Introduction to Machine Learning

Chapter 13: Trees

Bernd Bischl, Christoph Molnar

Department of Statistics – LMU Munich

Winter term 2017/18

TREES - INTRODUCTION

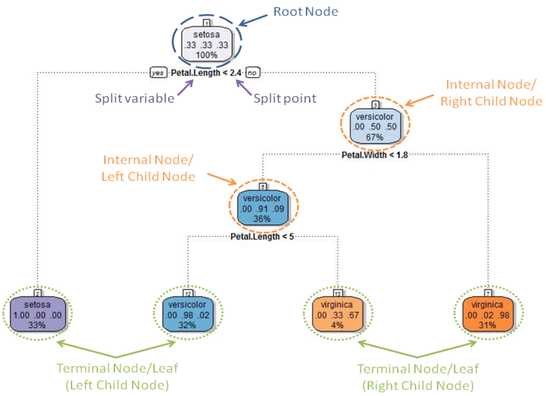
Can be used for classification, regression (and much more!)

Zoo of tree methodologies

- AID (Sonquist and Morgan, 1964)
- CHAID (Kass, 1980)
- CART (Breiman et al., 1984)
- C4.5 (Quinlan, 1993)
- Unbiased Recursive Partitioning (Hothorn et al., 2006)

CART

- Classification and Regression Trees, introduced by Breiman
- Binary splits are constructed top-down
- Only constant prediction in each leaf



CART

- In the greedy top-down construction, features and split points are selected by exhaustive search.
- For each node, one iterates over all features, and for each feature over all split points.
- The best feature and split point, which make both created child nodes most pure, measured by a split criterion, are selected.
- The procedure then is applied to the child nodes in a recursive manner.

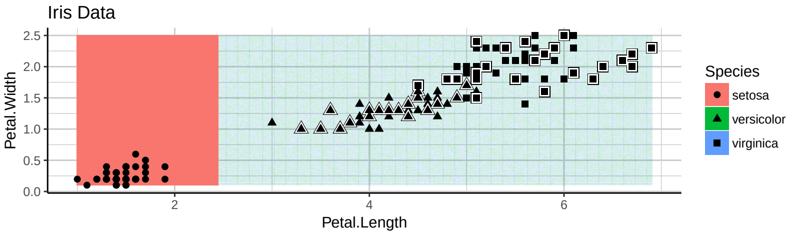
CART

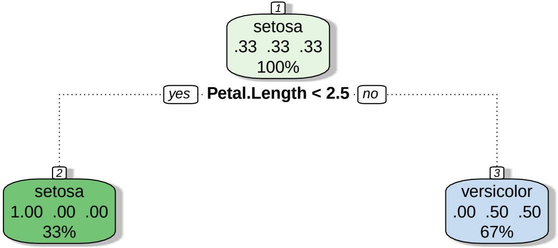
- Trees divide the feature space $\mathcal{X}$ into rectangles and fit simple models (e.g: constant) in these:

$$f(x) = \sum_{m=1}^{M} c_m \mathbb{I}(x \in R_m),$$

where M rectangles R_m are used. c_m is a predicted numerical response, a class label or a class distribution.

