Data Science 5

TILBURG UNIVERSITY

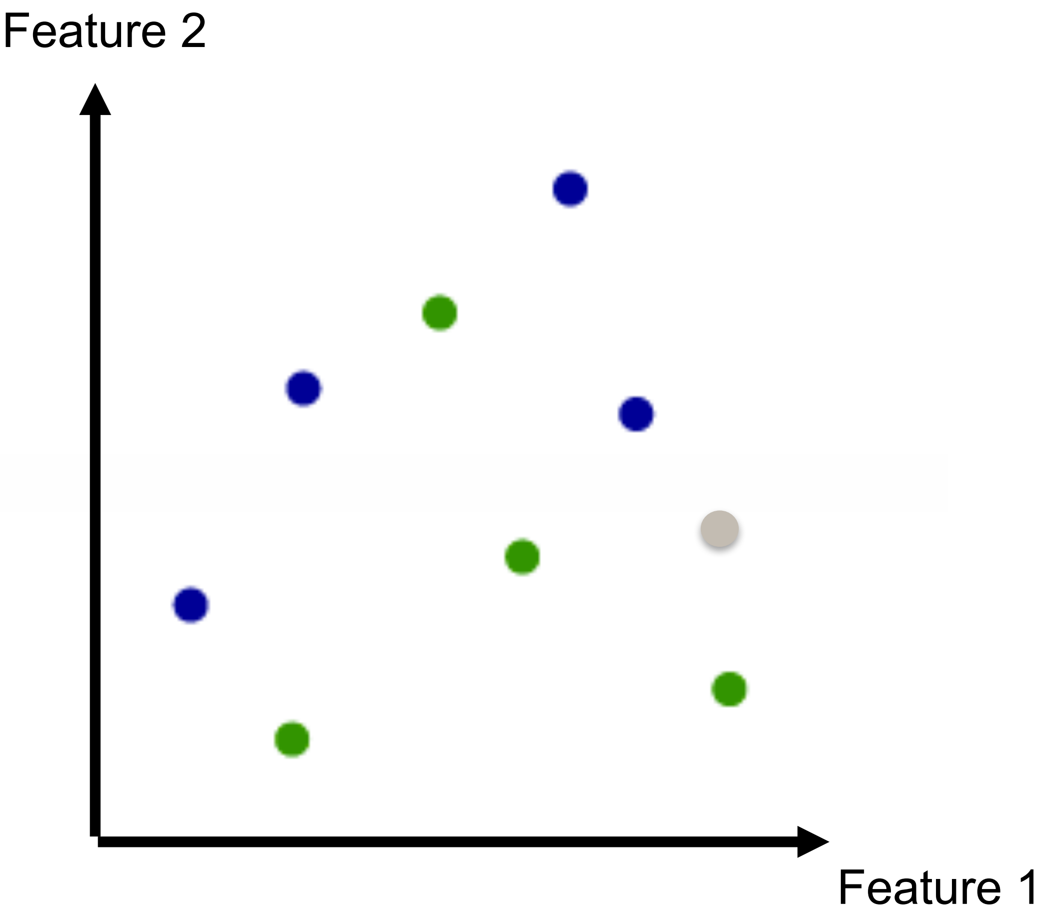
Understanding Society

Overview

- Decision Trees
- Random Decision Forests

- Decision Trees and Random Forests in Orange

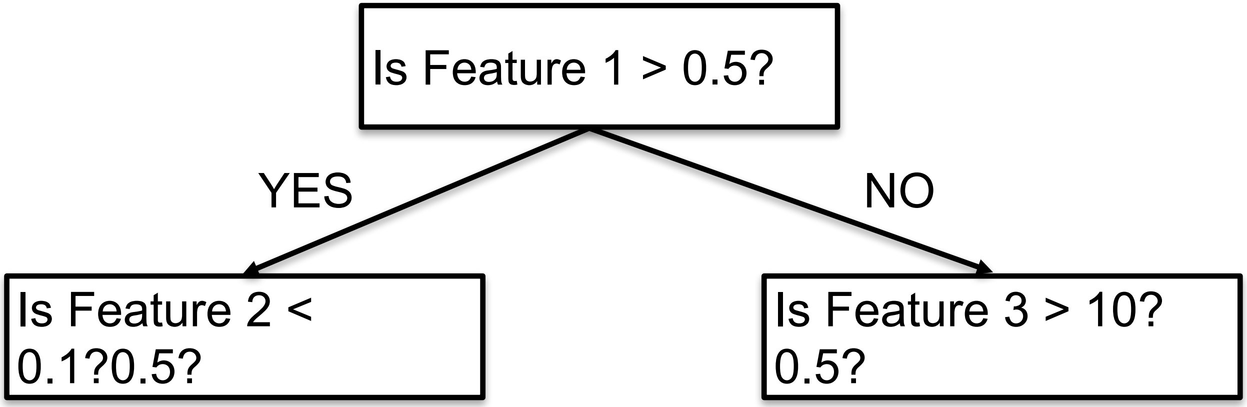
Classification Problem (blue or green?)



