Session 11

Assignment 1

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

   SOLUTION:

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

   (a) Model Building - To build a machine learning model, the first step is to form the machine learning problem and identifying what the model should predict. Based on this, data has to be collected, cleaned and prepared to be consumed by the machine learning models. Analyze the data to see what identify any issues and to gain insights to the data being used. Feed the data to the machine learning algorithm to build models.

   (b) Model Testing - After building the model it is important to test the performance of the model by using unseen data. Use the model to predict the data on the test set to compare to the actual truth. Various metrics should be collected to measure the accuracy of a model.

   (c) Applying the Model - Once the Machine learning metrics of the testing looks good, it can be used to make new predictions.2. What is the standard approach to supervised learning?

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

---

2. What is the standard approach to supervised learning?

SOLUTION:

Supervised learning is a type of machine learning in which the model is trained using the input and the output data.

The learning algorithm receives a set of inputs along with the corresponding correct outputs, and the algorithm learns by comparing its actual output with correct outputs to find errors. It then modifies the model accordingly.

Supervised learnings can be of classification, regression or anomaly detection types.

The steps involved in a supervised learning include:

(a) Determine what kind of data is to be used as a training set.

(b) 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.

(c) Determine the input feature representation of the learned function. The accuracy of the learned function depends strongly on how the input object is represented.

(d) Determine the structure of the learned function and corresponding learning algorithm.

(e) Complete the design. Run the learning algorithm on the gathered training set.

(f) 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?

SOLUTION:

Training Set: This is the set of data that is used to train the machine learning model. In training the model, specific features are picked out from the training set. These features are then incorporated into the model. The model would learn from the features provided in the training set. Hence the data provided here should be of very high standard.

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, the model should be adjusted till an optimum performance is obtained. Since the model was trained using the Training set, the same set should not be used to test, since that would give misleading results.

---

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

SOLUTION:

Ensemble model combines multiple individual and diverse models together and delivers superior prediction power. 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 or Bootstrap Aggregating is a method in ensemble for improving unstable estimation or classification schemes. First, create random samples of the training data set (sub sets of training data set). Then, build a classifier for each sample. Finally, results of these multiple classifiers are combined using average or majority voting. Bagging can reduce errors by reducing the variance term.

Boosting method are used sequentially to reduce the bias of the combined model. The first predictor is learned on the whole data set, while the following are learnt on the training set based on the performance of the previous one. It starts by classifying original data set and giving equal weights to each observation. If classes are predicted incorrectly using the first learner, then it gives higher weight to the missed classified observation. Being an iterative process, it continues to add classifier learner until a limit is reached in the number of models or accuracy.

---

5. How can you avoid overfitting?

SOLUTION:

Following are the methodologies that can be used to avoid overfitting:

(a) Cross-Validation: Use your initial training data to generate multiple mini train-test splits. Use these splits to tune your model.

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

(c) Early Stopping: While training a learning algorithm iteratively, measure how well each iteration of the model performs. Up until a certain number of iterations, new iterations improve the model. After that point, however, the model's ability to generalize can weaken as it begins to overfit the training data. Early stopping refers stopping the training process before the learner passes that point.

(d) Pruning: Pruning removes the nodes which add little predictive power for the problem in hand.

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

(f) Ensembling: Ensembles are machine learning methods for combining predictions from multiple separate models.
