An Intuitive Introduction to Decision Trees

June 26, 2020 Piyush & Rishabh (https://www.newtechdojo.com/author/piyush/)
Machine Learning (https://www.newtechdojo.com/category/machine-learning/)

(https://www.newtechdojo.com/intuitive-introduction-decision-trees/)

Decision Trees are one of the most powerful yet easy to understand machine learning algorithm. It lets the practitioner ask a series of questions helping her decide to choose between multiple alternatives at hand. Decision trees are ubiquitous in day-to-day life. We use them daily knowingly or unknowingly. Its algorithm assumes that the data follows a set of rules. These rules are identified by using various mathematical techniques. Decision Trees find their application in both the Classification (This or That) and the Regression (How much of This?) settings.

In this article, I will just introduce a basic decision tree, its intuition, its various elements, and techniques of building a tree. For starters, it must be noted that a decision tree is similar to a flowchart. We come across these charts almost every day in offices but with a decision at the end of it.

> *"Decision Trees are everywhere."*

An intuition into Decision Trees

Suppose you are out to buy a new laptop for yourself.

After reaching a shop, you are confused about which one to buy among so many options. So, you asked the shopkeeper to help you decide. The shopkeeper then asks you a series of questions. These questions help you decide which laptop to buy.

Q1: "How much storage space are you looking for?"

"Umm... around 1 TB space, preferably with an HDD."

Q2: "Perfect, and how about the RAM?"

"Definitely 8 GB or more."

Q3: "Alright, and any preferences on the Graphics Card?"

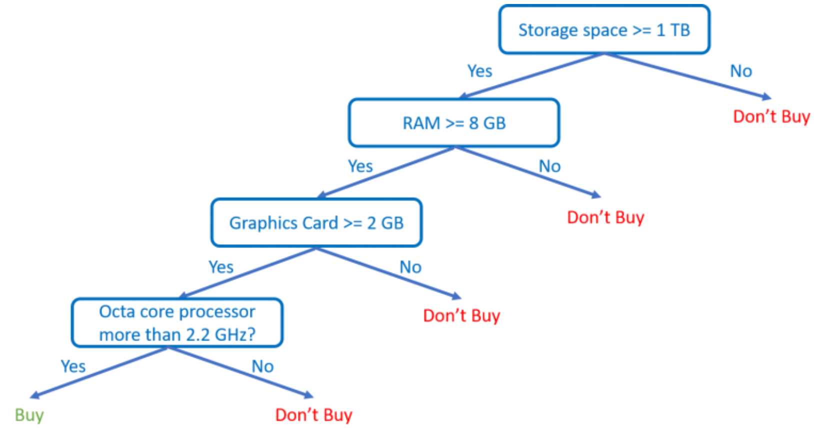
"I want at least 2 GB GPU."

Q4: "Alright, and any preferences on the Processor?"

"I want an octa-core processor with at least 2.2 GHz speed."

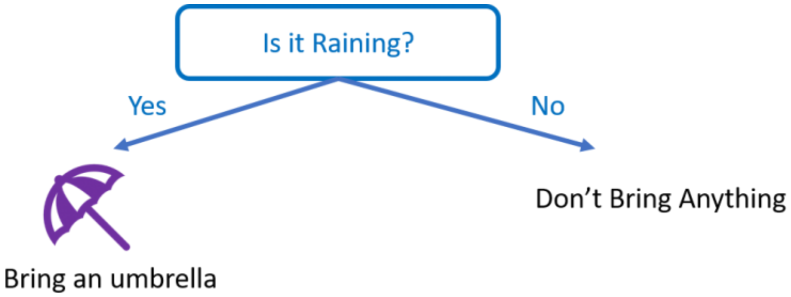
"Sure! I've got the perfect laptop for you." And he hands over a laptop to you.

What the shopkeeper just did was to help you walk through alternatives to narrow down your choices. Pictorially, we can represent this process as:

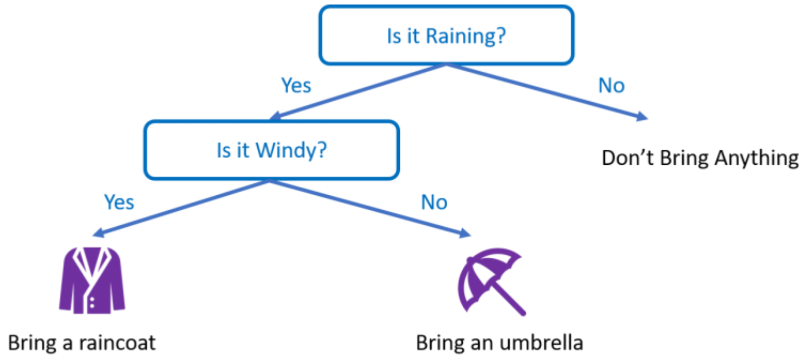


This figure corresponds to a decision-making process. This structure is called a **Decision Tree**.