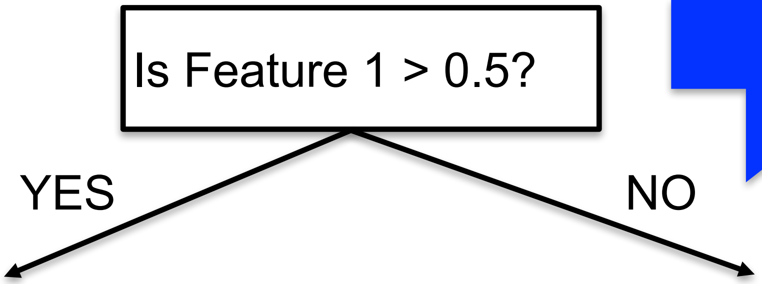
- Train instances
  - blue and green
- Test instance
  - gray
- Classifier induced from the data defines decision boundaries

Decision Trees

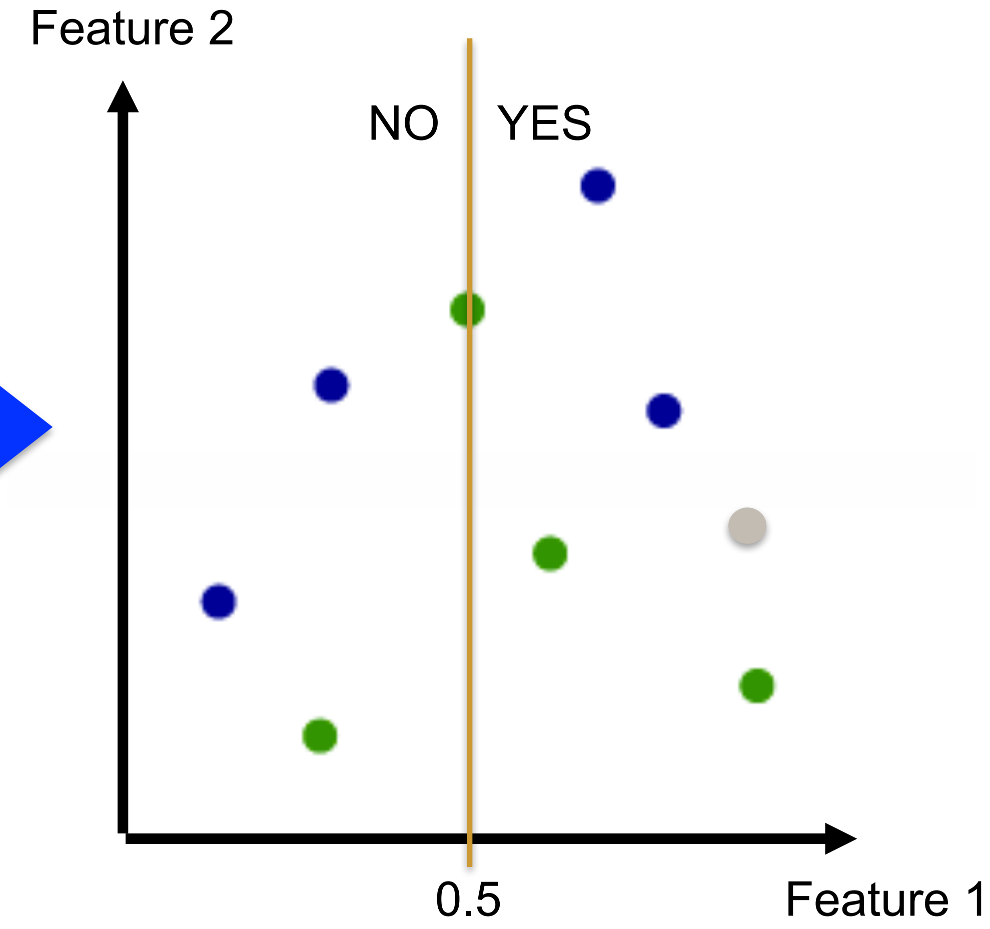
- Decision Trees take one feature at a time and test a binary condition
  For instance: is the feature larger than 0.5?
  If the answer is YES, grow a node to the left
  If the answer is NOW grow a node to the right



This results in the following decision Boundary



Is Feature 1 > 0.5?

