A beginner's guide to decision tree classification
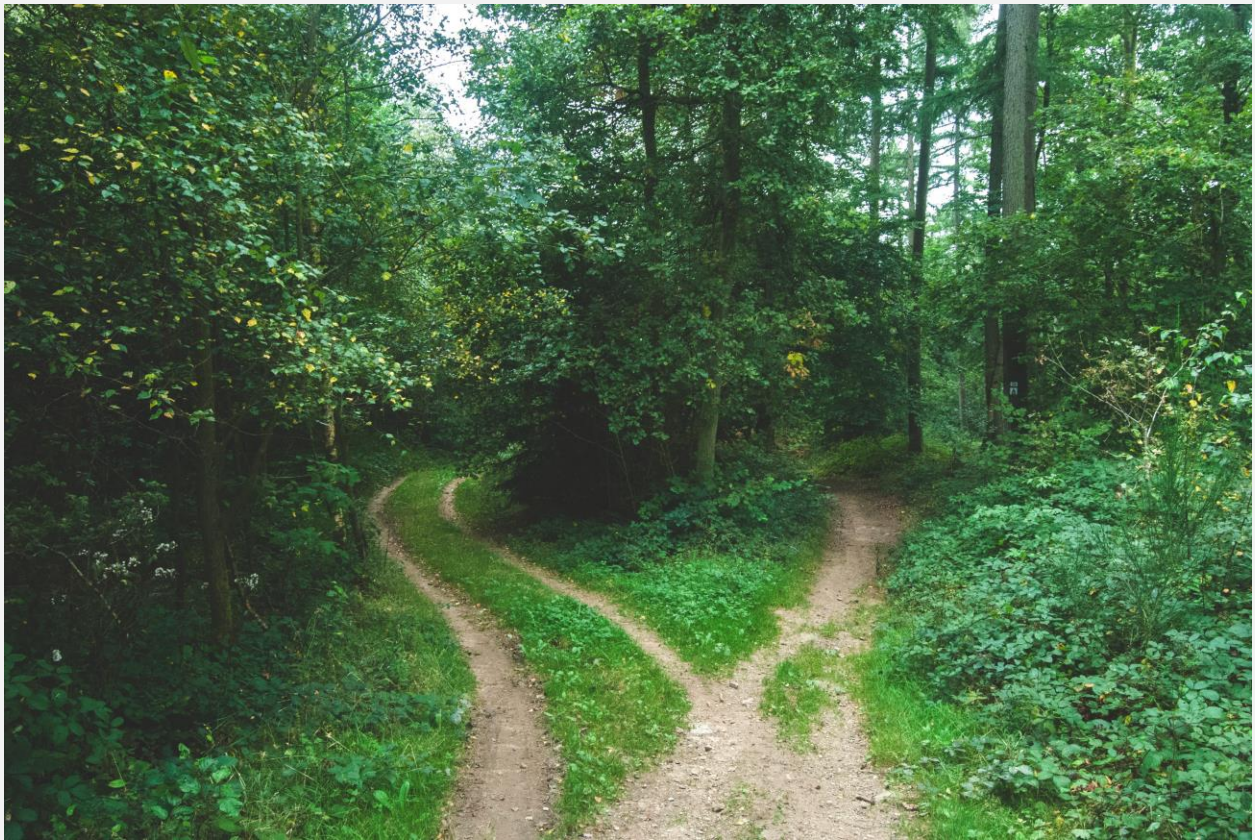
Charlie Bickerton

Follow

Decision trees are one of the most popular machine learning algorithms but also the most powerful. This article is going to explain how they work from a non-technical perspective.

One of the reasons they are so powerful is because they can be easily visualised so that a human can understand whats going on. Imagine a flowchart, where each level is a question with a yes or no answer. Eventually an answer will give you a solution to the initial problem. That is a decision tree. Everybody subconsciously uses decision trees all the time for most menial tasks. Decision trees in machine learning take that ability and multiply it to be able to artificially perform complex decision making tasks.

A lot of the time it can be very difficult to understand how a machine learning algorithm comes to its decision, making them unusable for many scenarios. This is especially true for cases where the decision might need to be contested, such as in the criminal justice system, health industry and for strategic business decisions. This is even more of a factor for customer-facing machine learning algorithms now that customers have a "right to an explanation" under GDPR. A decision tree's ability for human comprehension is a major advantage.

