Decision Trees

Decision trees are a supervised, probabilistic, machine learning classifier that are often used as decision support tools. Like any other classifier, they are capable of predicting the label of a sample, and the way they do this is by examining the probabilistic outcomes of your samples' features. Decision trees are named after their tree graph structure, although perhaps a better name for them is "decision roots", unless they're supposed to look like an upside down tree.

Before even describing decision trees, visit this New York Times page from 2012 that beautiful demonstrated an interactive one, based on electoral votes within the U.S. for determining the 'then' presidential election.

Decision trees search for a structured set of rules to help the rapid classification of your dataset. Imagine coming home from work to see your favorite vase laying broken on the floor. What caused it? Did someone break it? To find out what happened, you'd have to look for some clues. If no one else in the family came home before you, then chances are, it was a thief or earthquake. If anything else is missing from the house then, perhaps you had an intruder. If other items fell from high shelves, and the news reports an earthquake, then perhaps that's what ruined your vase.

If neither of these is the answer, then you have to keep digging. Were family members around at the time of the breakage? What else is physically placed near the vase? You notice that there is a comic book, and some cat food. You can't talk to the cat and get back a coherent response, so you ask your son Timmy instead. He says he had nothing to do with it, but you notice he has chocolate chip cookie

Dive Deeper

stains on his hands, mouth, and face. The only chocolate chip cookies in the house are in the cookie jar that was located right next to the vase. It's looking pretty certain Timmy is the culprit now. Had he not had cookie residue on his face and hands, then perhaps it might have been Neko the cat.

Each question asked by your decision trees helps to remove some ambiguity about the problem you're trying to solve (who broke the vase). The better the question you ask, the more rapidly you're able to find out the answer to your question. Decision trees have a few key attributes that make them unique:

- The decision tree is structured as a flowchart diagram. It has a single, starting root node, and can contain one or more leaf nodes, and internal nodes.
- Each node of the decision tree represents the test of single feature's value. For example, one node might test the feature named 'age' to see if the user is greater than 18 or not.
- Each branch of the decision tree connects two nodes, an originating node, and a destination node. A branch represents an outcome of the test of the branch's originating node. In the case of the age > 18 test, there are two possible branches: True and False. Each of these two branches would connect the age > 18 node, to two destination nodes.
- Each leaf node of the decision tree represents a classification. The end goal of the tree is to label your input samples, so each leaf fulfills that goal.