SESSION 14: ASSIGNMENT 1

Q1. What are the three stages to build the hypotheses or model in machine learning?

Answer:

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

Q2. What is the standard approach to supervised learning?

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

Q3. What is Training set and Test set?

Answer: 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

Q4. What is the general principle of an ensemble method and what is bagging and boosting in ensemble method?

Answer:

Ensemble methods: Ensemble methods are techniques that create multiple models and then combine them to produce improved results. Ensemble methods usually produces more accurate solutions than a single model would. This has been the case in a number of machine learning competitions, where the winning solutions used ensemble methods.

Bagging: The name Bootstrap Aggregating, also known as "**Bagging**", summarizes the key elements of this strategy. In the bagging algorithm, the first step involves creating multiple models. These models are generated using the same algorithm with random sub-samples of the dataset which are drawn from the original dataset randomly with bootstrap sampling method. In bootstrap sampling, some original examples appear more than once and some original examples are not present in the sample. If you want to create a sub-dataset with m elements, you should select a random element from the original dataset m times. And if the goal is generating n dataset, you follow this step n times.

In bagging, each sub-samples can be generated independently from each other. So generation and training can be done in parallel.

Boosting: The term "boosting" is used to describe a family of algorithms which are able to convert weak models to strong models. The model is weak if it has a substantial error rate, but the performance is not random (resulting in an error rate of 0.5 for binary classification). Boosting incrementally builds an ensemble by training each model with the same dataset but where the weights of instances are adjusted according to the error of the last prediction. The main idea is forcing the models to focus on the instances which are hard. **Unlike bagging, boosting is a sequential method, and so you cannot use parallel operations here**

Q5. How can you avoid over fitting?

Answer:

The possibility of over fitting happens as the criteria used for training the model is not the same as the criteria used to judge the efficiency of a model.

Over fitting can be avoided by using a lot of data over fitting can be avoided, over fitting 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 data points 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.