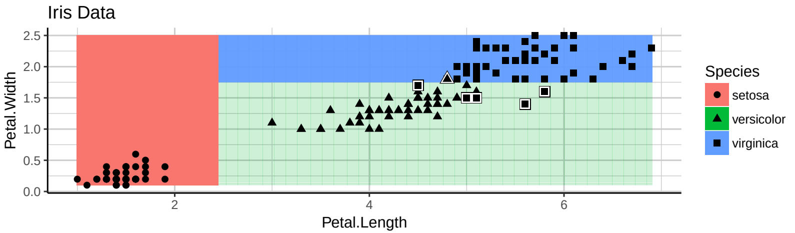
Example for Classification: Iris-Data

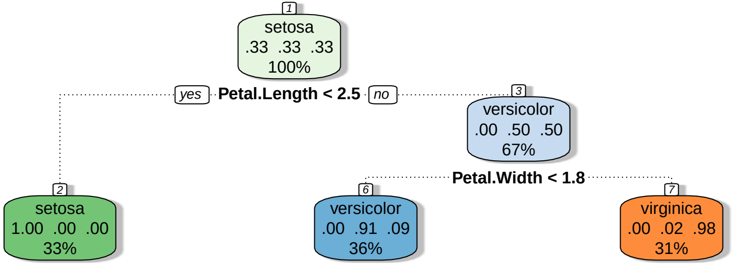




CART

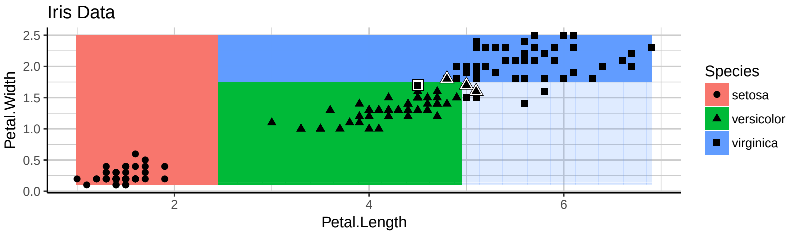
Example for Classification: Iris-Data

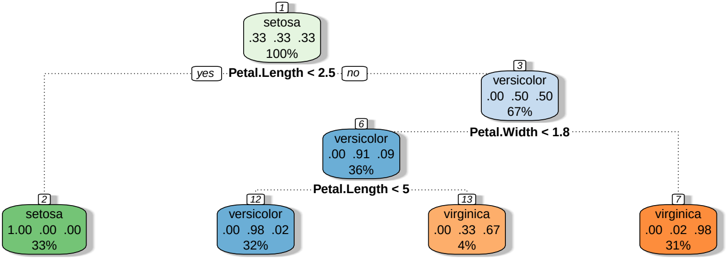




CART

Example for Classification: Iris-Data

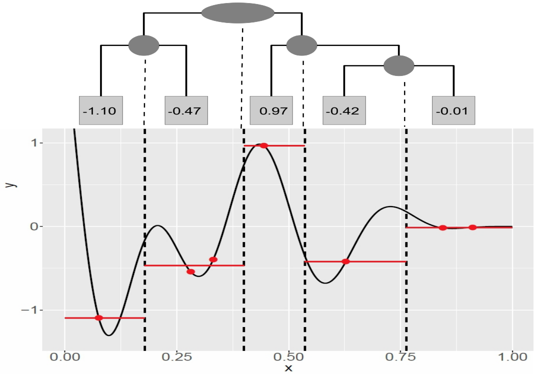




CART

Example for Regression:

x	y
0.075	-1.095
0.281	-0.543
0.331	-0.396
0.445	0.965
0.628	-0.421
0.845	-0.018
0.912	-0.011

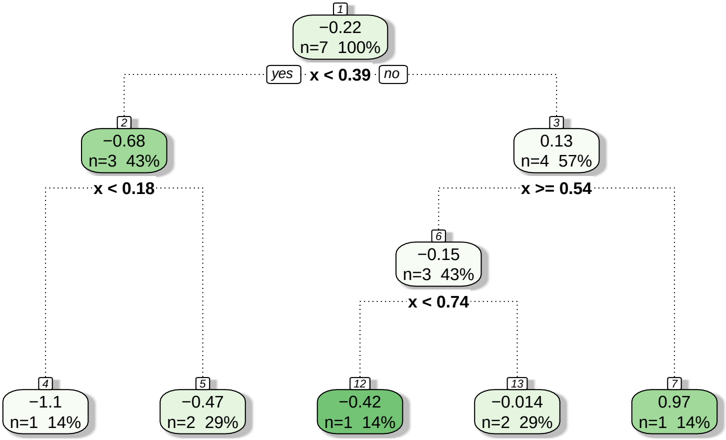


Data points (red) were generated from the underlying function (black):

$$sin(4x - 4) * (2x - 2)^2 * sin(20x - 4)$$

CART

Example for Regression:



CART: SPLIT CRITERIA

Let $\mathcal{N} \subseteq \mathcal{D}$ be a parent node with two child nodes $\mathcal{N}_1$ and $\mathcal{N}_2$.
Dividing all of the data with respect to the split variable x_j at split point t, leads to the following half-spaces:

$$\mathcal{N}_1(j,t) = \{(x,y) \in \mathcal{N} : x_j \leq t\} \text{ and } \mathcal{N}_2(j,t) = \{(x,y) \in \mathcal{N} : x_j > t\}.$$

Assume we can measure the impurity of the data in node $\mathcal{N}$ (usually the label distribution) with function $I(\mathcal{N})$. This function should return an "average quantity per observation".
Potential splits created in a node $\mathcal{N}$ are then evaluated via impurity reduction:

$$I(\mathcal{N}) - \frac{|\mathcal{N}_1|}{|\mathcal{N}|} I(\mathcal{N}_1) - \frac{|\mathcal{N}_2|}{|\mathcal{N}|} I(\mathcal{N}_2)$$

$|\mathcal{N}|$ means number of data points contained in (parent) node $\mathcal{N}$.

