Machine Learning & Data Mining @ NuIEEE

Tree Based Methods and Random Forests

Miguel Sandim & Paula Fortuna

01-03-2016

0 - Intro

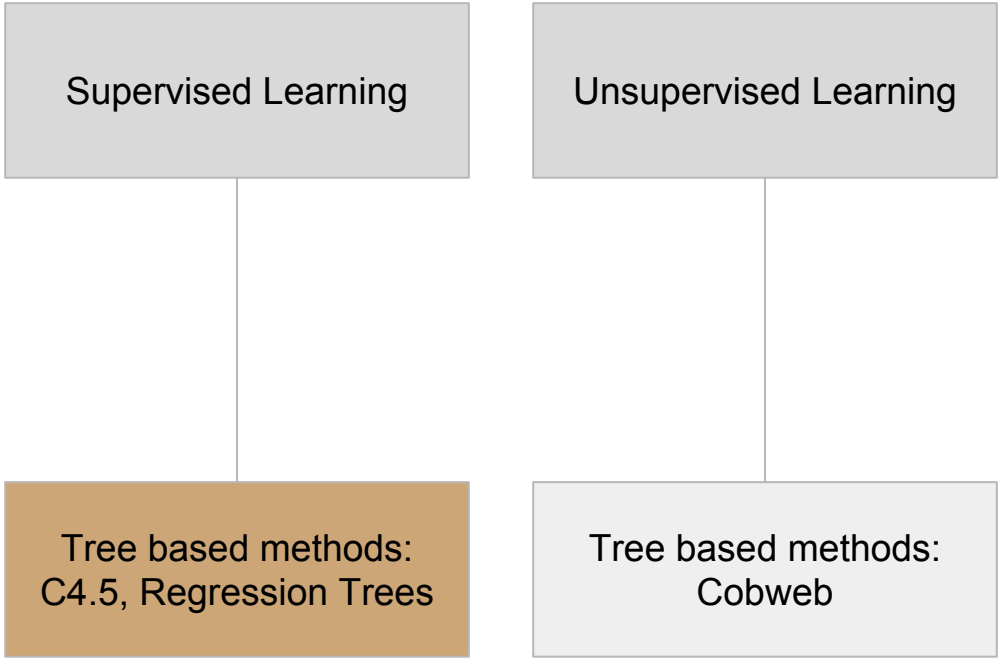
Attributes

Class or DV

	crim	zn	indus	chas	nox	medv
1	0.00632	18.0	2.31	0	0.538	24.0
2	0.02731	0.0	7.07	0	0.469	21.6
3	0.02729	0.0	7.07	0	0.469	34.7
4	0.03237	0.0	2.18	0	0.458	33.4
5	0.06905	0.0	2.18	0	0.458	36.2
6	0.02985	0.0	2.18	0	0.458	28.7
7	0.08829	12.5	7.87	0	0.524	22.9
8	0.14455	12.5	7.87	0	0.524	27.1
9	0.21124	12.5	7.87	0	0.524	16.5

