JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue
JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)
An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Decision Tree Algorithm for Data Science

Ram Dulari (Student)
Department of Computer Science
D.P.G. Institute of Technology and Management, Gurgaon 122001
Gurgaon, India
erbhardwajpriya@gmail.com

Richa Nehra (guide)
Deptt. of Computer Science & Engineering, DPG
Institute of Technology & Management
richa.dpgitm@gmail.com

ABOUT THE AUTHORS

Ram Dulari is a student of Computer Science & Engineering at DPG Institute of Technology and Management. She is currently doing her master's from this Institute and Degree is about to complete in Jun 2021. Her research area is in Artificial Intelligence, Data Science. Before her master's she is having a bachelor's degree in computer science and Engineering. And moreover, a technical software industry experience. Two years experience with NSDC Council in agriculture and food sector as a Coordinator (Government).

Ms Richa Nehra is an assistant professor of CSE Department DPG Institute of Technology and Management. Her research includes in software Engineering and Data Analysis, etc. She has published paper widely; She has 3 year of teaching and research experience. Her interest is in data fields. Before her master's degree she is having a bachelor's degree (B.Tech) in Computer Science and Engineering Department. She has an excellent academic record.

ABSTRACT

A decision tree is a tree whose internal nodes can be taken as tests (on input data patterns) and whose leaf nodes can be taken as categories (of these patterns). These tests are filtered down through the tree to get the right output to the input pattern. Decision Tree algorithms can be applied and used in various different fields. It can be used as a replacement for statistical procedures to find data, to extract text, to find missing data in a class, to improve search engines and it also finds various applications in medical fields. Many Decision tree algorithms have been formulated. They have different accuracy and cost effectiveness. It is also very important for us to know which algorithm is best to use. The ID3 is one of the oldest Decision tree algorithms. It is very useful while making simple decision trees but as the complications increases its accuracy to make good Decision trees decreases. Hence IDA (intelligent decision tree algorithm) and C4.5 algorithms have been formulated.

Key words : Decision trees, Feature selection, Information gain, Keyword advertising.Decision Tree

The goal of this algorithm is to create a model that predicts the value of a target variable, for which the decision tree uses the tree representation to solve the problem in which the leaf node corresponds to a class label and attributes are represented on the internal node of the tree. ...

Decision Tree is an algorithm used for supervised learning problems such as segregation or reversal. Each leaf of a tree has a label with a section or distribution of opportunities over

classes. The tree can be "read" by dividing the source set into the subsets based on the value assessment

The Decision Tree algorithm belongs to a family of supervised learning algorithms. The goal is to use the Decision Tree to create a training model that can be used to predict a category or number of target variations by studying the rules of simple decisions determined from previous data (training data).

Solution to data science:

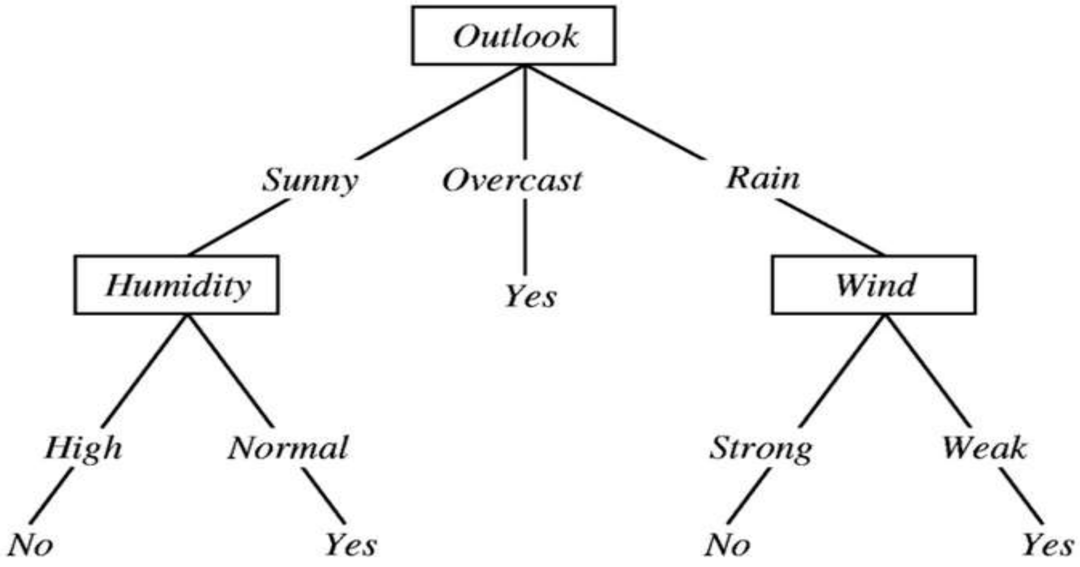
Decision tree is a flowchart-like structure in which each internal node represents a test element (e.g. they represent a combination of elements that lead to that phase.

- Decision trees are used to manage incompatible data sets.
- The drug decision tool is used in real life in many areas, such as engineering, community planning, law and business.
- Decision trees can be divided into two types; flexible and continuous evergreen trees.

Decision tree is a flowchart-like structure in which each internal node represents a test element (eg represents the combination of elements that lead to those class labels. The routes from root to column represent the rules of separation.), No Rain (No)).

Decision tree uses the tree representation to solve the problem in which each leaf node corresponds to a class label and attributes are represented on the internal node of the tree. We can represent any boolean function on discrete attributes using the decision tree.

Decision Trees are a non-parametric supervised learning method used for both classification and regression tasks. Tree models where the target variable can take a discrete set of values are called classification trees



Decision Tree for Rain Forecasting

Decision trees are constructed in an algorithmic way that identifies ways to separate data from different scenarios. It is one of the most widely used and practical methods of surveyed learning. Tree choices are a paramount study of the parameter that is used for partition and subtraction activities.

