Decision Tree algorithms in Machine Learning

Mrs.N.Anuradha, Assistant Professor & Head,

Dept of CS(SSS), Subbalakshmi Lakshmipathy College of Science, Madurai

Introduction

Machine Learning is a set of techniques to extract knowledge from available data and use that knowledge to make decisions. Machine learning involves making machines learn things like humans do. It involves several stages to take a decision, (ie) Data Pre-processing, Training the model, Inference. Many algorithms are available to prepare the dataset, train the model and to give the inference. Some algorithms are Linear Regression , Logical Regression, Decision Tree, Random Forest, Support Vector Machine(SVM) etc.

Decision Tree

The Machine Learning techniques are classifed as Supervised Learning and Unsupervised Learning. In a supervised learning model, the algorithm is given a labeled dataset and answer key. The algorithm can evaluate its accuracy on training data. An unsupervised model, in contrast, the algorithm is provided unlabeled data and the algorithm tries to make sense of by extracting features and patterns on its own.

The supervised Learning model uses two techniques such as Regression and Classification. The decision tree algorithm follows the supervised learning method. The decision trees are known as eager learners, because it learn the data set quickly by segregating.

Classifications of Decision tree algorithms

There are several decision tree algorithms available to train the data set and find the insights from the given data set. They are:

- ID3
- C4.5
- CART etc.

Comparison of various decision trees

S.No	Features	ID3	C4.5	CART
1	Types of data	Categorical	Continues & Categorical	Continues and Nominal
2	Speed	Low	Faster than ID3	Average
3	Boosting	Not Supported	Not Supported	Supported
4	Pruning	No	Pre-Pruning	Post-Pruning
5	Missing values	Can't deal with	Can't deal with	Can deal with
6	Formula	Use Information entropy and Information Gain	Use split info and gain ratio	Use Gini diversity index

Advantages of Decision trees

Decision trees are easy to understand, interpret and visualize. They implicitly perform variable screening or feature selection. They can handle numerical and categorical data. It can also handle multi-output problems (ie) multi-class or multi-regression problems. It require relatively little effort from users for data preparation.

Disadvantages of Decision trees

It may create over-complex trees that do not generalize the data well, which will lead to overfitting problem. It may be unstable because small variations in the data might result in a completely different tree being generated. This is called variance, which needs to be lowered by methods like bagging and boosting. Greedy algorithms can not guarantee to return the globally optimal decision tree. This can be mitigated by training multiple trees, where the features and samples are randomly sampled with replacement. Sometimes, it may create biased trees if some classes dominate. It is therefore recommended to balance the data set prior to fitting with the decision tree. The up-sampling and down-sampling techniques are used to solve this problem.

Applications of Decision tree

The decision trees are used in the following applications:

- Business Management
- Customer Relationship Management
- Fraudulent Statement Detection
- Energy consumption
- Fault Diagnosis
- Health care managment

Conclusion

Though there are several algorithms like linear regression, logical regression, Support Vector Machine, etc are available to train the data set model, the decision trees are used in many applications . The Random forest Algorithm is the extended version of decision tree.