Instance

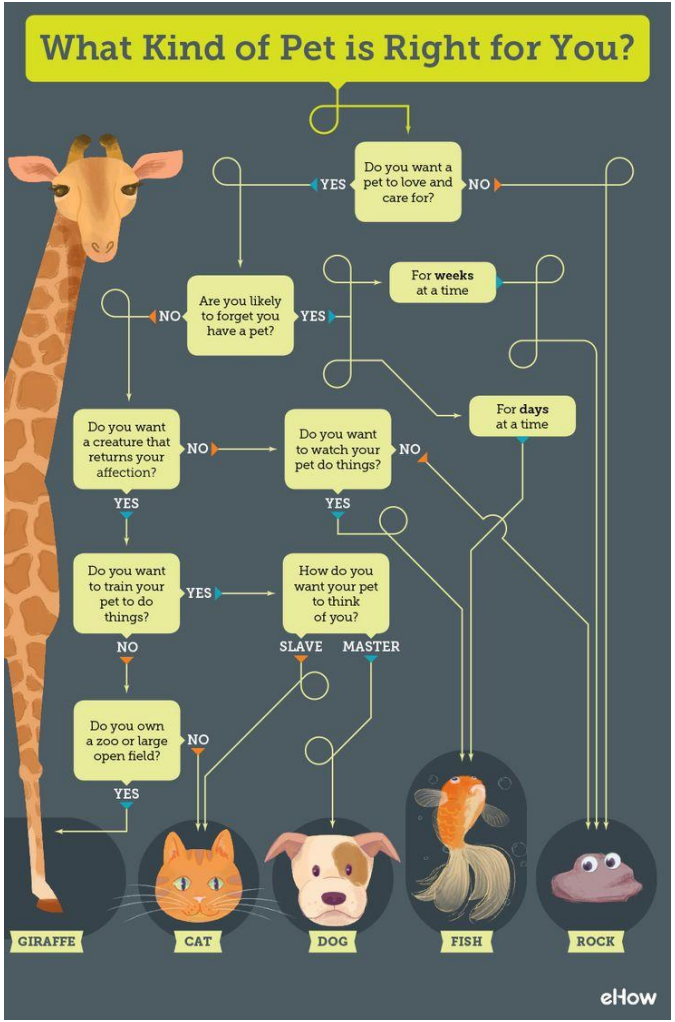
0 - Intro


Supervised Learning
Unsupervised Learning
Tree based methods:
C4.5, Regression Trees
Tree based methods:
Cobweb

Decision trees

What is?

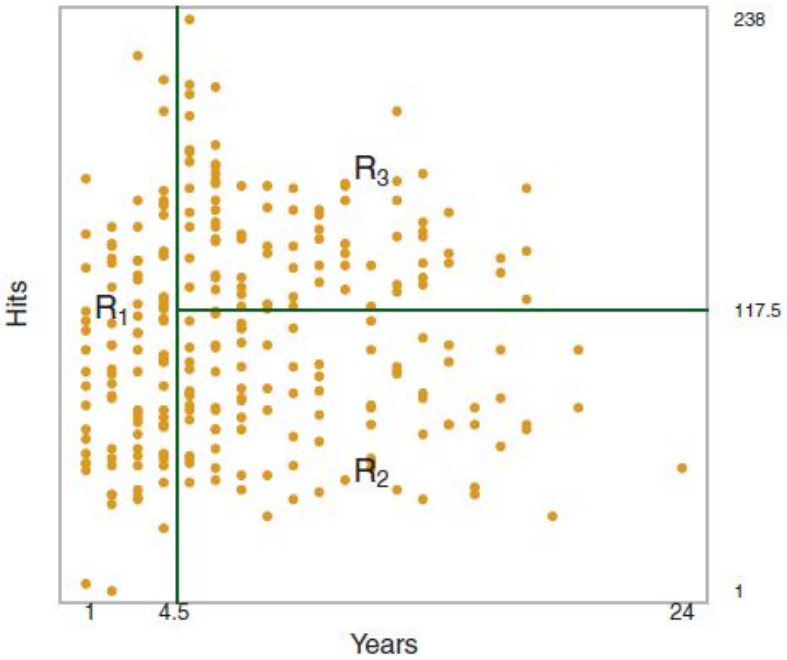
- Root
- Internal nodes
- External nodes - leafs



Decision trees

Regression Trees

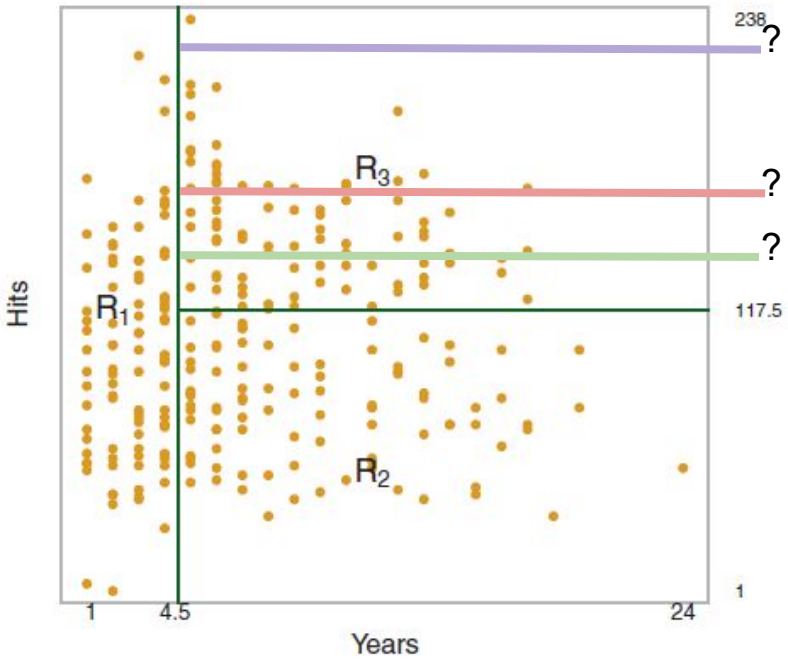
- We divide the predictor space - that is, the set of possible values for X1, X2, . . ., Xp - into J distinct and non-overlapping **regions**, R1, R2, . . . , RJ.
- For every observation that falls into the region Rj, we make the same prediction, which is simply the **mean** of the response values for the training observations in Rj.