Types of trees where the target variant can take a set of values called classified trees. Decision trees where the target variable can take continuous values (usually real numbers) are called decentralized trees. Separation from the Pressure Tree (TIME) is a common term for this.

In all of this post I will try to explain using examples.

Data Format

The data comes in form records.

(x, Y) = (x1, x2, x3,., xk, Y)

The variation that depends on it, Y, is the target variable that we are trying to understand, divide or integrate. Vector x is made up of elements, x1, x2, x3 etc., which are used for that function.

Steps to Making a Decision

- Obtain a list of lines (data) for consideration to make a decision-making solution (repeated for each node).
- Calculate the uncertainty of our data or Gini pollution or how mixed our data is etc.
- Generate a list of all the questions that need to be asked in that node.
- Divide lines into true lines and false lines based on each question asked.
- Calculate data gain based on gin contamination and separate data from the previous step.
- Update the highest information based on each question asked.
- Revive an excellent question based on knowledge acquisition (high knowledge acquisition).
- Divide the node into the best query. Repeat from step 1 and until we find a pure node (leaf nodes).

Types of trees where the target variant can take a set of values called classified trees. Decision trees where the target variable can take continuous values (usually real numbers) are called decentralized trees. Separation from the Pressure Tree (TIME) is a common term for this.

Benefits of Resolution

- Easy to use and understand.
- Able to manage sector and numerical data.
- Resistant retailers, which is why it requires less data processing.

Disadvantages of Decision Tree

- You need to be aware of parameter adjustment.
- You can make educated trees that discriminate against them if you control other categories.

How can you avoid overreacting to the decision tree model

Overuse is one of the biggest problems for every model in machine learning. If the model is not skipped it will be badly built on new samples. Avoiding the tree that decides in excessive use removes branches that use low-value features. This method is called pruning or pruning. In this way we will reduce the weight of the tree, which is why we improve the accuracy of speculation by reducing overgrowth.

Pruning should reduce the size of the reading tree without reducing the accuracy of the prognosis as measured by the crossing confirmation set. There are two main ways to pruning.

- Minor error: The tree is pruned back when the shortcut confirmed error is small.
- Smallest Tree: The tree is pushed back slowly over a small mistake. Technically the pruning creates a decision-making tree with a verification error between 1 common error and a minor error.

The Decision Tree in Real Life Choosing a plane to fly

Suppose you need to choose your next flight. How do we do that? We first check whether the flight is available that day or not. If it is not available, we will look at another date but if there is one we will be looking at flight time. If we want to have direct flights only then we look at whether the price of that plane is in your pre-defined budget or not. If it costs too much, we look at other airlines and subscribe to them!

The conclusion

Conclusion Decision trees assist **analysts** in evaluating upcoming choices. The tree creates a visual representation of all possible outcomes, rewards and follow-up decisions in one document

As the goal of a **decision tree** is that it makes the optimal choice at the end of each node it needs an algorithm that is capable of doing just that. That algorithm is known as Hunt's algorithm, which is both greedy, and recursive.

Finally, when it comes to building your own learning machine models, look at the various language development options, IDEs and Platforms. The next thing you need to do is start learning and mastering each machine learning process. The title is great, it means there is a wide range, but if you look at the depth, each article can be read in a few hours. Each topic is unique. You need to look at one topic at a time, read it, create it and apply algorithm / s to it using the language of your choice. This is a great way to start studying Learning Machines. Familiarizing yourself with one topic at a time, you will soon find the required range at the end of a typewriter.

References

[1] Leo Breiman, Random Forests. Machine Learning, Volume 45, Issue 1, October 1 2001, Pages 5-32

[2] Jerome H. Friedman, Greedy Function Approximation: A Gradient Boosting Machine, The Annals of Statistics, Vol. 29, No. 5 (Oct., 2001), pp. 1189-1232

[3] Trevor Hastie, Robert Tibshirani, Jerome Friedman, The Elements of Statistical Learning: Data mining, inference, and prediction. Second Edition (2009), Springer Series in Statistics, Pages 219-230

[4] Georg Heinze and Michael Schemper, A solution to the problem of separation in logistic regression, Statistics in Medicine, Volume 21, Issue 16 (30 August 2002), Pages 2409–2419.

[5] Andrew Gelman et al, A weakly informative default prior distribution for logistic and other regression models, The Annals of Applied Statistics 2008, Vol. 2, No. 4, Pages 1360– 1383

[6] Lawrence R. Rabiner, A tutorial on Hidden Markov Models and selected applications in speech recognition. Proceedings of the IEEE, Vol. 77, No. 2. (06 February 1989), Pages. 257- 286

[7] Charles C Holt, Forecasting Trends and Seasonal by Exponentially Weighted Averages. International Journal of Forecasting, Volume 20, Issue 1 (January–March 2004), Pages 5–10.

[8] M. Xie et al, A seasonal ARIMA model with exogenous variables for Elspot electricity prices in Sweden, 2013 10th International Conference on the European Energy Market (EEM), Stockholm (May 2013), Pages 1-4

[9] Cristina Stolojescu-Crisan, Data mining based wireless network traffic forecasting, 2012 10th International Symposium on Electronics and Telecommunications (ISETC), Timisoara (Nov 2012), Pages 115-118.

[10] Josef Kittler et al, On combining classifiers. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 20, No. 3 (March 1998), Pages 226-239.

[11] Papadopouli M, Evaluation of short-term traffic forecasting algorithms in wireless networks, 2006 2nd Conference on Next Generation Internet Design and Engineering, NGI,