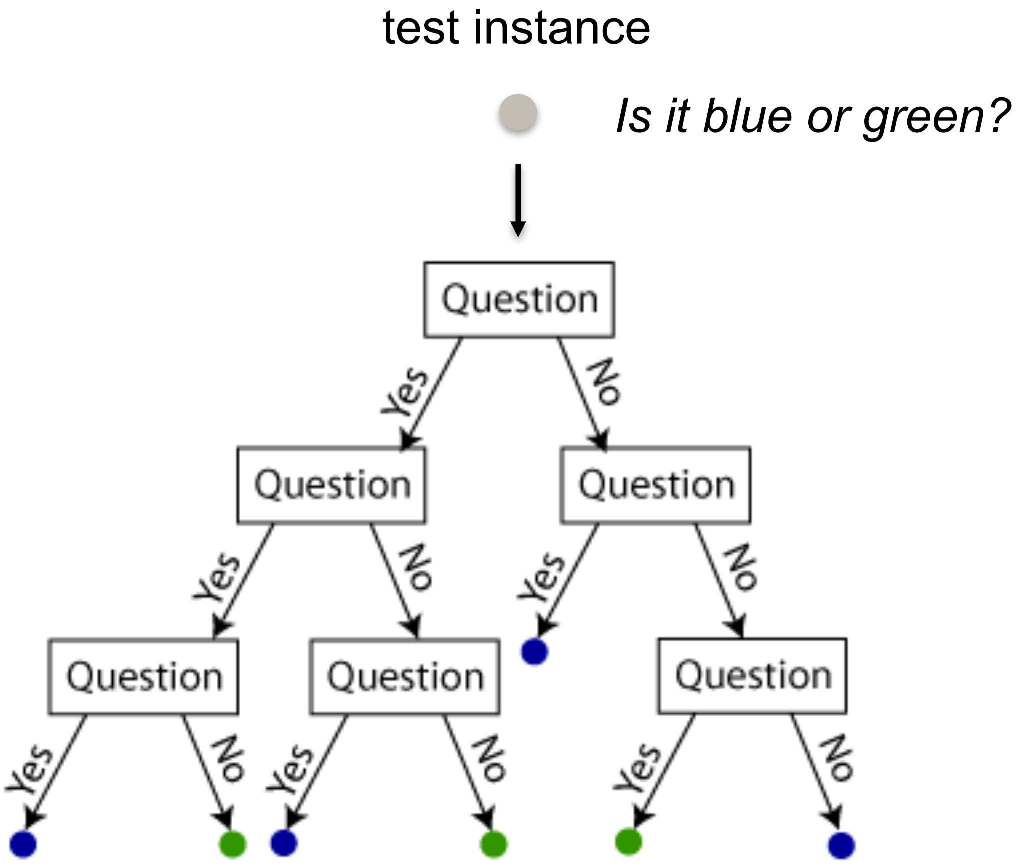


Decision Tree grows with each level of questions

- Each node (box) of the decision tree tests a condition on a feature
- The order of features is important
- It is like playing “20 questions”
  - “Guess the person”: it is better to start with the question “Is she female?”, rather than with “Is it Marie?”
  - The reason is that the answer to the first question maximises the information (“entropy”) gained from the answer.*
- In decision trees the order of features to be tested is determined by means of information theory (ID3 algorithm)

* Alternative: *Gini impurity* is a measure of how often a randomly chosen element from the data set would be incorrectly labeled if it were randomly labeled.

