Random Forest & Feature Selection

Charlie Beirnaert

Thomas More University of Applied Sciences
Business School
Data Science, Protection & Security

Data Science
Advanced

Outline

Random Forest

Feature Selection

Outline

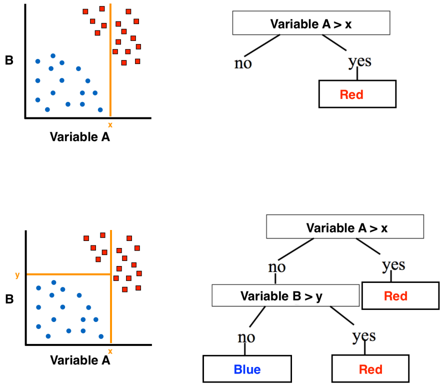
Random Forest

Feature Selection

Random Forest

- How does a decision tree work?
- How does a random forest work? (building process, decision proces, output)
- What are the main parameters and what do they do?

Decision Tree



1. Select the feature with the best separation between classes (how to quantify 'best'?)
2. Use this feature to make the next nodes (leaves) in the tree
3. Repeat process in each node until all samples have been classified (only one class at each endpoint)

Calculate separation quality

- quantify the Gini impurity in each node:

$$i = 1 - p_P^2 - p_N^2$$

- p_P is the proportion (fraction) of positives and vice verse for the negatives
- Use a feature to determine a split in the node and calculate the decrease in Gini impurity δi:

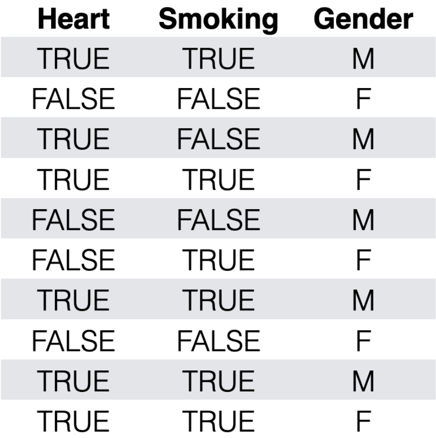
$$\delta i = i_{parent} - i_{child1} * f_{child1} - i_{child2} * f_{child2}$$

- f is the fraction of parent samples in that child

Gini impurity

- Lower Gini impurity signifies better separation between classes
- If only a single class occurs and the other is absent (perfect separation), the Gini impurity i will be 0 ($=$ 1 - 1 - 0)
- δi will be 0 if there is no improvement in the separation

Decision Tree example



Heart	Smoking	Gender
TRUE	TRUE	M
FALSE	FALSE	F
TRUE	FALSE	M
TRUE	TRUE	F
FALSE	FALSE	M
FALSE	TRUE	F
TRUE	TRUE	M
FALSE	FALSE	F
TRUE	TRUE	M
TRUE	TRUE	F

1. Make tree to classify heart disease
2. First node in tree: Smoking or Gender?

Decision Tree (dis)advantages

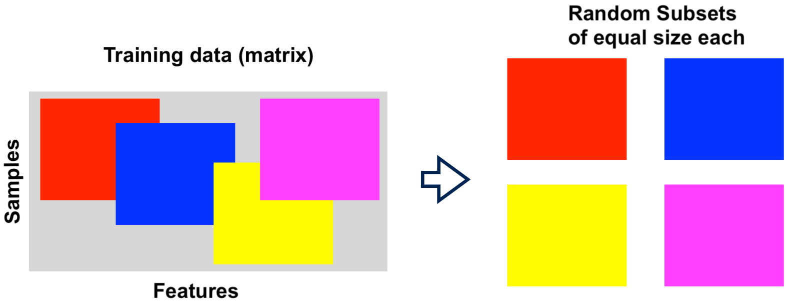
+ Intuitive approach, easy to implement and understand
+ Easy to check why a given sample is classified a certain way (so not a black box)
- Approach is ‘greedy’ and will continue until all samples are correctly classified (long and overfit trees)
- Tend to be very dependent on the input data and small changes can lead to very different trees

Random Forest

3 crucial concepts:

- Random subsets of data
- multiple trees
- voting to get result

Random Forest step 1



Training data (matrix)

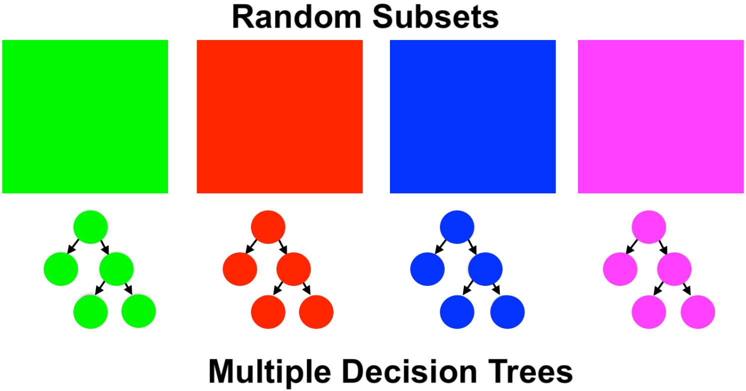
Samples

Features

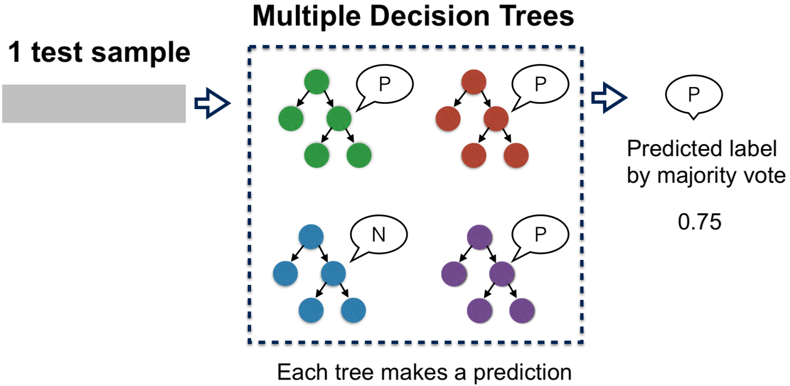
Random Subsets
of equal size each

Select random subsets of features and samples (with replacement)

Random Forest step 2



Random Forest step 3



Random Forest sources

1. https://towardsdatascience.com/why-random-forests-outperform-decision-trees-1b0f175a0
2. https://towardsdatascience.com/from-a-single-decision-tree-to-a-random-forest-b9523be

Outline

Random Forest

Feature Selection

Feature Selection

- Manual feature selection
- Statistical feature selection
- Model based feature selection
- Boruta feature selection

Boruta source

- https://towardsdatascience.com/boruta-explained-the-way-i-wish-someone-explained-it-t