:
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 ?

We can avoid overfitting by followings:

Cross-validation:
Use the initial training data to generate multiple mini train-test splits. Use these splits to tune the model.

Train with more data:
Training with more data can help algorithms detect the signal better.

Remove features:
Improve the generalizability of the training data by removing irrelevant input features.

Early stopping:
When you’re training a learning algorithm iteratively, you can measure how well each iteration of the model performs. Up until a certain number of iterations, new iterations improve the model.

Early stopping refers stopping the training process before the learner passes that point.

Regularization:
Regularization refers to a broad range of techniques for artificially forcing your model to be simpler.

The method will depend on the type of learner you’re using. For example, you could prune a decision tree, use dropout on a neural network, or add a penalty parameter to the cost function in regression.

Ensembling:
Ensembling is to combine the predictions of several models built with a given learning algorithm in order to improve robustness over a single model.