Decision trees

Regression Trees

- How to divide?



Decision trees

Regression Trees

- Goal
- Minimize **RSS =** $\sum_{j=1}^{J} \sum_{i \in R_j} (y_i - \hat{y}_{R_j})^2$

total number of boxes

mean response within the jth box

  - np problem -> we use a top-down **greedy** approach instead
  - In each iteration for each attribute we see what is the best cut point where:

$$R_1(j,s) = \{X | X_j < s\} \text{ and } R_2(j,s) = \{X | X_j \geq s\}.$$

minimize: $\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$

Decision trees

Regression Trees

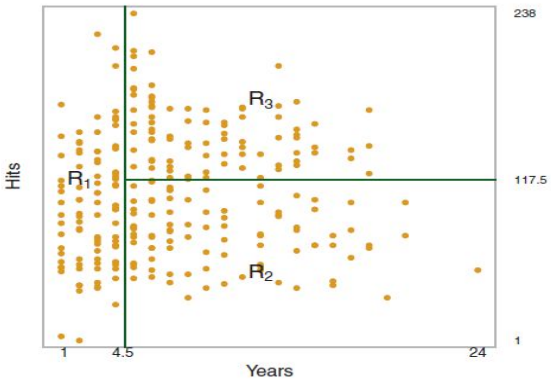
- How do we know we have reached a leaf?
  - We can limit the number of instances in each leaf to a minimum (e. g. 5)

	crim	zn	indus	chas	nox	medv
1	0.00632	18.0	2.31	0	0.538	24.0
2	0.02731	0.0	7.07	0	0.469	21.6
3	0.02729	0.0	7.07	0	0.469	34.7
4	0.03237	0.0	2.18	0	0.458	33.4
5	0.06905	0.0	2.18	0	0.458	36.2

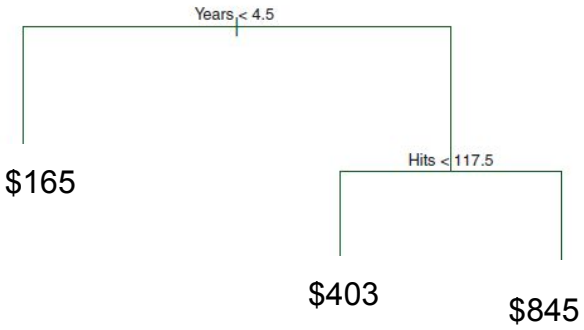
Decision trees

Regression Trees

- From **chart** to **decision rules** to **trees**



R1 ={X | Years<4.5}
R2 ={X | Years>=4.5, Hits<117.5}
R3 ={X | Years>=4.5, Hits>=117.5}



  - Years is the most important predictor.

Decision trees

Problems

- The process of making the tree is likely to overfit the data
  - because the resulting tree might be too complex, the solution is...
- Pruning
  - **Cost complexity pruning -** use a parameter α
  - For each value of α there corresponds a subtree T ⊂ T0 such that:

original tree

$$\sum_{m=1}^{|T|} \sum_{i:\, x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha |T|$$

number of leafs of tree *T*

  - When α = 0, then the subtree T will simply be equal to T0.
  - However, as α increases, there is a price to pay for having a tree with many leafs.