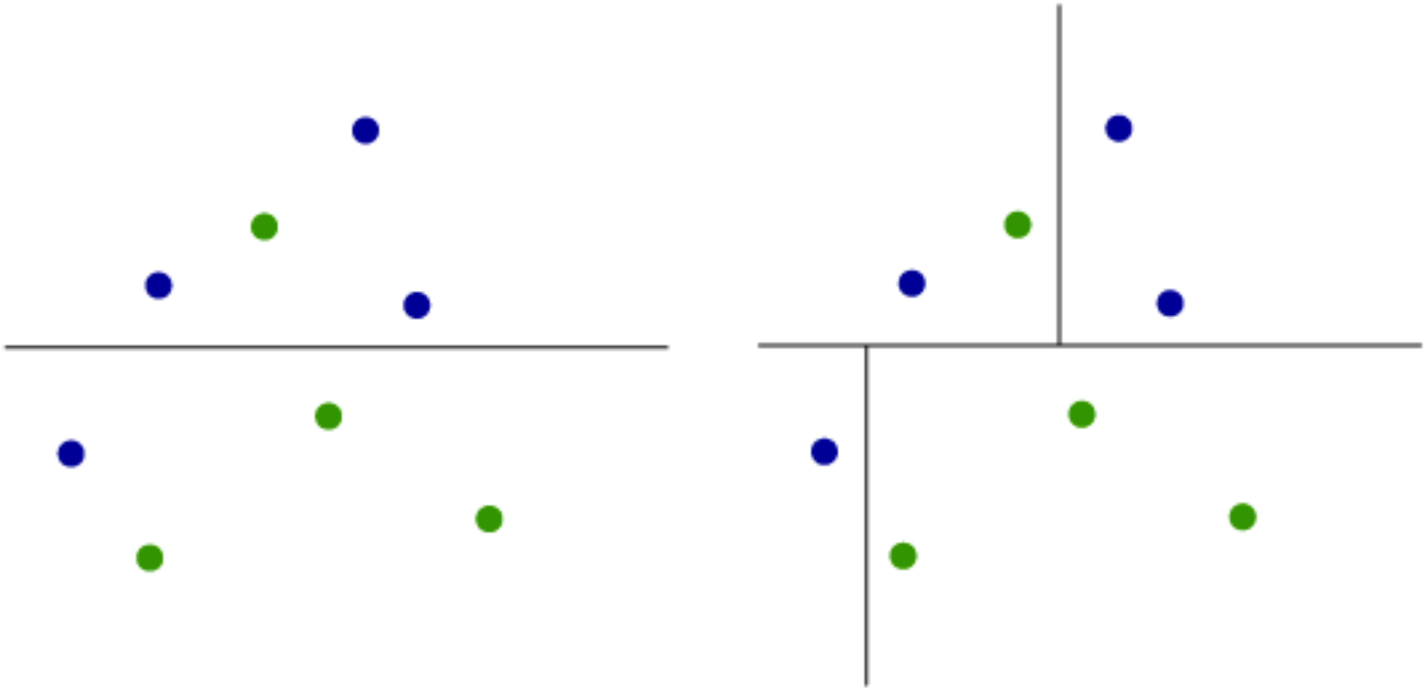
Decision Tree



Reproduced from: https://shapeofdata.wordpress.com/2013/07/02/decision-trees/

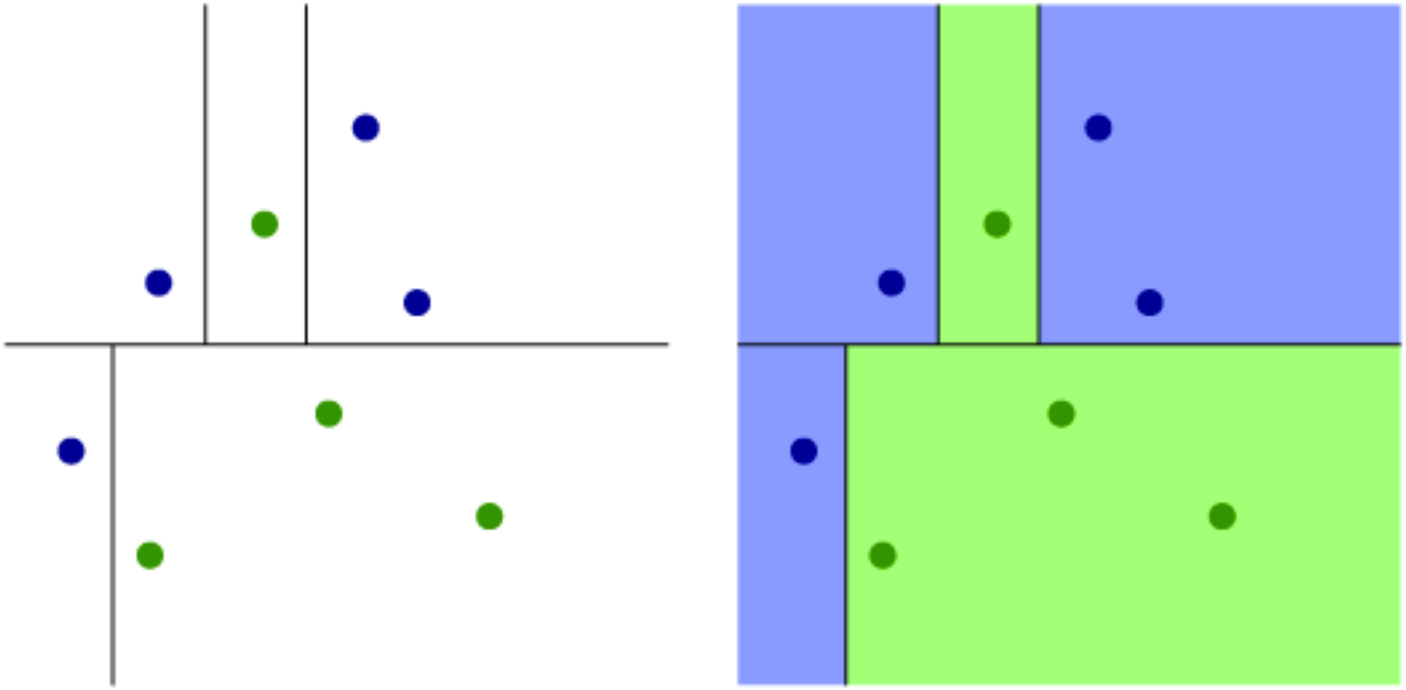
TILBURG UNIVERSITY

Each test (box) adds a decision boundary



Reproduced from: https://shapeofdata.wordpress.com/2013/07/02/decision-trees/

Adding another decision boundary

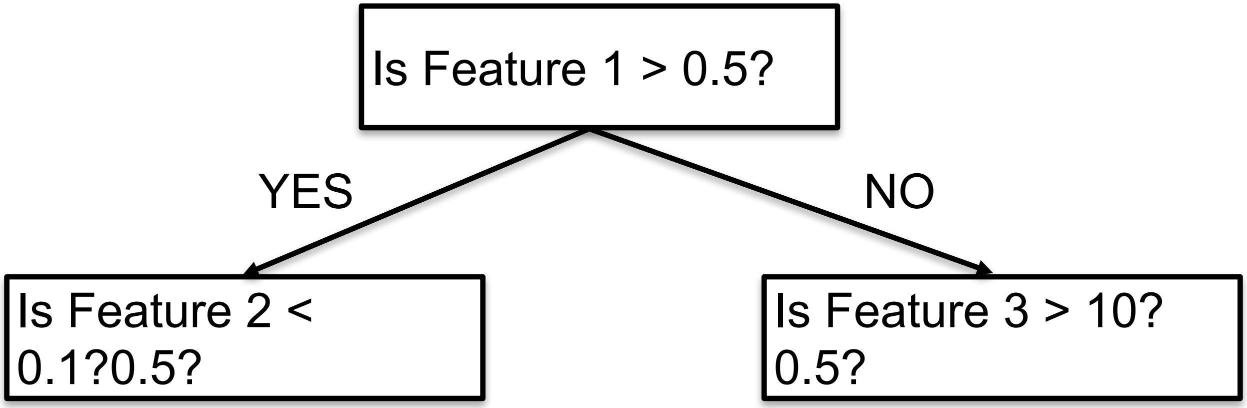


Reproduced from: https://shapeofdata.wordpress.com/2013/07/02/decision-trees/

TILBURG UNIVERSITY

Complexity of the induced model

- The complexity of the model induced by a decision tree is determined by the depth of the tree
- Increasing the depth of the tree increases the number of decision boundaries
- All decision boundaries are perpendicular to the feature axes, because at each node a decision is made about a single feature



Is Feature 1 > 0.5?