These decision trees can be built for almost any decision-making in day-to-day life. For example, imagine you want to pick between an umbrella, raincoat, or nothing while going out on a rainy day. This process can be described as:

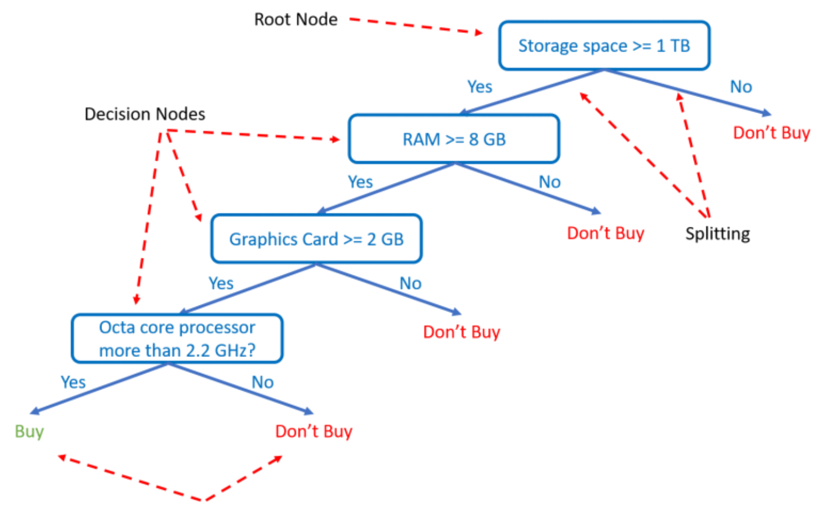


Now suppose we say, "if it is windy outside, I'll use a raincoat; otherwise, I'll use an umbrella." This statement adds a little detail to our tree in this way:



Tree Diagram

In the above section, we dealt with what a decision tree is and how we can build one with a simple intuitive process. Let us now discuss the different elements of a decision tree with our previous case.



In this Decision Tree diagram, we have:

1. Node: This is where we either ask a question or make a decision. These are on the ends of pointy arrows. For our case, Storage space, RAM, GPU, Processor, Buy and Don’t buy all of these are individual nodes.
2. Root Node: This is the place where the first separation takes place. In our case, the question about the Storage space of the laptops forms our root node.
3. Splitting: It is the process of dividing any node into two or more nodes. In our case, every question resulted in two splits, one that asks more questions and the other decides that we don’t want to buy.
4. Decision Node: If, after any split, a resulting node asks another question, then the resulting node is called the decision node. Here we have, RAM, GPU, and processor node as decision nodes.
5. Leaf: If, after any split, a resulting node outputs a decision (categorical or continuous value), it is called the leaf node. This node doesn’t ask further questions. In this example, Don’t buy and Buy nodes are leaf nodes.

Asking the Right Questions

So far, we have seen how we can build a simple decision tree and what its different elements are. But while creating a decision tree, it is crucial to ask the right questions at the correct stage in a tree. That is what essentially building a decision tree or decision tree learning means. Asking irrelevant questions can lead to complications in our problem.

For example: Imagine during the laptop purchase case, the shopkeeper asked if you wanted a laptop with or without a Graphics Card. Would that lead to a narrowing down of your options?

Mathematically, we have two commonly used techniques to determine what will be the best question to ask at any stage:

1. Entropy
2. Gini Index

These techniques help us decide what, when & where to start and stop asking questions. These techniques are popularly called the splitting criteria. I'll give a brief description of these concepts in the subsequent sections.

Entropy and Information Gain

Imagine the shopkeeper did not keep an ordered shelf, in our laptop purchase case. In other words, what if the shop was unordered with all quality of laptops on a single display. Would that make your decision making easier?

> *"Entropy is an indicator of how messy your data is."*

For a dataset, messiness corresponds to a mixture of available options (*target variable*).

In the decision tree learning, our goal is to separate this mixture. Entropy lets us decide between the right questions to ask to separate the desired outcome from all available options.

Higher the entropy of a dataset, the higher the degree of mixing, while lower entropy corresponds to a well-separated data.

