Introduction

Machine learning is programming computers to optimize a performance criterion using past experience or data. This is usually applicable to problems which change over time (behvaioural change on part of computers is required, this change can be temporal, spatial or both). It is also applicable to problems for which human experitse does not exist or where it is not explainable. If this is applied to large databases then it is known as data mining.

However in machine learning, contrary to popular belief, an entire algorithm is not devised by the computer, the algorithm (or model in machine learning vocabulary) is decided by a human expert, this model is applicable to capture the trend in the real life situation which the machine is trying to learn. The learning part is responsible for learning the parameters of this model or algorithm. We require to optimize the model and once the learning is done then prediction of new values or applicatin of that model should be fast enough as well.

Machine learning applications include:

- Learning association – The probability P(y|x) i.e. probability of occurrence of y given that x has already occurred.
- Classification – Again this can be explained as finding probability P(y|x) i.e. probability of an entity of belonging to class ‘y’ given that it possesses properties ‘x’.
- Regression – Problems where the output of the machine learning algorithm/model is quantitative are known as regression problems

Response surface design is about conducting experiments for creation of data which will guide the learning process. The experiments are conducted and a value (output) is chosen which needs to be maximized; based on the experiments conducted thus far parameters are adjusted so that output is maximized, this is also counted as experiment and used for further improvement. This process is continued till the output reaches some threshold value or it can be continued indefinitely as well.

In unsupervised learning, only the input is available in the data set and the aim is to build models which can find regularities in the input. In case of reinforcement learning the machine is expected to learn from its actions and generate a good policy, which is defined by its goal state (both the goal and the method used to reach the goal are important in this case).

Supervised Learning

Version Space

Given a hypothesis class and the probability distribution from which examples are drawn we can find the most specific class, i.e. the tightest class which contains all the positive examples and none of the negative examples and we can also find the most general class, i.e. the least tight rectangle which contains all the positive examples and no negative examples. All the classes between these two classes make up the version space. Any example falling in the version space can be considered as a case of doubt.
Using version space a method called candidate elimination is devised which draws one example then finds its version space, then the next one and repeats this process, this way version spaces are decreased by more and more examples till we arrive at the smallest one after all the examples have been listed.

Vapnik-Chervonenkis Dimension

Given N points 2^N different classification problems can be defined. If a hypothesis class can solve all of these 2^N problems then it is said to shatter N points. The highest number of N is known as VC dimension for the given hypothesis class. It is enough to find any one configuration of N points which are shattered by hypothesis class, not all configurations need not be shattered. For axis aligned rectangles the VC dimension is 4.

The real class C is capapble of having any shape, and hypothesis can at best be a very good approximation of the reality. If the VC is required to infinite then the shape is required to be formless.

Noise

Noise is any unwanted anomaly in data, due to noise the class can become more difficult to learn as zero error becomes infeasible. The noise can arise due to the following reasons:

- Imprecision in recording input attributes
- Error in labelling of data points
- Additional attributes exist which have not been taken into account

To accomodate for noise two solutions are possible, either to have a more complex hypothesis class which increases the number of parameters required to be trained; or one can keep the model simple and allow certain degree of error in the learning. The advantage of using the latter are as follow:

- Simple model is computationally faster in giving output
- Training is easier for simpler model due to lesser number of parameters (such a model has low variance as it does not vary much with new attributes, however such a model is more rigid to change and is thus said to have large bias, an optimum model seeks to have low variance as well as bias)
- Simple model is easier to explain i.e. knowledge extraction is easier with a simple model
- A simple model generalizes better than a complex model. Even according to Occam's razor any unneccesary complexity in the model can cause overfitting and should not be preferred over a simpler model.

Regression

If the output is of numeric value and no noise is associated then the task is known as interpolation and otherwise in presence of noise this is known as regression.

Model selection and generalization

An ill-posed problem is one which where the data used is not enough for learning the actual class. This is the case with most problems and thus to learn the class we make certain assumptions about the model, these are known as the inductive bias. The ability of the model trained on training data to predict the reality is known as generalization. If the hypothesis class is less complex when compared to the actual class then the hypothesis is said to underfit, for vice versa it is known as overfitting. In all types of learning the following three trade-offs are involved:

- Complexity of hypothesis
- Amount of training data
- Generalization on new examples

As the amount of training data increases the generalization error decreases, with increase in complexity the generalization error may decrease in the beginning but at one point of time it starts to increase (where overfitting starts to occur).

To take generalization into accout the training data is split to include a validation set as well. However this cannot be used in papers as often the validation set is part of chosing the correct model, for that we need another set of sample known as the publication set.

The complete task of supervised learning can be defined as:

- Obtain a model to learn its parameters based on data
- Use a loss function to compute the difference between model and the real class
- Have an optimization procedure, which reduces the value of the loss function

These three are the parameters in which difference in machine learning algorithms arise, it is either the model chose, the loss function or the optimization procedure.