The decision tree analyses a data set in order to construct a set of rules, or questions, which are used to predict a class. Let us consider a dataset consisting of lots of different animals and some of their characteristics. These characteristics can be used to predict their class. If we take an eagle and an elephant, a question that would split these two animals would be 'Does this animal have two legs?' or perhaps 'Does this animal weight under

500kg?'. The answer no for either of these questions would lead to the classification of an elephant, whereas yes would be an eagle.

These rules can be built up to create a model that can classify complex situations. To extend the animal classification example, consider the scenario of needing to classify a selection of animals into mammals, birds, fish. Look at the visualised decision tree to see how two simple questions can be used to split the data.
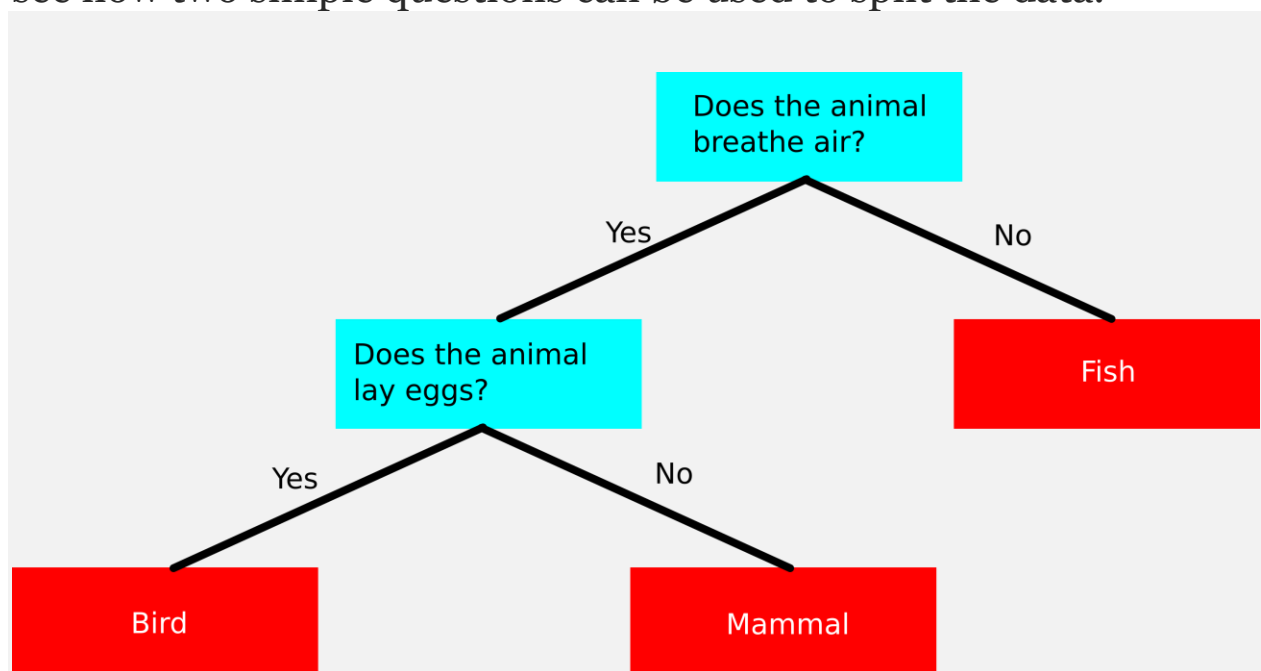
These simple questions layered one after another, allow the classification of a wide range of animals. This is the power of decision trees. Now if we give the trained decision tree a new animal, for example a dog, it will classify it. Does a dog breath air? Yes. Does a dog lay eggs? No. Therefore the model will classify it as a mammal, the correct answer!

When a human constructs a decision tree, the questions and answers are based off their logic and knowledge. In data science the creation of these rules is usually governed by an algorithm learning which questions to ask by analysing the entire data set. To put this into context we will come back to the animal example, the algorithm will look at all the animals to figure out that all animals that don't breath air and notice that they are all fish. This mathematically splits the dataset by it's class. This creates powerful algorithms that can classify new data into classes in a way that any human can understand.

Decision trees can become much more powerful when used as ensembles. Ensembles are clever ways of combining decision trees to create a more powerful model. These ensembles create state of the art machine learning algorithms that can outperform neural networks in some cases. The two most popular ensemble techniques are random forests and gradient boosting.

Thank you for reading. Please checkout my next article on [a beginners guide to deep learning](). Or if you are interested further articles on a beginner's guide to data science or applying AI to business please visit [Miminal]().

Towards Data Science

A Medium publication sharing concepts, ideas, and codes.
Follow

- Machine Learning
- Data Science
- Artificial Intelligence
- Computer Science
- Algorithms

108 claps