Machine Learning -1

Haribansh Kumar Agrawal

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

***Answer*:** Following are the three stages to build the model in machine learning:-

1- Data Pre-processing:- when we get a set of data to build the machine learning model, it requires different pre-processing action to clean the data like formatting, cleaning and sampling.

2- Build and testing the model: - when our data gets cleaned up we apply a fraction of that data on a suitable machine learning algorithm. Algorithm will train itself by that data and then builds a model. After model formation we apply unseen data to the model to test the model.

3- Applying the model: - After building and testing the model we apply that model in the real world problems.

****

What is the standard approach to supervised learning?

Answer: In supervised learning we get labelled data for building the model. So first we split the data in training and testing data. Build model on training data and then apply testing data on model to test the model.
We'll test the model by comparing the predicted value by the model to the labelled values of data.

****

:-What is Training set and Test set?

***Answer:* -** when we get a data set we divide them into Training and Test set.
Training set: - it is a dataset we use to train a model. We apply this data set onto a machine learning algorithm. if the training set is labelled correctly then the model should be able to learn the relationship between the data value(input) and their corresponding label(output).
Test set: - it is uses to test the accuracy of prediction generated by the model. it is the unseen dataset to the model. So we apply it on the model and see how well it predicted all values.

****

What is the general principle of an ensemble method and what is bagging and Boosting in ensemble method?

Answer:
Ensemble Learning Method: - It means grouping multiple week learning models to form a strong model so that we can obtain better prediction.
Begging: - It is uses in Low Bias - High Variance problem. This problem occur when the model over-fits like in Decision tree.

- In this case various models are built in parallel and each model gets trained on randomly selected samples.
- Then the various models vote to give the final prediction.
- Predictions will be averaged to get the final prediction (in case of regression) and in case of classification the final prediction will be the mode of the predicted ans.

Boosting: - It is uses in Low variance - High Bias problem. This problem occurs when model under-fits.

1. In this method we first sampled the input data to generate a set of training data.
2. Then we run an algorithm on this training data to get a trained model.
3. Then we take all our training data to test the model and we are going to discover that some of the points are not well predicted.
4. Now we'll build the second bag of sampled data. Here also the data will be randomly chosen, but now each data point is weighted according to error found in last model. So these values are more likely to get picked in this bag than any other else.
5. Now we'll build a model for this sample set also then we'll test it.
   Here the testing will be performed on both of the model and the result will be mode of the both result (in case of classification) in case of regression result will be mean of the both result.
6. Now again we'll find some values which are not predicted well. so we'll build one more bag and one more model and this process will continue.

****

How can you avoid overfitting?

Answer : when we train the model on training set to such a level that it starts predicting noise or outlier present in training data correctly. Then we say that model gets over-fit. Because in this case, model will show 100% accuracy for training set but its accuracy will be low on test data.
There are multiple methods by which we can avoid overfitting -
1- Ensemble method: - it is the best method to avoid overfitting and increasing accuracy.
2- Remove features: - its example is Pruning, which we do in case of decision tree model.
3- Cross validation