**TTA Home Learning Week 06 – Assignment**

An algorithm is a set of instructions that deals with a specific issue and how to resolve it. It is often represented by a workflow.

Human beings brains do work using algorithms. This understanding is underpinning the idea of machine learning. Indeed, as humans can learn new algorithms, scientists believe that this can be reproduced by machines. However, as scientists are not yet exactly sure on how human brains function and are able to reproduce and acquire learning, machine learning remains relatively simple, yet it is based on being able to learn and provide the desired outcome when confronted to new data. For instance, as a child, we learn what a "tree" looks like. Throughout our life, we see different "trees" and we are able to label those as "tree" despite differences in shape, size and colour. Each new "tree" we see is a brand new piece of data, but our brains are able to label the object correctly. This is what machine learning is trying to reproduce.

Machine learning refers to the ability of a machine to apply an algorithm (called a model) to new data and then to adjust this model to fit the new data into whatever it is trying to achieve.

In all instances, the model is always based on existing data and the process (model) provides the expected result. If we stayed at this level, the machine would not learn, it would just validate/check data from its existing reference.

There are several ways to get a machine to learn. The simplest and most reliable is called supervised learning. In this process, new data and the outcome are checked and confirmed or infirmed. This ensures that the machine is learning correctly. In other words, the machine add new data and its outcome to its existing reference data. In this context, each piece of data in the reference material has a specific label (categorised) and new data is checked whether it is correctly labelled according to the model. Each validated piece of data is integrated in the reference data. As such, the reference data grow and (hopefully) improves the model accuracy. This approach focusses on the result/outcome.

Another way to deal with machine learning is to use an unsupervised approach mostly based on statistics. Here again, there is a set of reference data that have been analysed using some statistical approach to distribute the data into various groups. The data in this instance is not labelled. When confronted to new data, the machine looks at which group best represents this new element and then add it to its referential. This approach focusses on the data.

Both approaches use statistics to reach a decision about the new data, but the use of labels in one instance and the absence of label in the other makes that their use is different. With labelling, we often have an A - !A dichotomy that will provide a yes or no answer (a result). This approach enables systems such as fraud detection (it this a fraudulent card activity? Yes – no; is there a car in this image? Yes-no; is the system running? Yes-no; etc.). Using some inferential statistics (mostly regression algorithms) , it becomes possible to predict or forecast previously unseen behaviours.

With unsupervised learning, we can have a categorisation based on statistics (cluster A or cluster B with return on out-of-cluster data) called clustering. This means that it is possible to group data into groups and to do this either directly (cluster) or indirectly (association). This approach also enables big data visualisation. Here, the type of statistics used will be slightly different.

As with everything, it is also possible to create hybrid models that will use both supervised and unsupervised machine learning to handle data.

As we started with algorithms, it is time for us to investigate further some of the most commonly used.

Linear regression is often used for an input ***a*** giving an output ***b***. It means that it is possible to plot the input value on the x-axis and the ouput value on the y-axis. Using statistics you can then draw a median line that will represent the most likely value. It is sometimes called "linear model". This model also works if there is more than one input (multiple input variables). This median line (in the case of a single input) provides a coefficient that will enable the user to create a data matrix that will return the optimal values for coefficients. When dealing with multiple inputs, you need to optimize the coefficients values. This is done by a process called gradient descent whereby you iterate through the values trying to minimize the error of the training data (or reference data). There exist also regularisation procedures to help stabilise the model and improve output accuracy. This is an excellent approach for supervised learning as this enables the user to forecast or predict outcome based on input.

In case of working with categorisation (A - !A), it is often wise to sellect a different algorithm called logistic regression. As for the linear regression, the input (on the x axis) will be allocated to an output (on the y axis), however, the y axis will only contain a given set of values. Each input will fall on one of those values. As such, you can have a simple binary category (A - !A) as well as multiple categories (multi-linear functions). In this case, the curve, instead of looking like an "S", looks like rounded stairs.

Sometimes, it is simpler and more effective to work with a decision tree. Contrary to the two previous algorithm, the decision tree will not be able to be represented visually by a graph, it will have to be a workflow (starting from the trunk) that will divide into branches based on the data labels. This is possible the most common model in supervised learning as each decision node leads to either terminals nodes or other decision nodes. Each decision node leads to a split in the outcome. To create a decision tree, the decision nodes must have a specific order (from generic to specific). Some decision nodes will require statistics to provide an outcome (provide a decision). The aim of this approach is to reach a high accuracy. Each attributes must be represented by a decision node. For this purpose several mesures are used:

1. Entropy: mesure the randomness of an event. The higher, the harder to draw a conclusion
2. Information gain: the level of how well this separate the data. This is based on entropy before the split and entropy after the split. It tends to prefer larger sets of different values.
3. Gini index: a function that favours larger dataset splits and gives a "A - !A" results.
4. Gain ratio: compensate for the information gain bias by taking the number of split into account.
5. Reduction in variance: use a standard variance to find the split with the lowest variance.

One of the key issue with the decision tree is that it will always have a value for the training data which might not be accurate when dealing with external data. This means that the model must find a way to compensate for this issue. One approach is called pruning.

This aims at reducing the number of branches (decision nodes) without affecting the outcome. For that purpose, the training data is split into 2 sections, a training part and a validation part. The validation data is used to reduce the number of decision nodes when those decisions do not lead to any improvement in the accuracy.

In some model, the pruning is replaced by the "random forest". In this instance, a random subset of the training data is used to build the tree and a random subset of attributes are picked to create the decision nodes. Bagging is then used to create a set of trees that will give combined predictions that can be summed or averaged. This approach often lead to extremely adaptable and accurate supervised machine learning models.