Machine learning assignment 1 session 19

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

a) Model building

b) Model testing

c) Applying the model

2. What is the standard approach to supervised learning?

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

Training set used to train the model in other terms make the learner to learn from the training data set.

Test data set is unseen data for the learner to test the model.

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 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?

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.

BAGGING

Several estimators are built independently on subsets of the data and their predictions are averaged. Typically the combined estimator is usually better than any of the single base estimator.

Bagging can reduce variance with little to no effect on bias.

ex: Random Forests

BOOSTING

Base estimators are built sequentially. Each subsequent estimator focuses on the weaknesses of the previous estimators. In essence several weak models "team up" to produce a powerful ensemble model. (We will discuss these later this week.)

Boosting can reduce bias without incurring higher variance.

ex: Gradient Boosted Trees, AdaBoost

5. How can you avoid overfitting ?

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. 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.

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.