Entropy ↑ Extent of Mixing in Data ↑ Entropy ↓ Separation in Data ↑

Once we have asked the right questions, we have narrowed down our options and know what not to choose from. Instead, we have more information about where to find the answer. For example, knowing that we needed a laptop with HDD >= 1 TB made sure that this is the bare minimum. We don't desire laptops that do not possess this quality.

This phenomenon of finding a desirable direction for our exploration is called as Information Gain. Entropy helps in calculating this gain numerically. We will skip these numerical parts in the article.

Gini Index

Similar to entropy, the Gini Index also helps us decide the right set of questions to ask. But instead of measuring the messiness of a dataset, it measures its impurity.

> *"Gini Index is the measure of how impure your data is"*

For a dataset, impurity corresponds to a mixture of decisions (target variable). If the dataset after a particular splitting remains mixed with all available options to choose from, we have impurity in the data. If we have reached a decision, it implies that we have data that is pure in terms of the options that we have.

While building a decision tree, we try to find out the series of questions that lead to the maximum decrease in the impurity of the dataset.

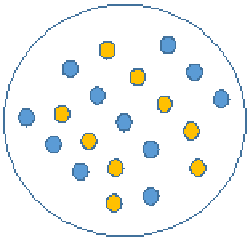
Higher Gini Index corresponds to a mixture (impure) while lower corresponds to separated data.

Gini Index ↑ Impurity/Mixture in Data ↑ Gini Index ↓ Purity in Data ↑

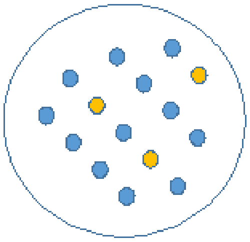
We can use either the Gini Index or the Entropy to build a decision tree.

Both these concepts can be quite confusing to grasp at first. So, let us do a simple exercise to understand these better.

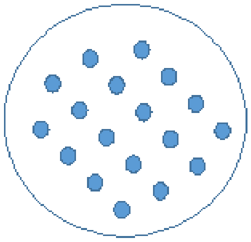
Suppose you have a bag of blue and yellow balls. Your target is to separate these two in different packs. Given below is a diagram of 3 such bags with a different combination of yellow and blue balls. Find out which one has:



A



B



C

1. Maximum Entropy
2. Minimum Entropy
3. Maximum Gini impurity
4. Minimum Gini impurity

Advantages and Disadvantages of using a Decision Tree:

Advantages:

1. Decision trees can be used both in Regression and in Classification settings.
2. No parameter required to know before the model building
3. Interpretability of decision trees is easy
4. Fast to build a decision tree
5. No scaling of features is needed to fit a decision tree

Disadvantages:

1. Very high chances of Overfitting if we keep on splitting
2. Decision trees are optimized at every split, which can sometimes lead to a wrong result.
3. Low usages for imbalanced dataset.

Answer to above questions: A, C, A, C respectively

Tags: **decision tree (https://www.newtechdojo.com/tag/decision-tree/)**
Entropy (https://www.newtechdojo.com/tag/entropy/) **Gini Index (https://www.newtechdojo.com/tag/gini-index/)**
Information Gain (https://www.newtechdojo.com/tag/information-gain/)
machine learning (https://www.newtechdojo.com/tag/machine-learning/)

Most People Like This Blog

(https://www.newtechdojo.com/modern-deep-learning-python/)

🕓 January 9, 2017 (https://www.newtechdojo.com/modern-deep-learning-python/) 🗀 Machine Learning (https://www.newtechdojo.com/category/machine-learning/)

Modern Deep Learning in Python (https://www.newtechdojo.com/modern-deep-learning-python/)

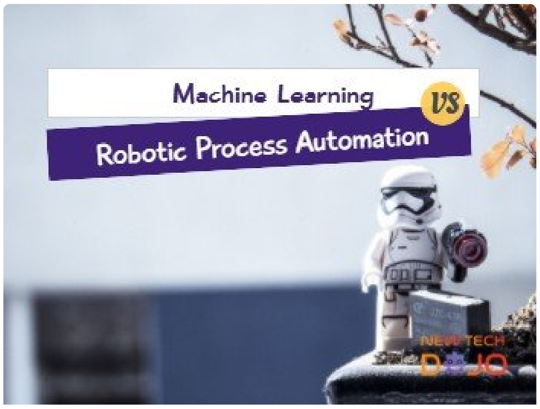
Read more (https://www.newtechdojo.com/modern-deep-learning-python/)

(https://www.newtechdojo.com/tensorflow-for-deep-learning-with-python/)

🕓 January 9, 2017 (https://www.newtechdojo.com/tensorflow-for-deep-learning-with-python/) 🗀 Machine Learning (https://www.newtechdojo.com/category/machine-learning/)