YES

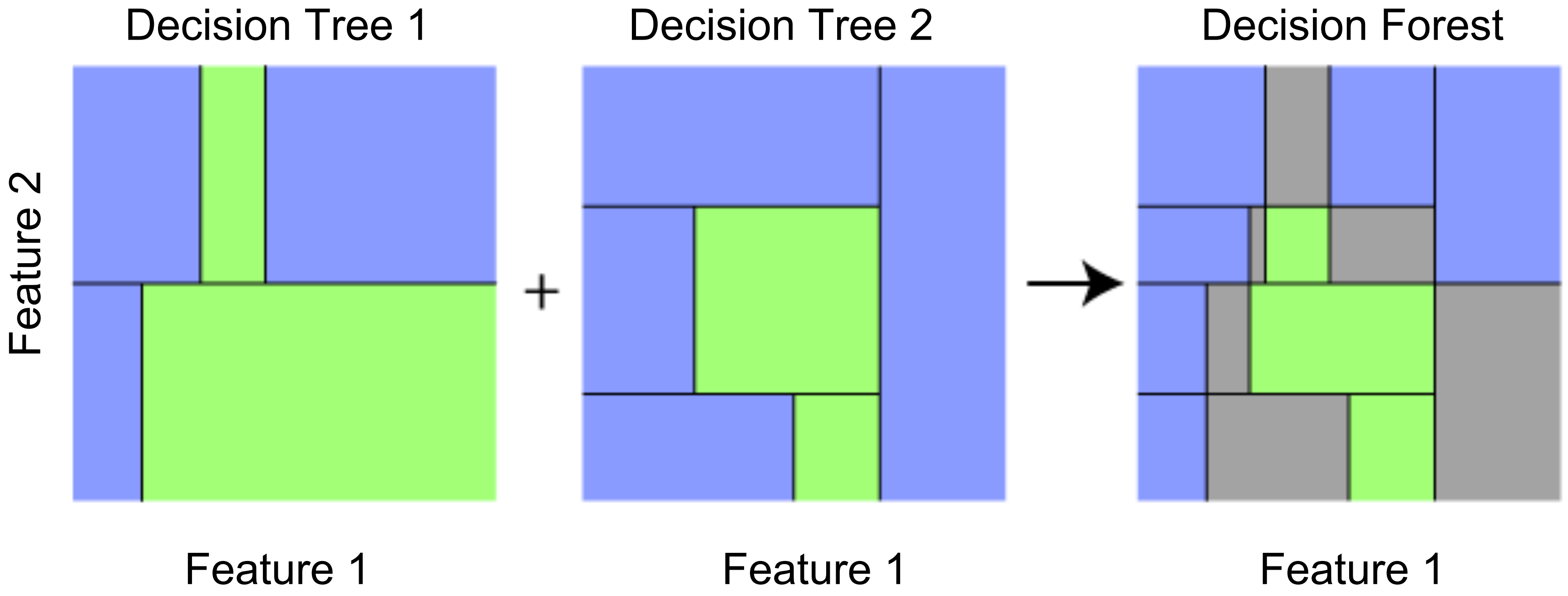
NO

Is Feature 2 < 0.1?0.5?

Is Feature 3 > 10? 0.5?

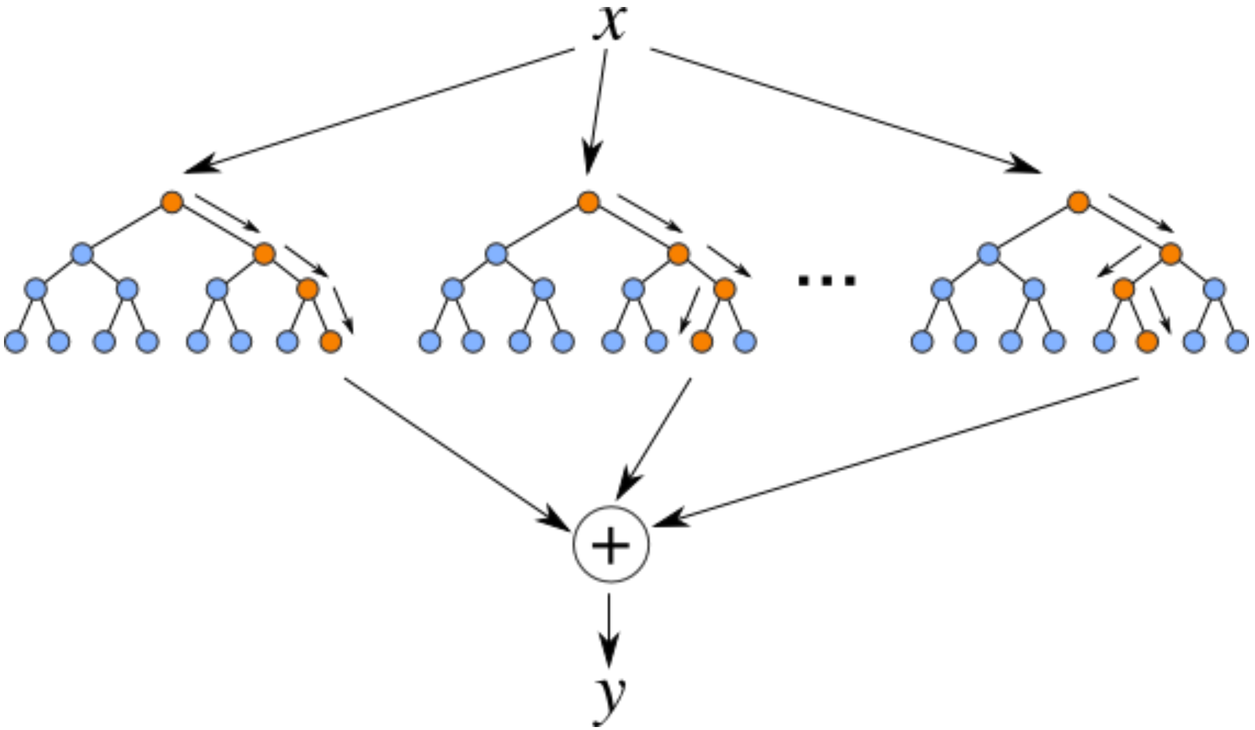
Random Decision Forests

- From one tree to many



Classification and Regression with RDFs

- Classification: the mode of the classes outputted by the trees.
- Regression: the mean of the values outputted by the trees.