Decision trees

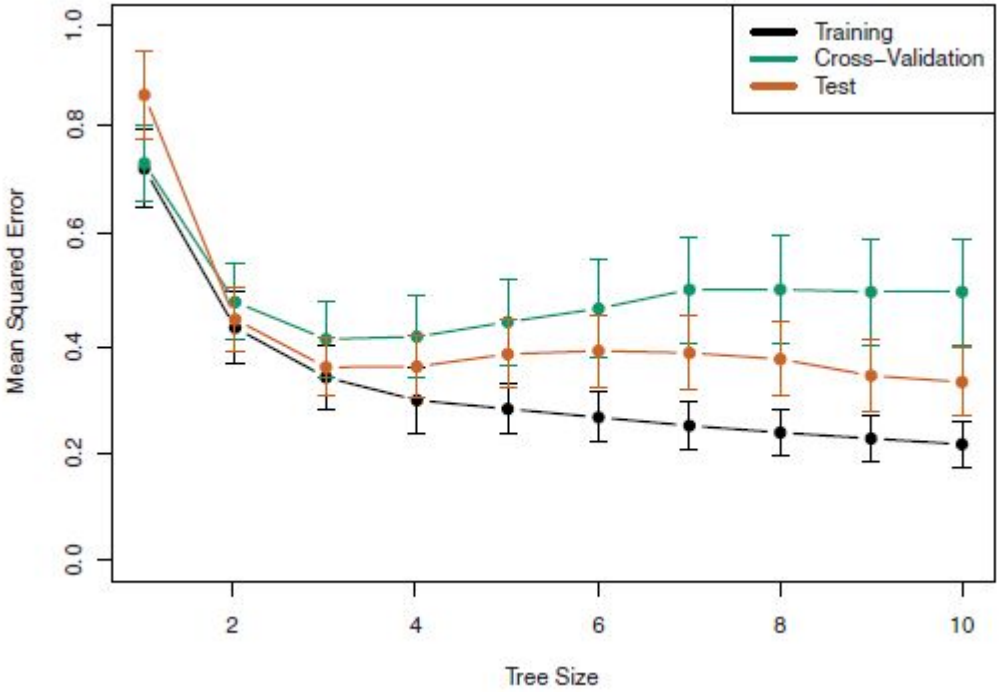
Pruning

Algorithm

- Build the tree (stop when have a minimum number of instances in leafs).
- Apply cost complexity pruning to the large tree in order to obtain a sequence of best subtrees, as a function of α.
- Use K-folds to decide the best α, based on MSE.
- Return the tree that corresponds to the chosen value of α.

Decision trees

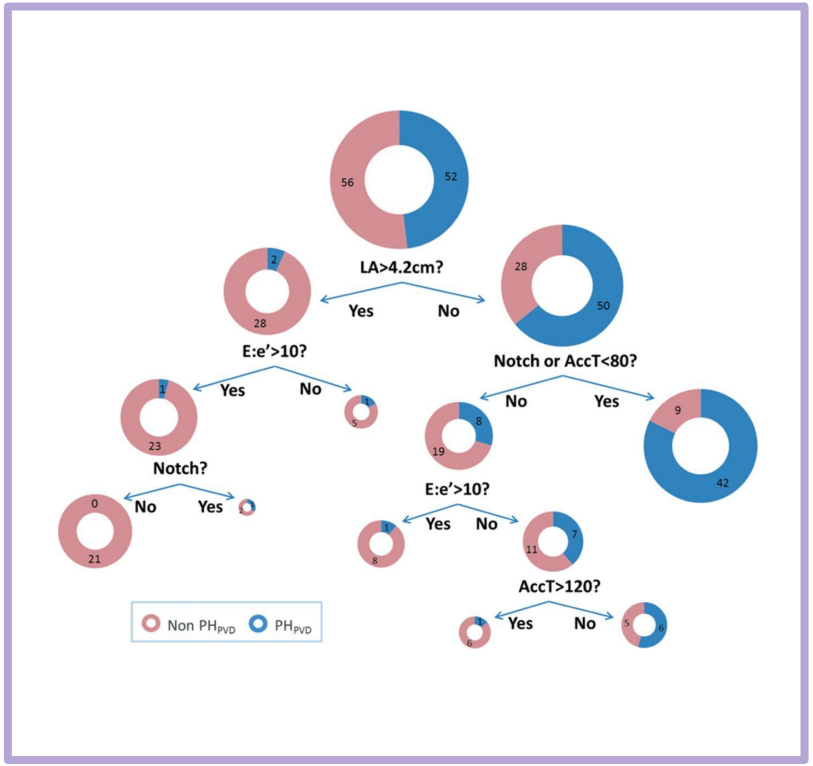
Pruning



Decision trees

Classification Trees

- Similar to regression trees.
- Qualitative dependent variables:
  - Instead of the mean, we use the most **commonly occurring class** of training in the region to which it belongs.
  - Also interested in the class **proportions** among the leaf.



