Introduction to Decision Trees

Nicole Donnelly

LEARNING OBJECTIVES

- Define decision trees
- Determine when a decision tree is appropriate

DATA SCIENCE WORKFLOW

DATA SCIENCE WORKFLOW

BUILD A DATA MODEL

- ☐ Select appropriate model
- ☐ Build model
- ☐ Evaluate and refine model

Decision Trees are machine learning algorithms for classification and regression. They make decisions by asking a series of questions to split data into homogenous sets.

Supervised	Non-parametric	Hierarchical

Decision Trees are machine learning algorithms for classification and regression. They make decisions by asking a series of questions to split data into homogenous sets.

Supervised	Non-parametric	Hierarchical
We model our data with known target values.		

Decision Trees are machine learning algorithms for classification and regression. They make decisions by asking a series of questions to split data into homogenous sets.

Supervised	Non-parametric	Hierarchical
We model our data with known target values.	We start with no assumed parameters such as distribution or error. We do not have coefficients to tune.	

Decision Trees are machine learning algorithms for classification and regression. They make decisions by asking a series of questions to split data into homogenous sets.

Supervised

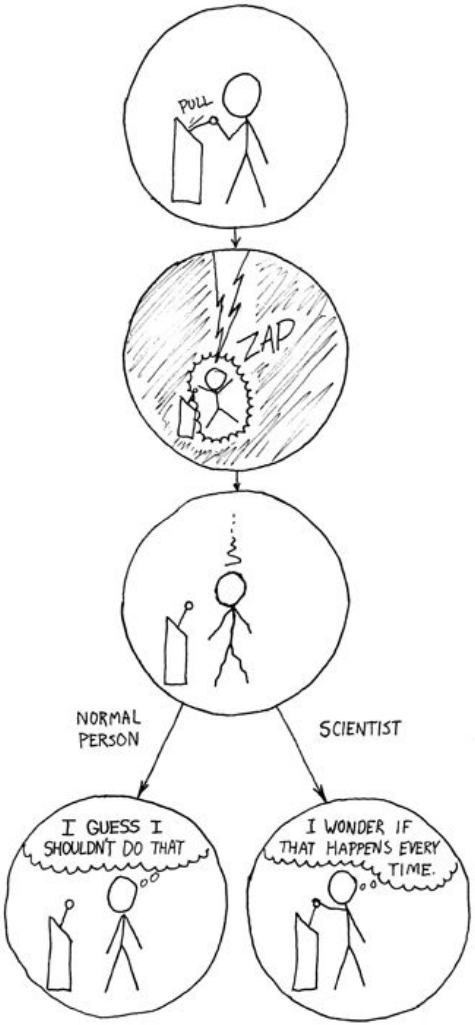
We model our data with known target values.

Non-parametric

We start with no assumed parameters such as distribution or error. We do not have coefficients to tune.

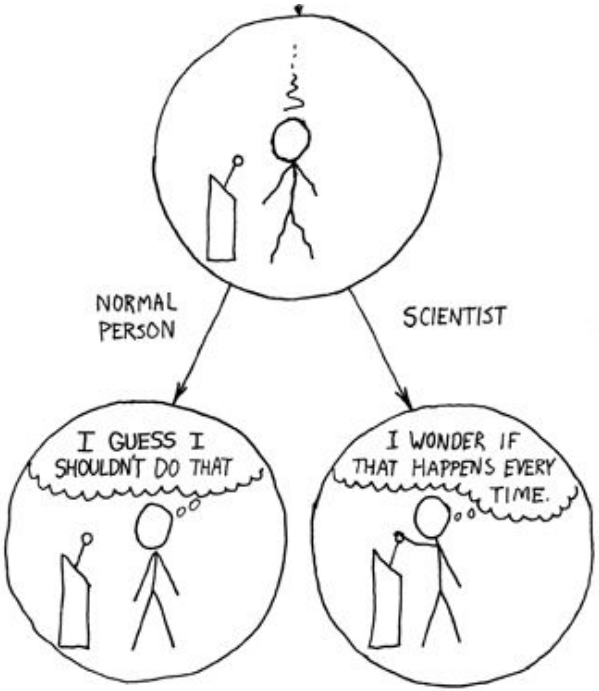
Hierarchical

We are predicting a target value via recursive splits in a top-down fashion. This essentially acts like a series of if-then statements.



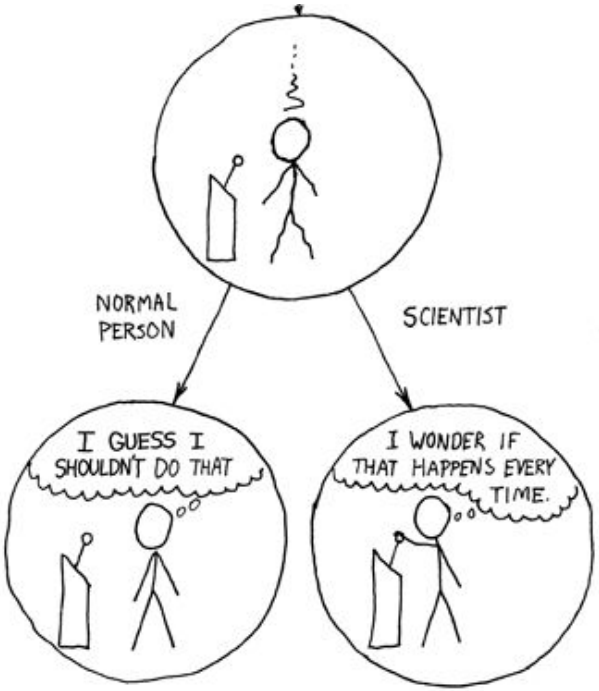
IS THIS A SCIENTIST?

