CASOS

ORA Machine Learning

Michael J Kowalchuck

mkowalch@andrew.cmu.edu

Carnegie Mellon

Center for Computational Analysis of
Social and Organizational Systems
http://www.casos.cs.cmu.edu/

Carnegie Mellon

Algorithms

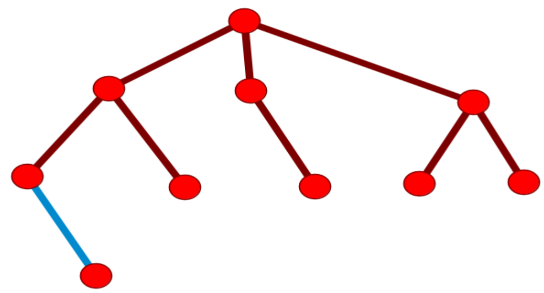
- Decision Tree
- Random Forest
  - A forest of decision trees
- JRip (coming soon)
  - Tree of rules

CASOS

Carnegie Mellon

Trees

- Connected acyclic graph with a root

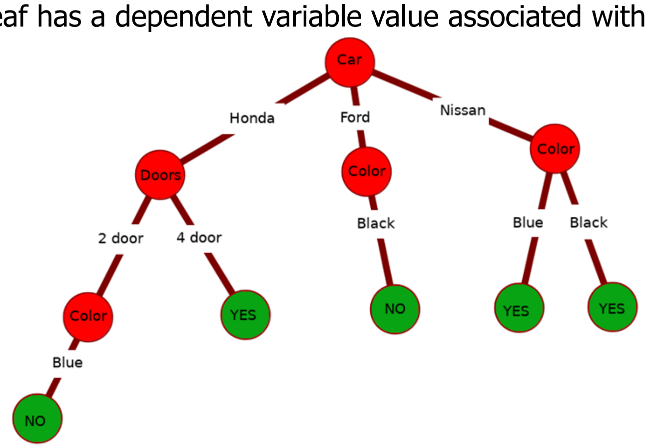


CASOS

Carnegie Mellon

Decision Tree

- Every node is associated with a variable in the data
- Every branch is a value that the parent node can take
- Every leaf has a dependent variable value associated with it



CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Decision Tree Overview

- Can be used for classification or regression problems
  - Ora's decision tree can only do Classification for now
    - Classification is where we are predicting a variable with discrete categories
- Still useful with unbalanced data (almost all positives or almost all negatives)
- Useful for finding the most important variables
- Weak learner
- Tends towards overfitting
- Walking up the tree from a leaf gives interesting subgroups

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Tennis Data

Outlook	Temperature	Humidity	Windy	Play
Sunny	Hot	High	FALSE	No
Sunny	Hot	High	TRUE	No
Overcast	Hot	High	FALSE	Yes
Rainy	Mild	High	FALSE	Yes
Rainy	Cool	Normal	FALSE	Yes
Rainy	Cool	Normal	TRUE	No
Overcast	Cool	Normal	TRUE	Yes
Sunny	Mild	High	FALSE	No
Sunny	Cool	Normal	FALSE	Yes
Rainy	Mild	Normal	FALSE	Yes
Sunny	Mild	Normal	TRUE	Yes
Overcast	Mild	High	TRUE	Yes
Overcast	Hot	Normal	FALSE	Yes

CASOS

Carnegie Mellon

DT Algorithm

- Pick the splitting variable
  - Pick variable with highest information gain or gini impurity
- Break data set into subsets, one for each value splitting variable can take
- create child nodes for each subset and repeat the process for each child

Carnegie Mellon

Choosing the Splitting Variable

- Information gain
  - Difference in entropy between parent and entropy of all child nodes
  - Problem: if there are too many unique values for a variable
- Gini impurity
  - How accurate the current split is
  - Useful in regression
  - Not gini coefficient that is something else

Carnegie Mellon

DT as Weak Learner

- Weak learners generally do not perform very well
  - Sometimes barely above random chance
- Decision trees performance can improve by picking the right parameters
- Weak learners can do well in bagging

Carnegie Mellon

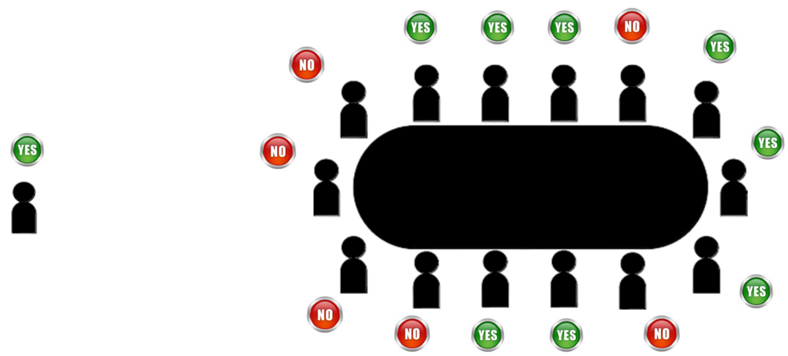
Decision Tree Paramters

- Minimum samples per node
- Maximum tree depth
- Without these the algorithm will perfectly overfit the data

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Random Forest

- Random Forest is based on Bagging
  - Ensemble method
  - Bagging is Bootstrap Aggregation



CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Bootstrap

- If you have a sample of some population that is independent and identically distributed
  - Resample from your sample with replacement (so the same sample can be taken multiple times) until you get the a new sample of the same size as the original
  - For each resample calculate your statistic

CASOS

Carnegie Mellon
isr institute for SOFTWARE RESEARCH

Build Random Forest

- For every tree in the forest, create a subsample of the data with replacement
  - Size of the forest is a parameter to the algorithm
- For every subsample, create a decision tree
  - These trees are allowed to overfit
  - When deciding which variable to split on only a subset of available variables is considered
    - The size of this subset is sqrt(variable_count)
    - This is the difference from bagged decision trees and random forest

CASOS