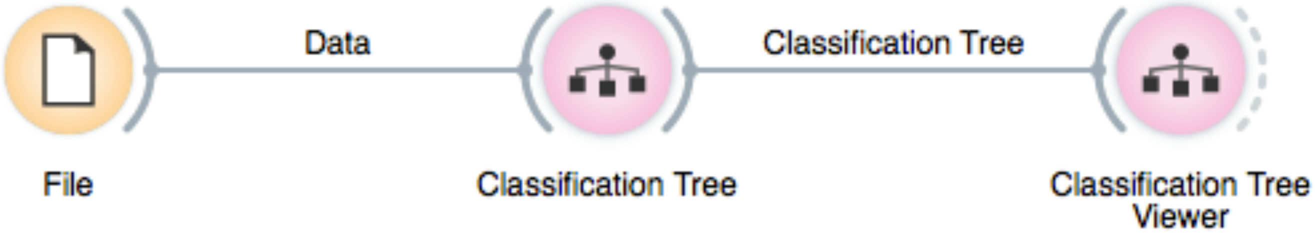
Complexity of Random Decision Forests

- The complexity of RDFs is determined by the number of trees (and their depths)

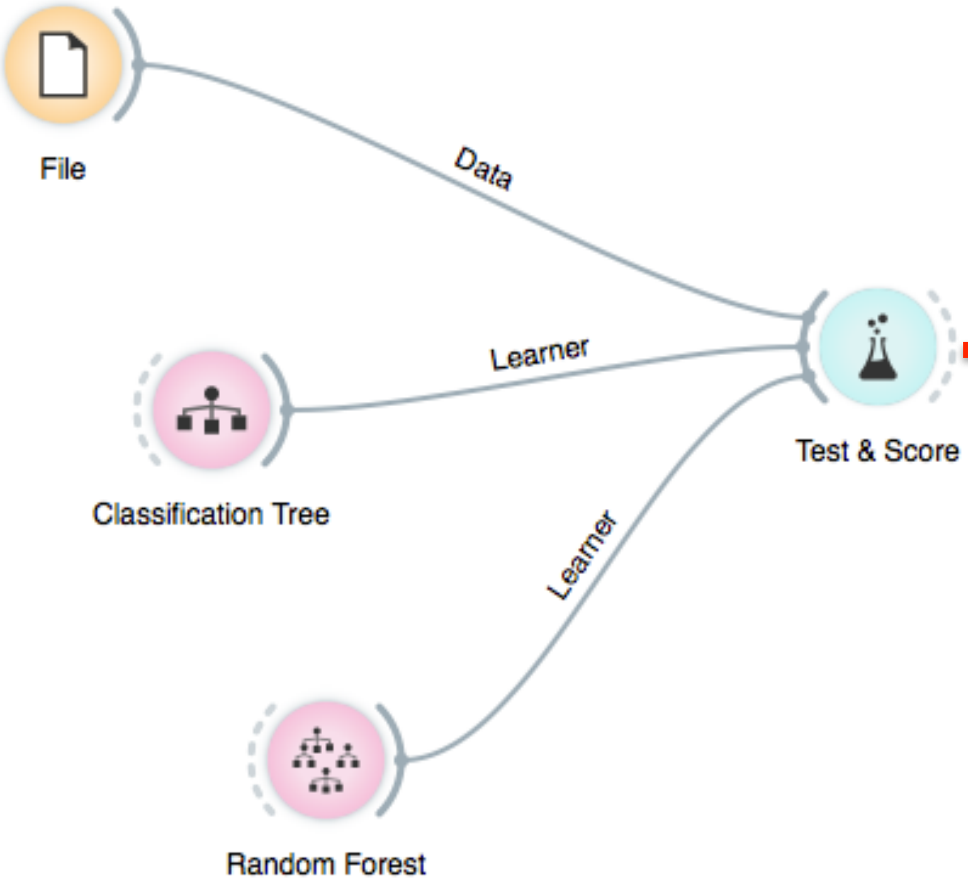
- In some decision forests trees are induced on the same complete set of features

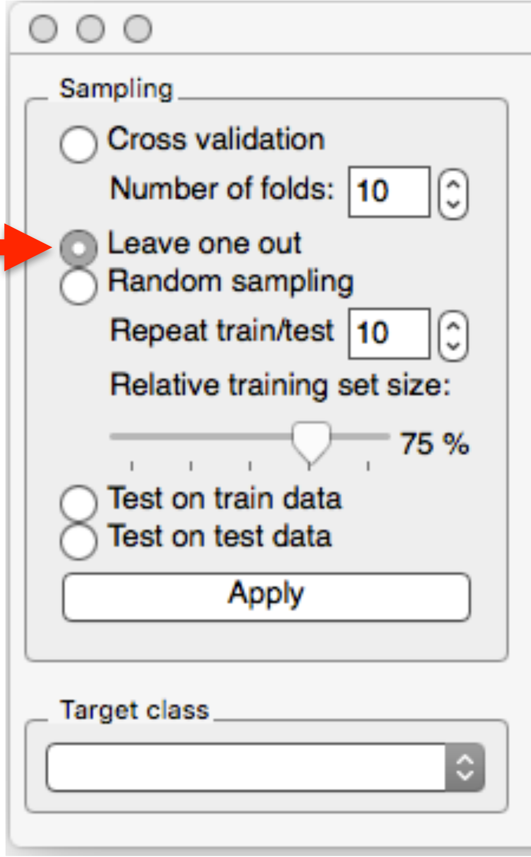
- In random decision forests, trees are induced on randomly selected subsets of features