Complete Guide to TensorFlow for Deep Learning with Python (https://www.newtechdojo.com/tensorflow-for-deep-learning-with-python/)

Read more (https://www.newtechdojo.com/tensorflow-for-deep-learning-with-python/)



(https://www.newtechdojo.com/machine-learning-vs-robotic-process-automation/)

January 3, 2018 (https://www.newtechdojo.com/machine-learning-vs-robotic-process-automation/) Machine Learning (https://www.newtechdojo.com/category/machine-learning/)

Machine Learning vs Robotic Process Automation (https://www.newtechdojo.com/machine-learning-vs-robotic-process-automation/)

Read more (https://www.newtechdojo.com/machine-learning-vs-robotic-process-automation/)

NewTechDojo is an on-demand marketplace to learn from the Best and experienced industry Experts. Get trained from the Top Data Science consultants and Programmers. Take this opportunity, explore your career in Data Science and learn from the skilled and upbeat Mentors.

Tags

accuracy (https://www.newtechdojo.com/tag/accuracy/)

Artificial Intelligence (https://www.newtechdojo.com/tag/artificial-intelligence/)

classification (https://www.newtechdojo.com/tag/classification/) clustering (https://www.newtechdojo.com/tag/clustering/)

Coca Cola (https://www.newtechdojo.com/tag/coca-cola/) Computer Vision (https://www.newtechdojo.com/tag/computer-vision/)

Confusion matrix (https://www.newtechdojo.com/tag/confusion-matrix/)

continuous (https://www.newtechdojo.com/tag/continuous/) decision tree (https://www.newtechdojo.com/tag/decision-tree/)

deep learning using Python (https://www.newtechdojo.com/tag/deep-learning-using-python/)

discrete (https://www.newtechdojo.com/tag/discrete/) Entropy (https://www.newtechdojo.com/tag/entropy/)

F1 score (https://www.newtechdojo.com/tag/f1-score/) Facebook (https://www.newtechdojo.com/tag/facebook/)

Gini Index (https://www.newtechdojo.com/tag/gini-index/)

GPU for Machine Learning (https://www.newtechdojo.com/tag/gpu-for-machine-learning/)

histograms (https://www.newtechdojo.com/tag/histograms/) IBM Watson (https://www.newtechdojo.com/tag/ibm-watson/)

Information Gain (https://www.newtechdojo.com/tag/information-gain/)

Kinds of machine learning (https://www.newtechdojo.com/tag/kinds-of-machine-learning/)

KNN (https://www.newtechdojo.com/tag/knn/) knn algorithm (https://www.newtechdojo.com/tag/knn-algorithm/)

knn algorithm example (https://www.newtechdojo.com/tag/knn-algorithm-example/)

knn algorithm in ml (https://www.newtechdojo.com/tag/knn-algorithm-in-ml/)

knn classification (https://www.newtechdojo.com/tag/knn-classification/)

linear regression (https://www.newtechdojo.com/tag/linear-regression/)

machine learning (https://www.newtechdojo.com/tag/machine-learning/)

Machine Learning Algorithms (https://www.newtechdojo.com/tag/machine-learning-algorithms/)

machine learning with tensorflow (https://www.newtechdojo.com/tag/machine-learning-with-tensorflow/)

Precision (https://www.newtechdojo.com/tag/precision/) Probability (https://www.newtechdojo.com/tag/probability/)

probability distribution (https://www.newtechdojo.com/tag/probability-distribution/)

random forest (https://www.newtechdojo.com/tag/random-forest/) Recall (https://www.newtechdojo.com/tag/recall/)

reinforcement learning (https://www.newtechdojo.com/tag/reinforcement-learning/)

Specificity (https://www.newtechdojo.com/tag/specificity/)

supervised learning (https://www.newtechdojo.com/tag/supervised-learning/) SVM (https://www.newtechdojo.com/tag/svm/)

Ted Talks on Machine Learning (https://www.newtechdojo.com/tag/ted-talks-on-machine-learning/)

tensorflow course (https://www.newtechdojo.com/tag/tensorflow-course/)

tensorflow deep learning (https://www.newtechdojo.com/tag/tensorflow-deep-learning/)

tensorflow tutorial (https://www.newtechdojo.com/tag/tensorflow-tutorial/)

types of machine learning (https://www.newtechdojo.com/tag/types-of-machine-learning/)

unsupervised learning (https://www.newtechdojo.com/tag/unsupervised-learning/)

YouTube (https://www.newtechdojo.com/tag/youtube/)