An Intro to Decision Trees

13 December 2016 | Denis Vrdoljak

LEARNING OBJECTIVES

After this lesson, you will be able to: - describe how decision trees work, - when to use them, - how to impliment a decision tree model in SKLearn, and - how to visualize the actual Decision Tree.

STUDENT PRE-WORK

Before this lesson, you should already be able to: - Experience with sckit-learn classifiers, - Know how to load, manipulate, and prep data for modeling, and - install a .dot (graph) file viewer, such as GraphViz (http://graphviz.org).

INSTRUCTOR PREP

Before this lesson, instructors will need to: - install a .dot (graph) file viewer, such as GraphViz (http://graphviz.org). - review the instructor and student .py files on GitHub, get familiar with the dataset (also on GitHub: https://github.com/denisvrdoljak/GAdenis

TIMING	TYPE	TOPIC
5 min	Introduction	Decision Trees Overview
5 mins	Practice	Implementing DT's in SKLearn
5 mins	Wrap Up	Conclusions, Lessons Learned

Decision Trees Overview (5 mins)

Recall how to set up data and models, and how to evaluate ML models from previous lessons.

Check: When is F-Score maximized?

- Decision trees are non-linear, and are based on aggregations of decision boundaries.
- They are weak learners that are easy to overfit.
- They are black-boxes, and their complexity makes the tree itself of little use in understanding the data. But, we can create and export a visualization.

Hands On: Modeling DT's in SKLearn (5 mins)

Here we will use a dataset related to Breast Cancer Detection and create a Decision Tree to predict Malignant vs. Benign cases. Then, we will export the actual decision tree and analize the branches and decision nodes to see if we can glean any useful information from it. Here, we will see examples of decision branches that result in very little information value, i.e., splits with a single datapoint going to one child while the rest of the data goes to the other child.

The data has already been pre-processed for you, and fit to a Naive Bayes model. We'll use this template to save time, and to highlight the similarities between setting up different Machine Learning models.

Wrap Up: Analysis and Conclusions (5 mins)

Questions to consider here are:
What did we learn about the data?
Was a Decision Tree a good model?
Did it overfit the data? If so, how?
What did we learn about the data from the visualization of the Decision Tree? What was the root decision? What was an example of a decision that provided little information?