XKCD http://xkcd.com/242/



IS THIS A SCIENTIST?

XKCD http://xkcd.com/242/



IS THIS A SCIENTIST?

NAME	ZAPPED	PULL_AGAIN	SCIENTIST
Nicole	Yes	Yes	1
Cow Girl	Yes	No	0
Hugo	Yes	No	0
Vince	Yes	Yes	1

XKCD http://xkcd.com/242/

Decision Tree Terminology

ROOT Node: The entire population or sample, which is divided into two or more homogeneous sets

Splitting: Dividing a node into two or more sub-nodes

Decision Node: A node with further sub-nodes

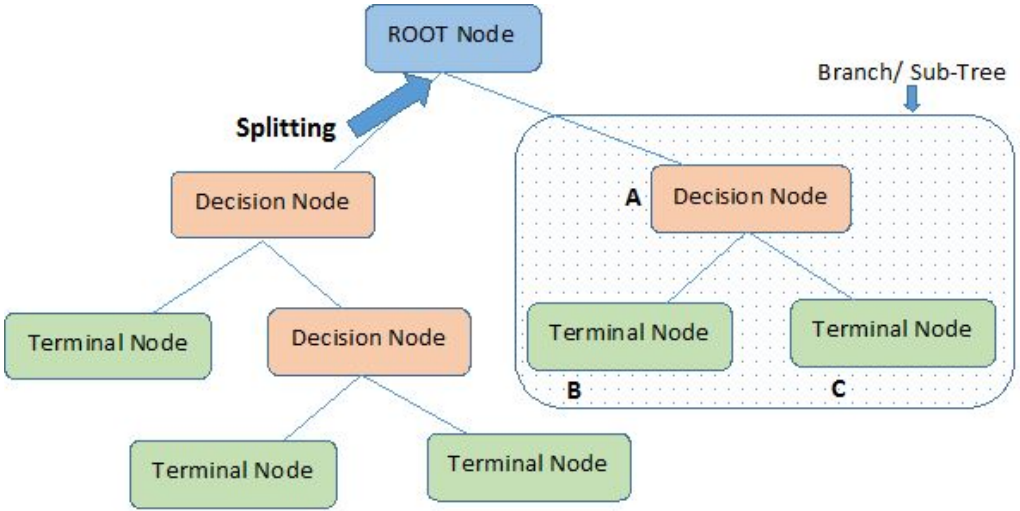
Leaf/ Terminal Node: A node without a split

Pruning: Remove sub-nodes of a decision node, the opposite process of splitting

Branch / Sub-Tree: Sub-section of a tree

Parent and Child Node: Describes the relationship between two nodes

Directed Acyclic Graph: Another name for our tree



From http://www.analyticsvidhya.com/blog/2015/01/decision-tree-simplified/

YOU DO: WORK IN PAIRS TO DRAW A DECISION TREE

- TAKE 5 MINUTES
- Some data ideas
  - Iris data
  - Titanic data
  - Will I (ride my bike | play tennis)- temperature, sun, wind, rain

- You may also hear Decision Trees referred to as CART
- CART stands for Classification and Regression Tree
- CART is one algorithm used to implement Decision Trees
- CART is the underlying algorithm used in scikit-learn
- Other Decision Tree algorithms include:
  - ID3 (Iterative Dichotomiser 3)
  - C4.5
  - C5.0

Some Advantages

Some Disadvantages

Some Advantages

- Simple to understand, interpret and visualize
- Requires little data preparation
- We can mix numeric and categorical data and have multiple outputs
- White Box - we can observe the model and explain it with boolean logic
- We can validate our model with statistical tests.

Some Disadvantages

Some Advantages

- Simple to understand, interpret and visualize
- Requires little data preparation
- We can mix numeric and categorical data and have multiple outputs
- White Box - we can observe the model and explain it with boolean logic
- We can validate our model with statistical tests.

Some Disadvantages

- Overfitting - we can have an overly complex tree that doesn’t adequately generalize data
- Unstable - small variations in data can produce different trees
- There are concepts that are too hard for Decision Trees to learn
- If one class is dominant, the Decision Tree can be biased
- Greedy
- Based on heuristic algorithms with locally optimal decisions made at the nodes

Why use Decision Trees?

Decision Trees are the foundation for ensemble methods, which mitigate many of their disadvantages (Bagging, Boosted Trees, Random Forest)

How are they used?

BP's GasOIL system for separating gas and oil on offshore platforms - decision trees replaced a hand-designed rules system with 2500 rules. C4.5-based system outperformed human experts and saved BP millions.

Try this Decision Tree Visualization