Inducing and Visualising a Tree



Evaluating and Comparing Tree and Forest




Sampling

○ Cross validation

Number of folds: 10

● Leave one out

○ Random sampling

Repeat train/test 10

Relative training set size: 75 %

○ Test on train data

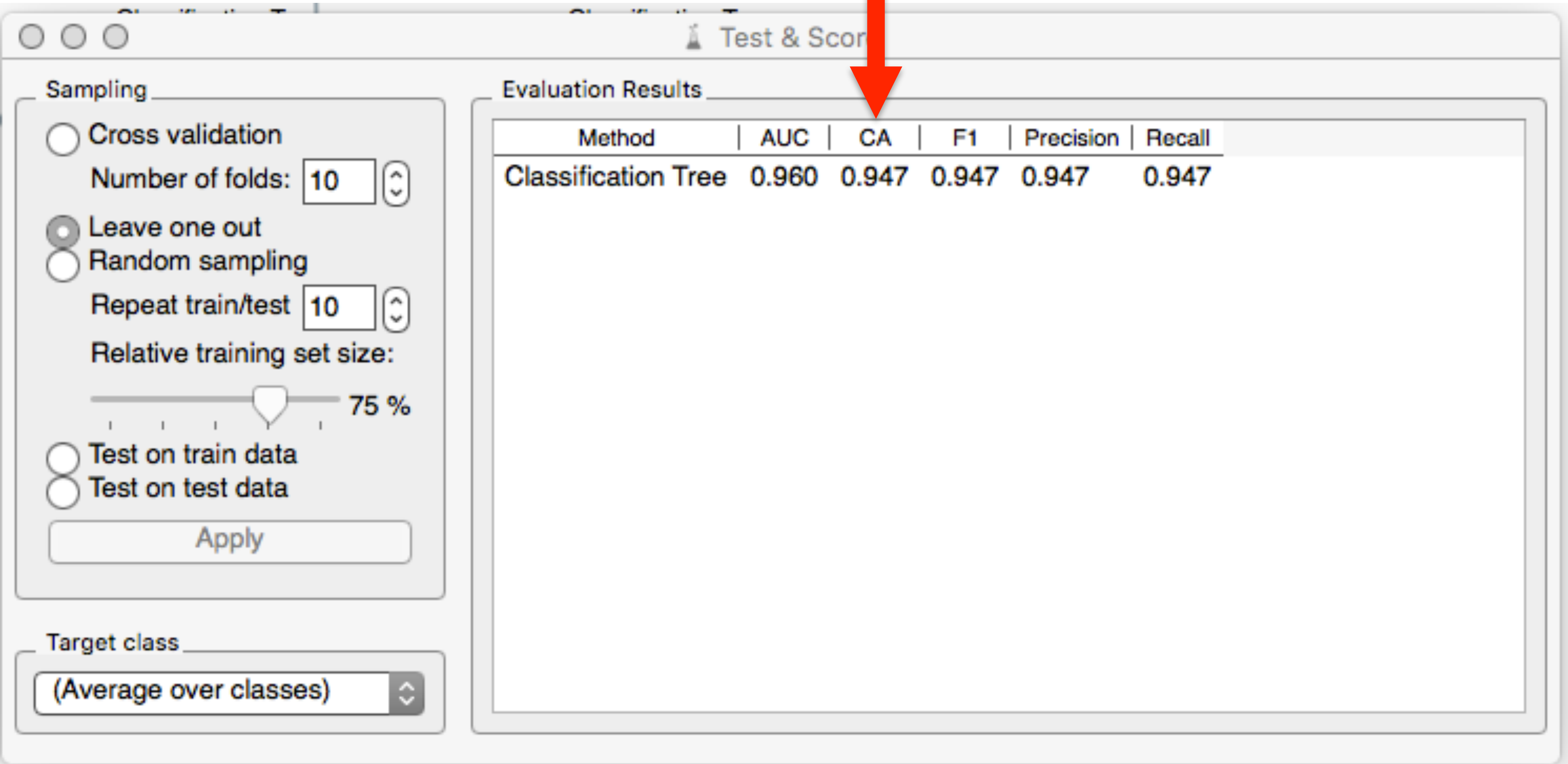
○ Test on test data

Apply

Target class

Performance Measure: Classification Accuracy

CA: proportion correctly classified



Sampling

○ Cross validation

Number of folds: 10

● Leave one out

○ Random sampling

Repeat train/test 10

Relative training set size: 75 %

○ Test on train data

○ Test on test data

Target class

(Average over classes)

Evaluation Results

Method	AUC	CA	F1	Precision	Recall
Classification Tree	0.960	0.947	0.947	0.947	0.947