CART: SPLIT CRITERIA

- **Continuous targets:** mean-squared error / variance

$$I(\mathcal{N}) = \frac{1}{|\mathcal{N}|} \sum_{(x,y) \in \mathcal{N}} (y - \bar{y}_\mathcal{N})^2$$

with $\bar{y}_\mathcal{N} = \frac{1}{|\mathcal{N}|} \sum_{(x,y) \in \mathcal{N}} y$.

Hence, the best prediction in a potential leaf $\mathcal{N}$ is the mean of the contained y-values, i.e. impurity here is variance of y-values.

We can also obtain this by considering:

$$\min_{j,t} \left(\min_{c_1} \sum_{(x,y) \in \mathcal{N}_1} (y - c_1)^2 + \min_{c_2} \sum_{(x,y) \in \mathcal{N}_2} (y - c_2)^2 \right).$$

The inner minimization is solved through: $\hat{c}_1 = \bar{y}_1$ and $\hat{c}_2 = \bar{y}_2$

CART: SPLIT CRITERIA

- **Categorical targets (K categories):** "Impurity Measures"
  - Gini index:

$$I(\mathcal{N}) = \sum_{k \neq k'} \hat{\pi}_k^{\mathcal{N}} \hat{\pi}_{k'}^{\mathcal{N}} = \sum_{k=1}^{g} \hat{\pi}_k^{\mathcal{N}} (1 - \hat{\pi}_k^{\mathcal{N}})$$

  - misclassification error:

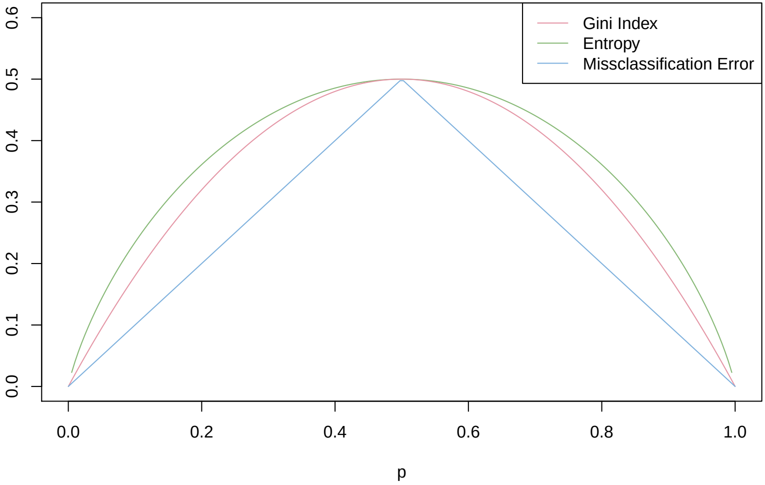
$$I(\mathcal{N}) = 1 - \max_k \hat{\pi}_k^{\mathcal{N}}$$

  - Shannon entropy:

$$I(\mathcal{N}) = -\sum_{k=1}^{g} \hat{\pi}_k^{\mathcal{N}} \log \hat{\pi}_k^{\mathcal{N}} ,$$

where $\hat{\pi}_k^{\mathcal{N}}$ corresponds to the relative frequency of category k of the response.

CART: SPLIT CRITERIA



IMPURITY MEASURES

- In general the three proposed splitting criteria are quite similar.
- Entropy and Gini index are more sensitive to changes in the node probabilities.
- **Example:** two-class problem with 400 obs in each class and two possible splits:

Split 1:

	class A	class B
Left node	300	100
Right node	100	300

Split 2:

	class A	class B
Left node	400	200
Right node	0	200

IMPURITY MEASURES

Split 1:

	class A	class B
Left node	300	100
Right node	100	300

Split 2:

	class A	class B
Left node	400	200
Right node	0	200

- Both splits produce a misclassification rate of $\frac{200}{800} = 0.25$
- Split 2 produces a pure node and is probably preferable.
- The average node impurity after a split based on x_1 is 0.375 (Gini) or 0.406 (Entropy) and $\frac{1}{3}$ (Gini) or 0.344 (Entropy) after a split based on x_2.
- Both criteria prefer split 2 and *choose* the result with a pure node.

IMPURITY MEASURES

- For metric features the exact split points can be ambiguous.
- If the classes of the response (for classification trees) are completely separated regarding the value range of the feature, a split can be done anywhere between the extreme values of the feature in the classes and the impurity measures stay the same.
- Look again at the Iris data and the classes *setosa* and *versicolor*: