Data Science Masters :Assignment 19

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

Answer:
The three stages to build the hypotheses or model in machine learning are:
1. Model building - Once the data is cleaned up the next step would be choosing
appropraite ML algorithm and developing the Model.

2. Model testing  - Once the Model is built the next step would be testing the same with
the test data set.

3. Applying the model & Deployment - After evaluating the ML model and the results are
acceptable then deploy the model and use it.
```

```
# 2. What is the standard approach to supervised learning?

Answer:

The standard approach to supervised learning is to split the given data set into training
and test sets. It is good to use 70-80% of data for training the machine and use the
remaining % to test the model (Assumed we have enough volume of data to split and use)
```

```
# 3. What is Training set and Test set?

Answer:
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 i.e. a set of data that is used to discover the potentially
predictive relationship known as ‘Training Set’.

The test set is a dataset used to measure how well the model performs at making
predictions on that test set.

The important note here is that 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?

Answer:
The principle of an ensemble method is to combine the predictions of several models built
with a given machine learning algorithm in order to improve accuracy over a single model.


Bagging is a method in ensemble for improving unstable estimation or classification
schemes.

Boosting method are used sequentially to reduce the bias of the combined model.

Boosting and Bagging both can reduce errors by reducing the variance term.
```

```
# 5. How can you avoid overfitting ?

Answer:
By using a good volume 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 where the given
dataset would be splitted into two sets(Train and test sets).

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.
```