Decision trees

Classification Trees

- How to evaluate the model?
  - Error Rate

$$E = 1 - \max_k(\hat{p}_{mk})$$

  - Gini Impurity

$$G = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk})$$

  - Cross entropy

$$D = -\sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk}$$

Decision trees

Classification Trees

- deviance (reported in summary)

number of occurrences from class m in box k

$$-2\sum_m\sum_k n_{mk}\log\hat{p}_{mk}$$

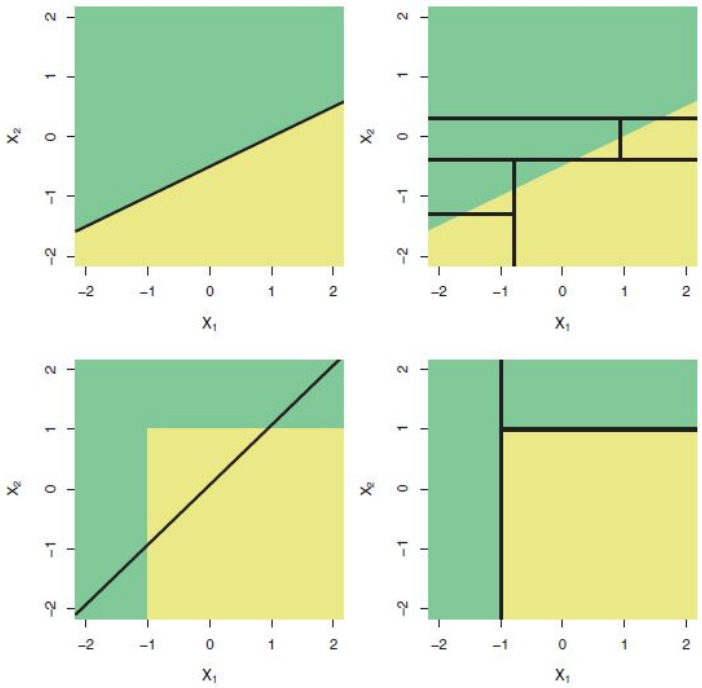
- Residual mean deviance

$$\frac{-2\sum_m\sum_k n_{mk}\log\hat{p}_{mk}}{n-|T_0|}$$

Decision trees

VS. Linear Regression Models

- It depends.
- If the relationship is linear an approach such as linear regression will outperform a method such as a regression tree.
- If instead there is a highly non-linear and complex relationship, then decision trees may outperform classical approaches.



Decision trees

Advantages

- Easy to understand.
- Explained graphically.
- Can handle qualitative predictors.

Disadvantages

- Do not have the same level of predictive accuracy.
- Non-robusts. Small changes in data can cause changes in the model. **!!!!!**

Decision trees

Methods to improve predictive performance

- Bagging
  - Bootstrap, aggregation, or bagging, is a general-purpose procedure.
  - Take many training sets from the population, build a separate prediction model using each training set, and average the resulting predictions.
  - Reduces variance by averaging a set of observations.
  - Particularly useful and frequently used in the context of decision trees.
- Random forests
  - Random forests provide an improvement over bagged trees by way of a random small tweak that decorrelates the trees.
- Boosting
  - Boosting works in a similar way, except that the trees are grown sequentially: each tree is grown using information from previously grown trees.

Decision trees

Random forests

Decision Tree

Some Algorithms

MIEIC

- C4.5 (J48 in Weka)
- COBWEB

Others

- REPtrees

Final session topics

- Data preprocessing (e. g. missing values, outliers)
- How to choose a family of algorithms?
- How to evaluate if an algorithm is working
- Overfitting & underfitting
- Validation Methods - K-folds