Introduction to statistical learning

3.6 Tree based methods

V. Lefieux

June 2018

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Table of contents

Preamble: performance of a binary classifier

Principles

Classification tree

Regression tree

Stop growing a tree

Advantages and disadvantages of trees

Models averaging

Table of contents

Preamble: performance of a binary classifier

Principles

Classification tree

Regression tree

Stop growing a tree

Advantages and disadvantages of trees

Models averaging

Table of confusion 1/2

		Predictive condition	
		Positive	Negative
True condition	Positive	True positive (TP)	False negative (FN)
	Negative	False positive (FP)	True negative (TN)

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Table of confusion 2/2

The sensitivity, or true positive rate (TPR), is:

$$TPR = \frac{TP}{TP + FN} .$$

The specificity (SPC), or the true negative rate (TNR), is:

$$SPC = \frac{TN}{FP + TN} .$$

The precision, or the positive predictive value (PPV), is:

$$PPV = \frac{TP}{TP + FP} .$$

The accuracy (ACC) is:

$$ACC = \frac{TP + TN}{TP + FN + FP + TN} .$$

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Receiver Operating Characteristic

The ROC (Receiver Operating Characteristic) curve plots the sensitivity (TPR) in function of the specificity (FPR) for different decision threshold.

The AUC (Area Under the ROC) measures how well a decision rule can classify:

- $AUC = \frac{1}{2}$: worse case,
- $AUC = 1$: better case.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

R libraries

- ROCR.
- PROC.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Table of contents

Preamble: performance of a binary classifier

Principles

Classification tree

Regression tree

Stop growing a tree

Advantages and disadvantages of trees

Models averaging

Goal

- To:
  - Forecast a continuous variable $Y \in \mathbb{R}$.
  - Classify a categorical variable Y with K classes $\{1, \ldots, K\}$.
- Based on p predictors $(X_1, \ldots, X_p)$

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Principles

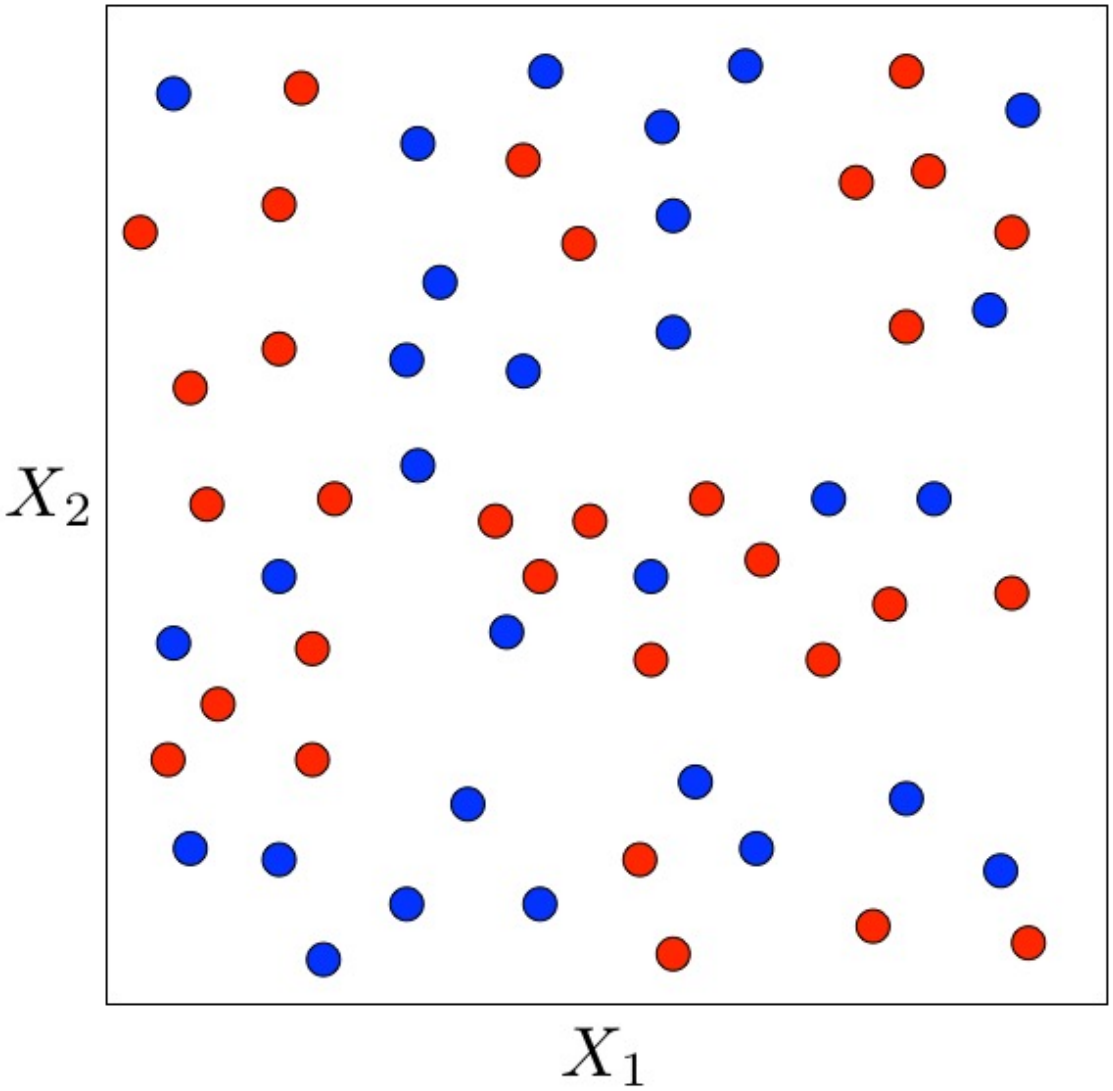
A Classification And Regression Tree, CART (Breiman et al., 1984), is a recursive method:

- At the root is the whole sample.
- Each node of the tree separates the sample into 2 branches, according to a discrete, continuous or ordinal variable (threshold) or a nominal variable (set of categories).
- A terminal node is called leaf.

We obtain a partition of the feature space into rectangles (recursive binary partitions), and then fit a simple model (average, majority) in each rectangle.

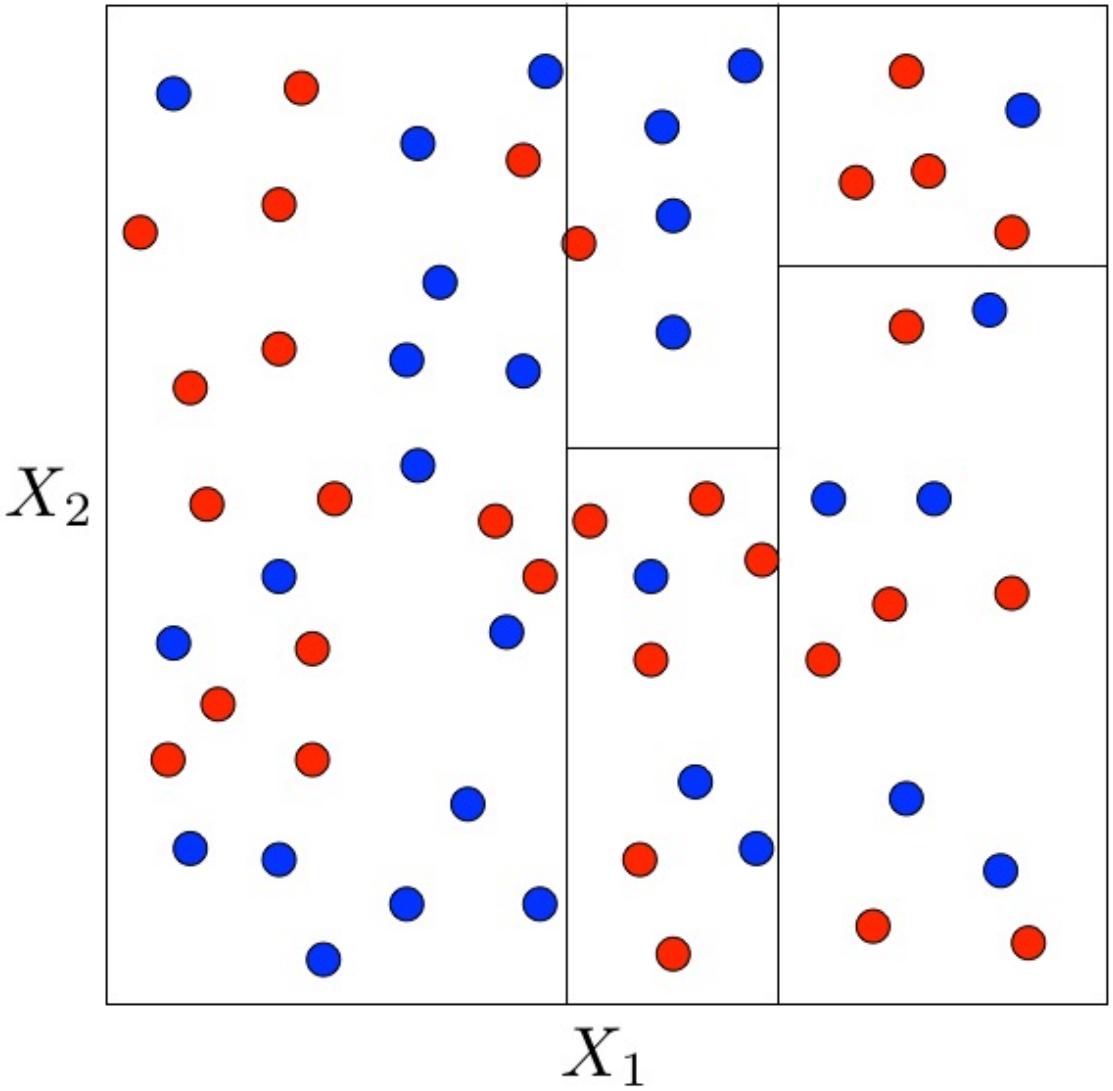
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 1/15



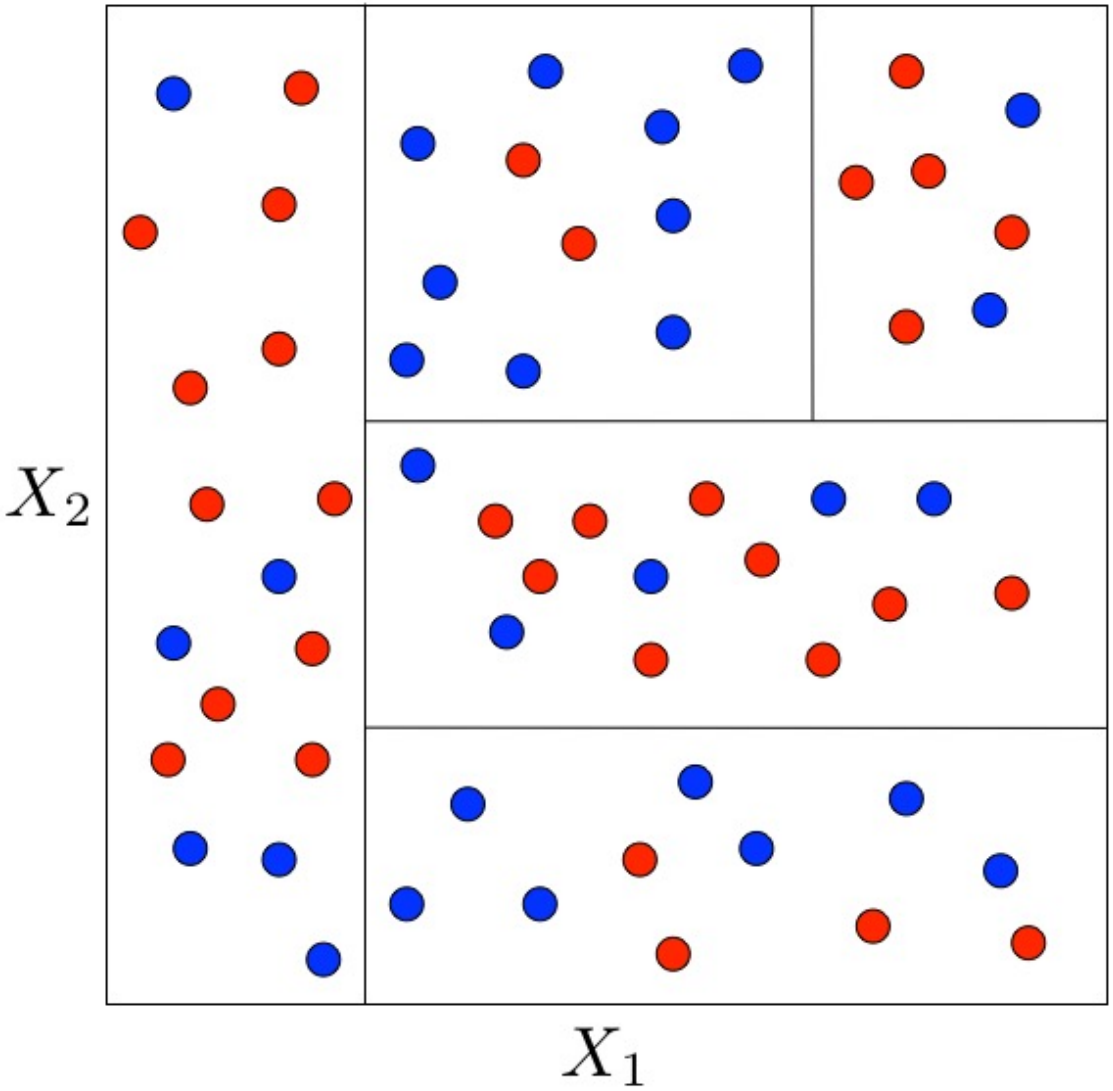
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 2/15



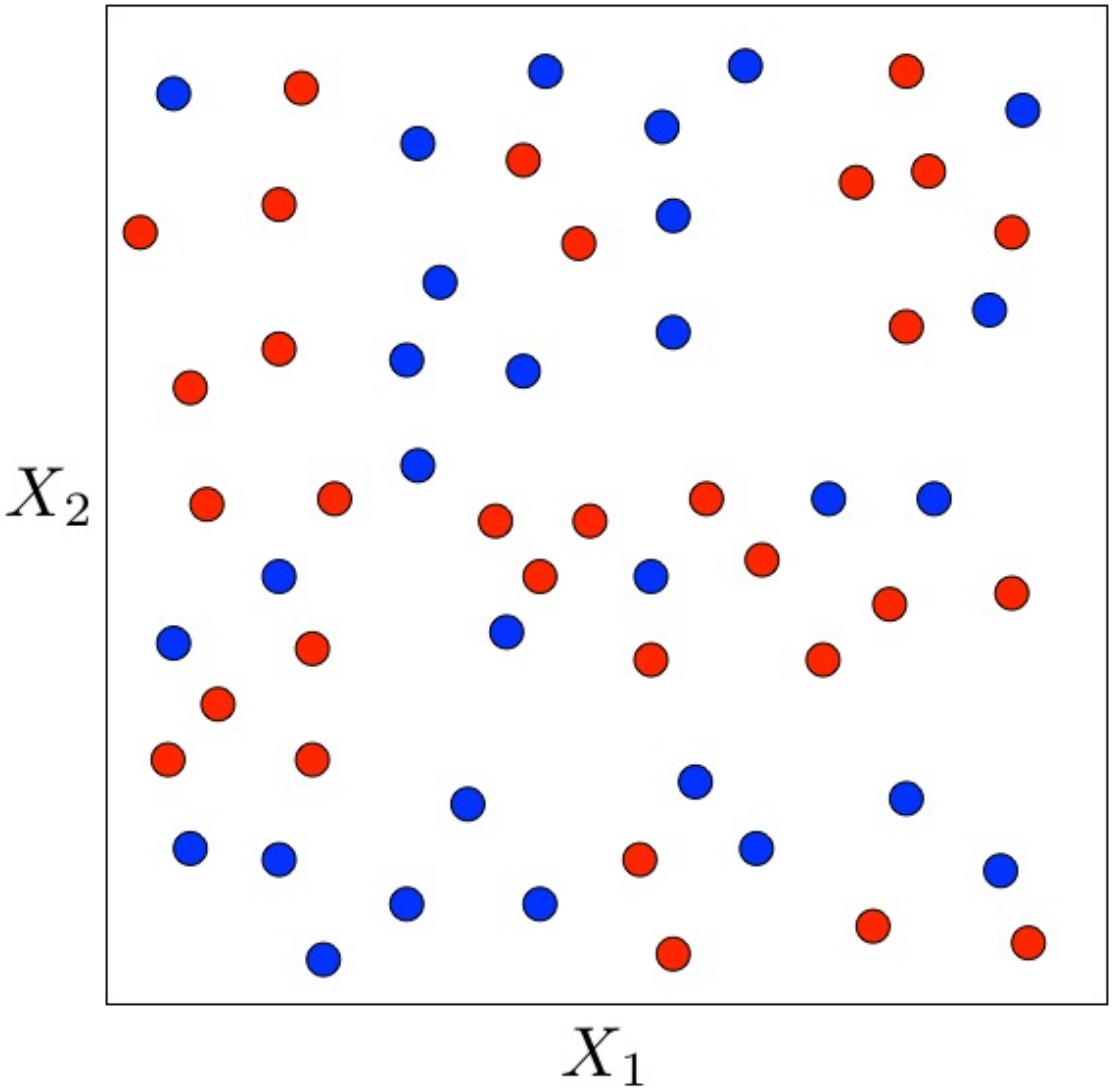
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 3/15



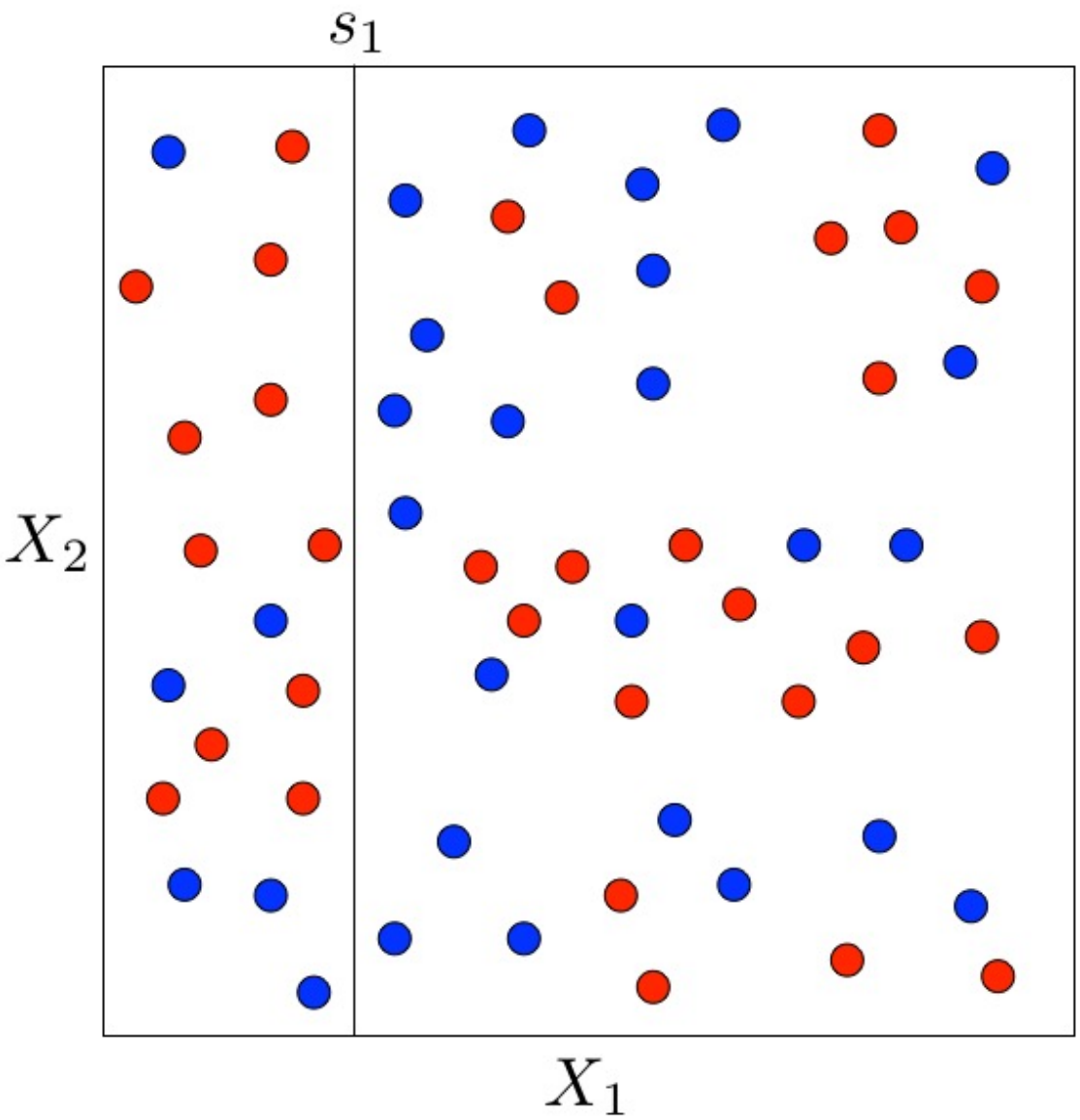
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 4/15



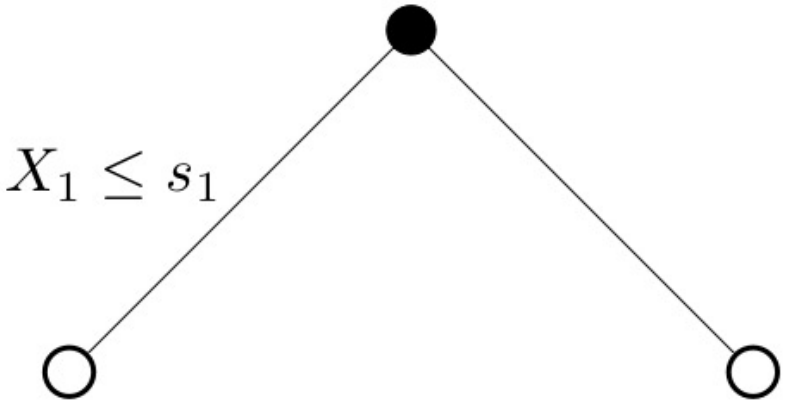
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 5/15



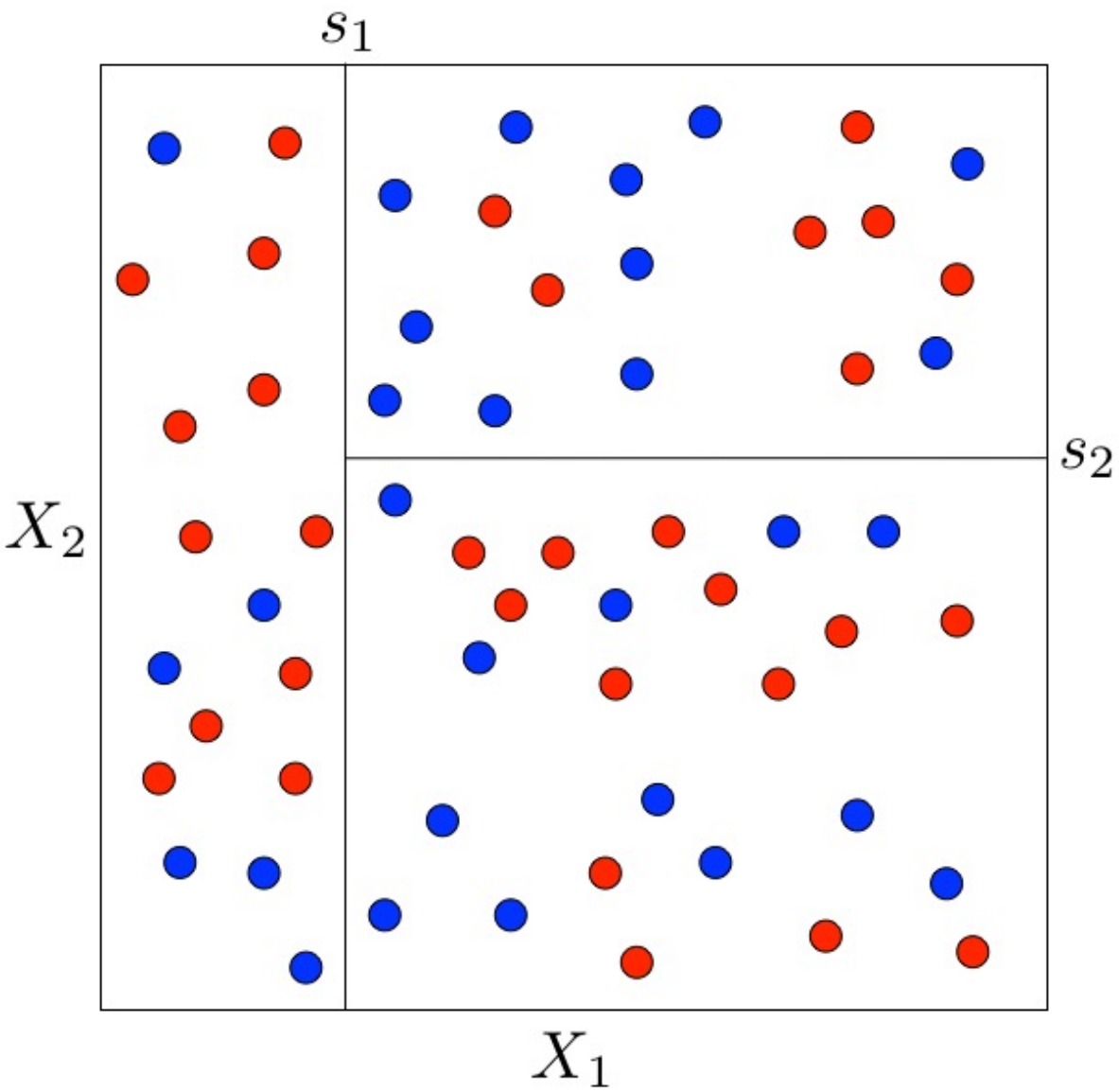
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 6/15



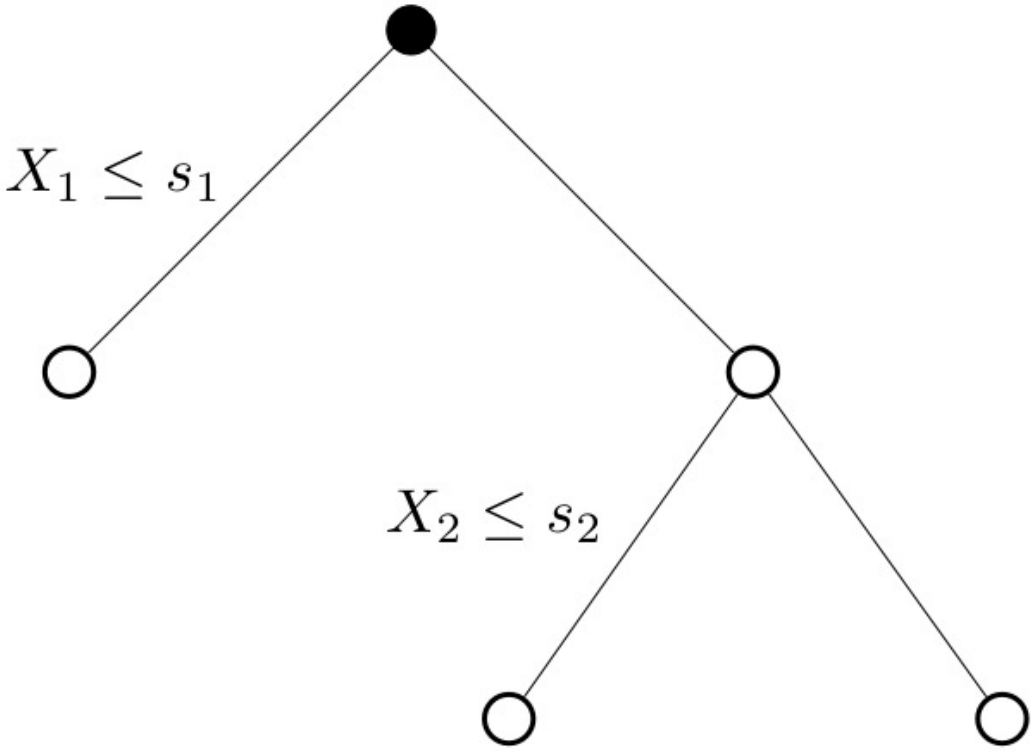
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 7/15



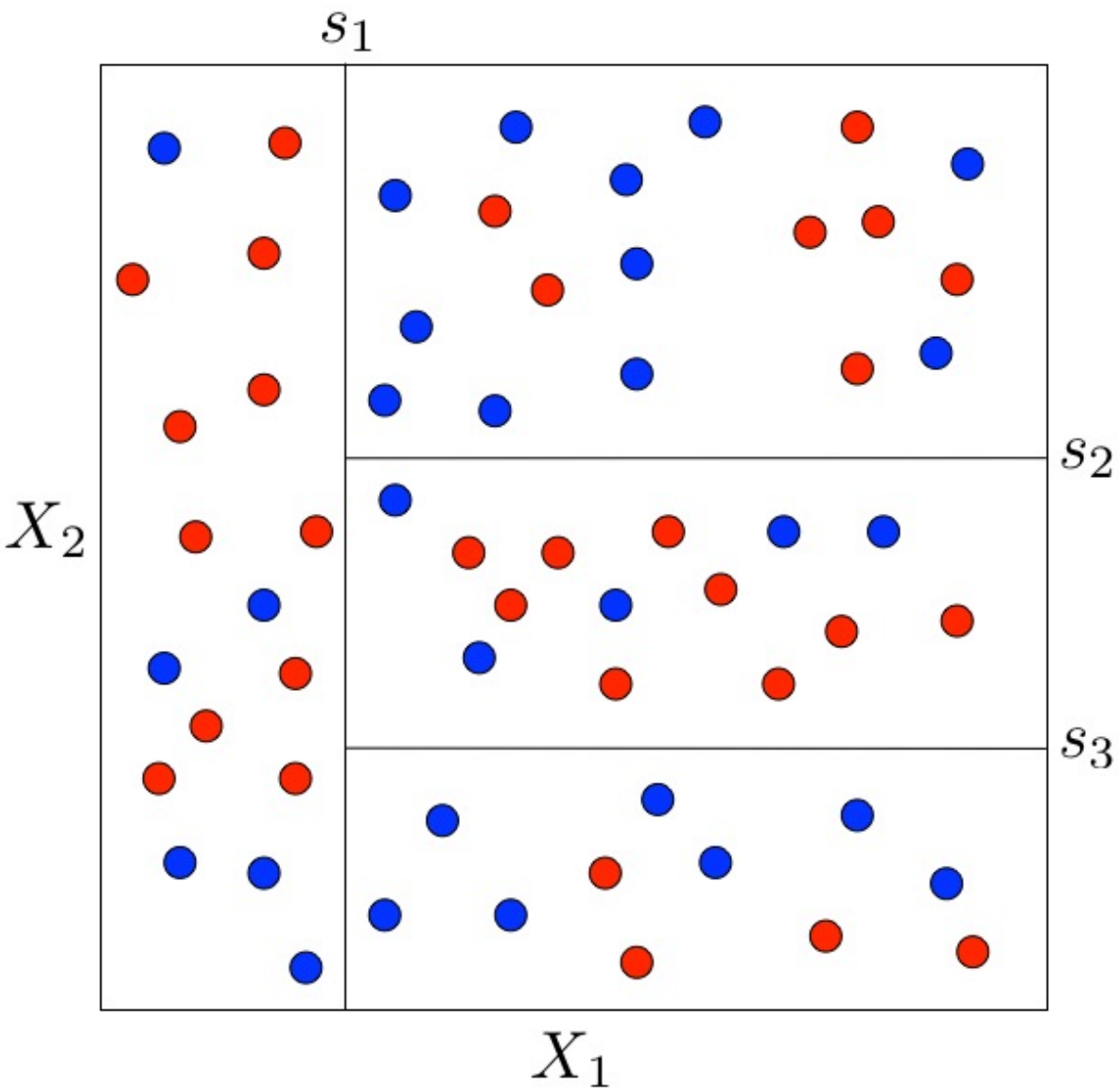
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 8/15



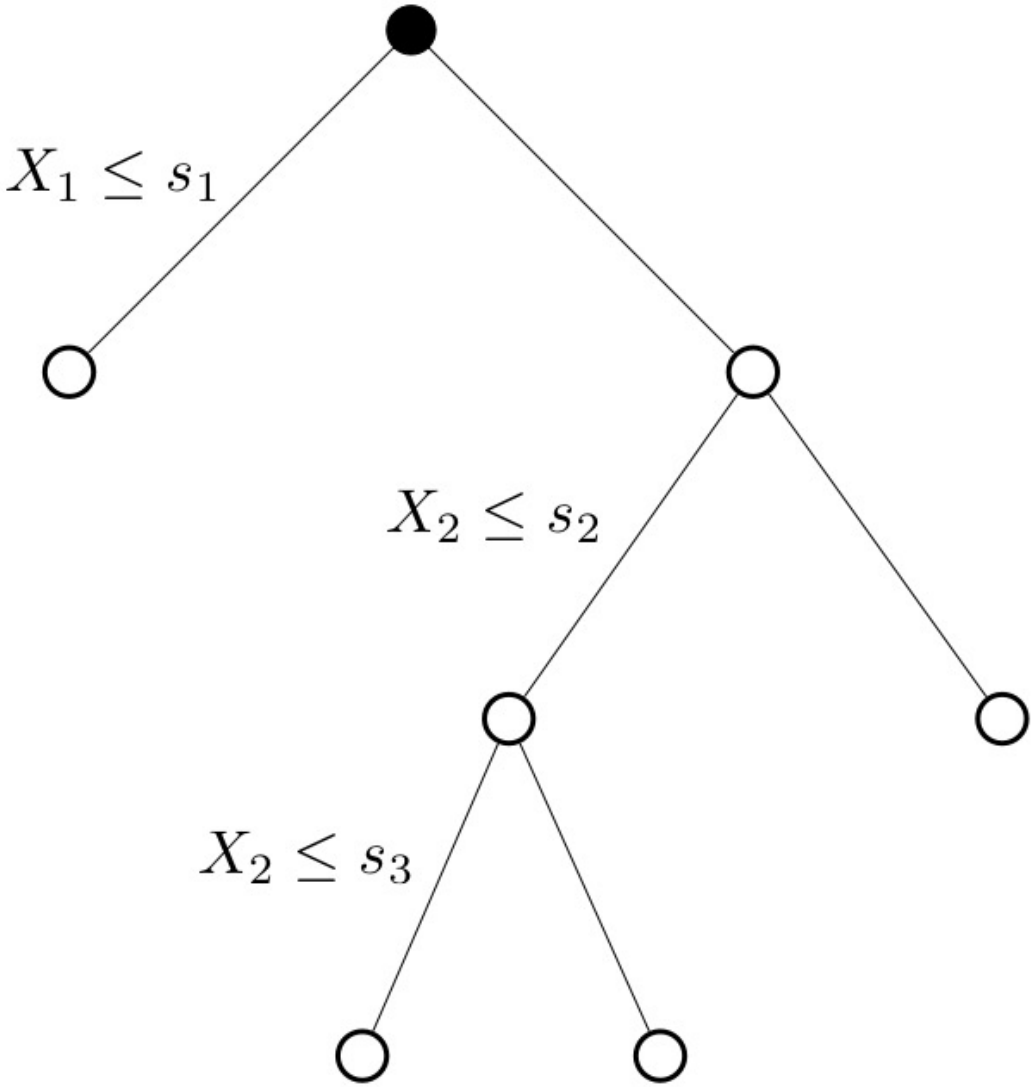
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 9/15



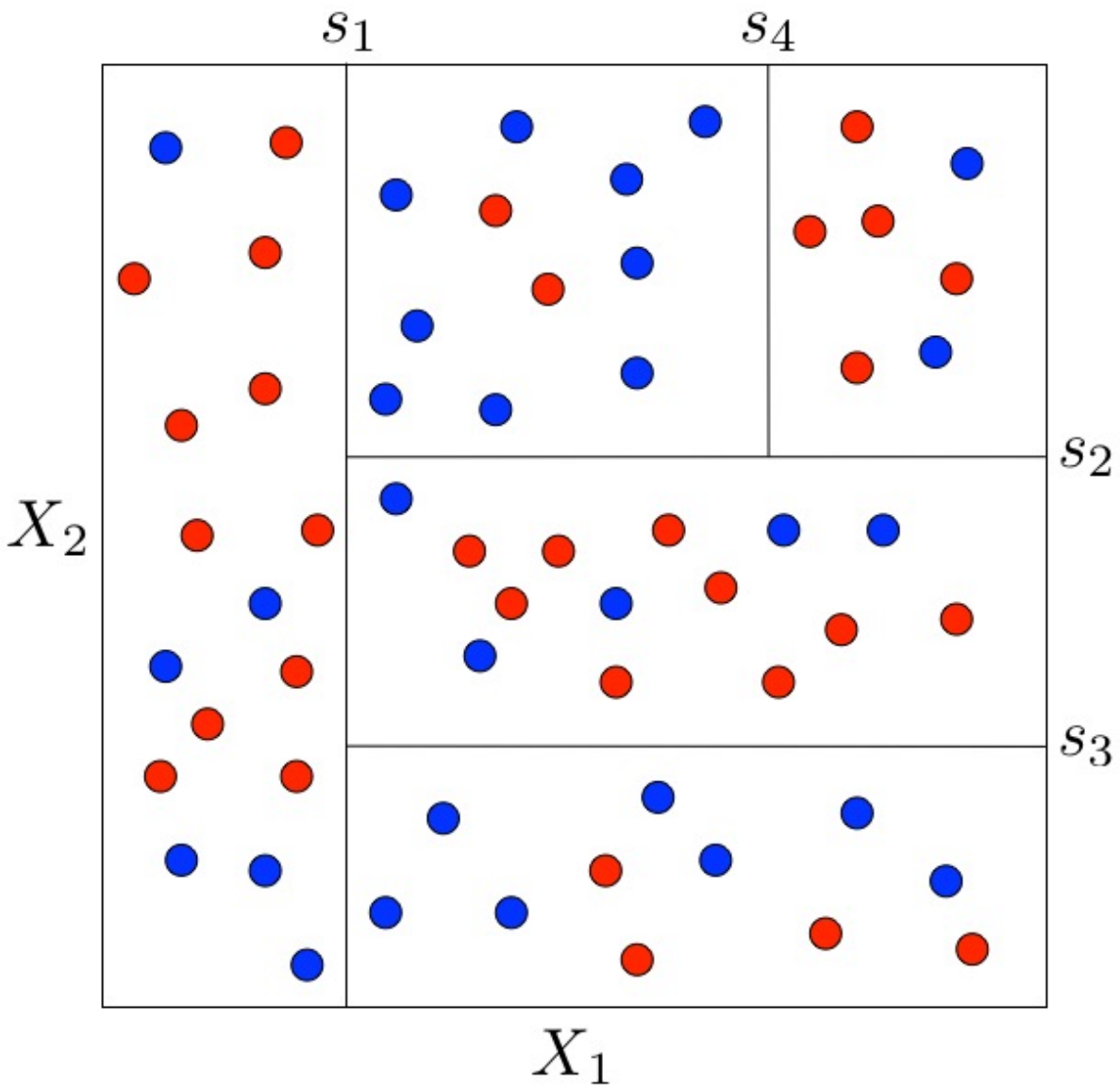
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 10/15



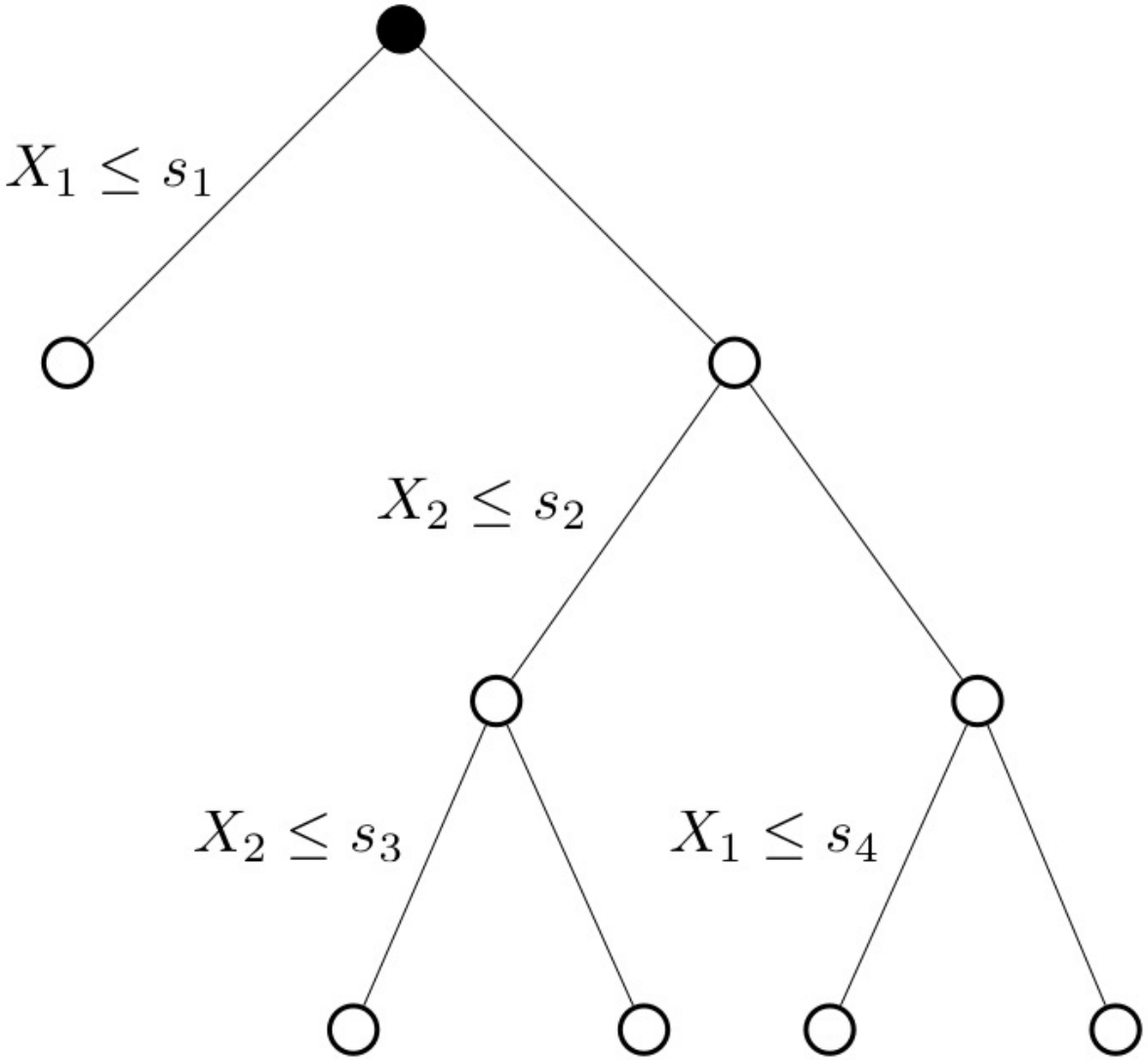
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 11/15



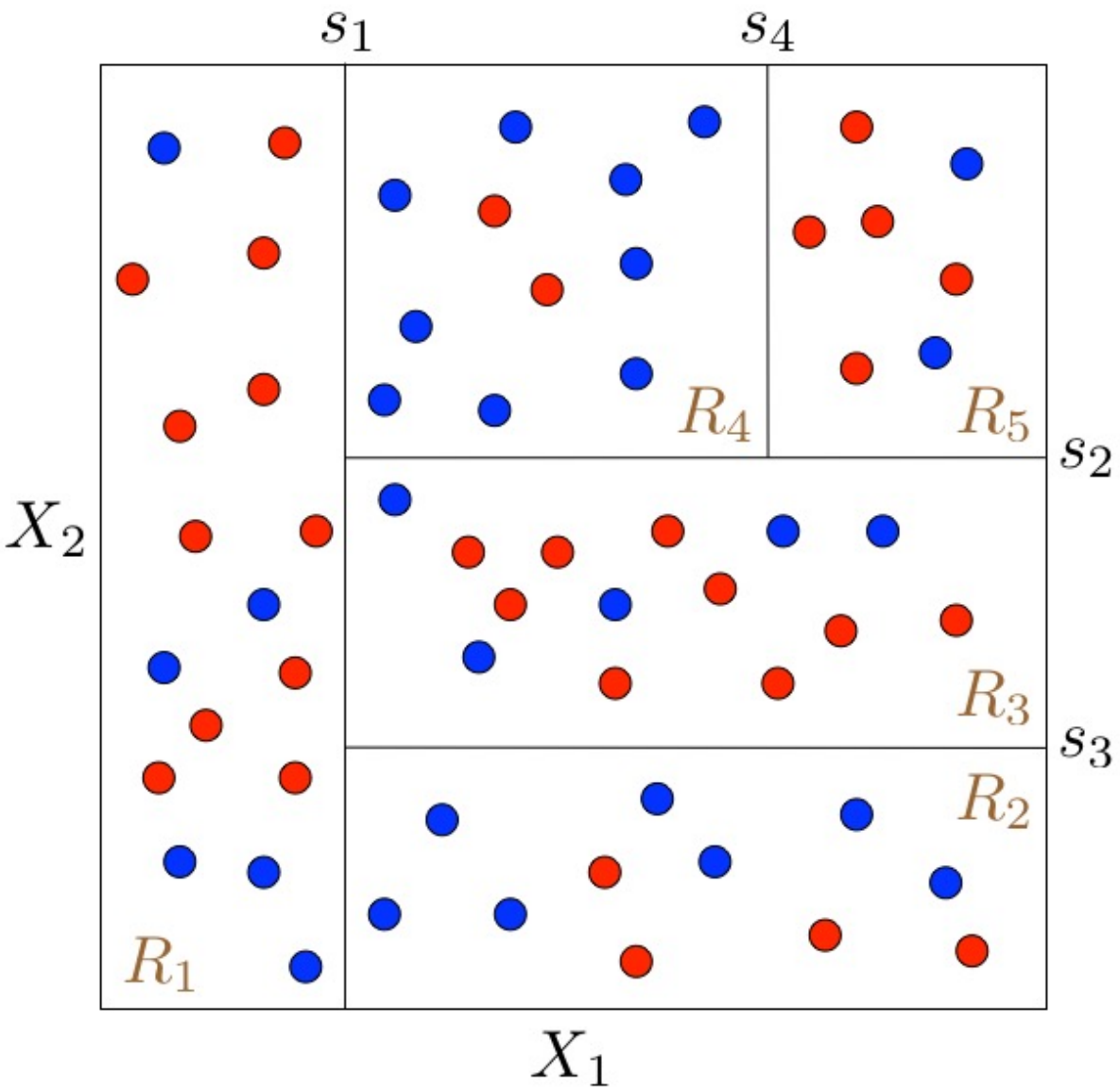
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 12/15



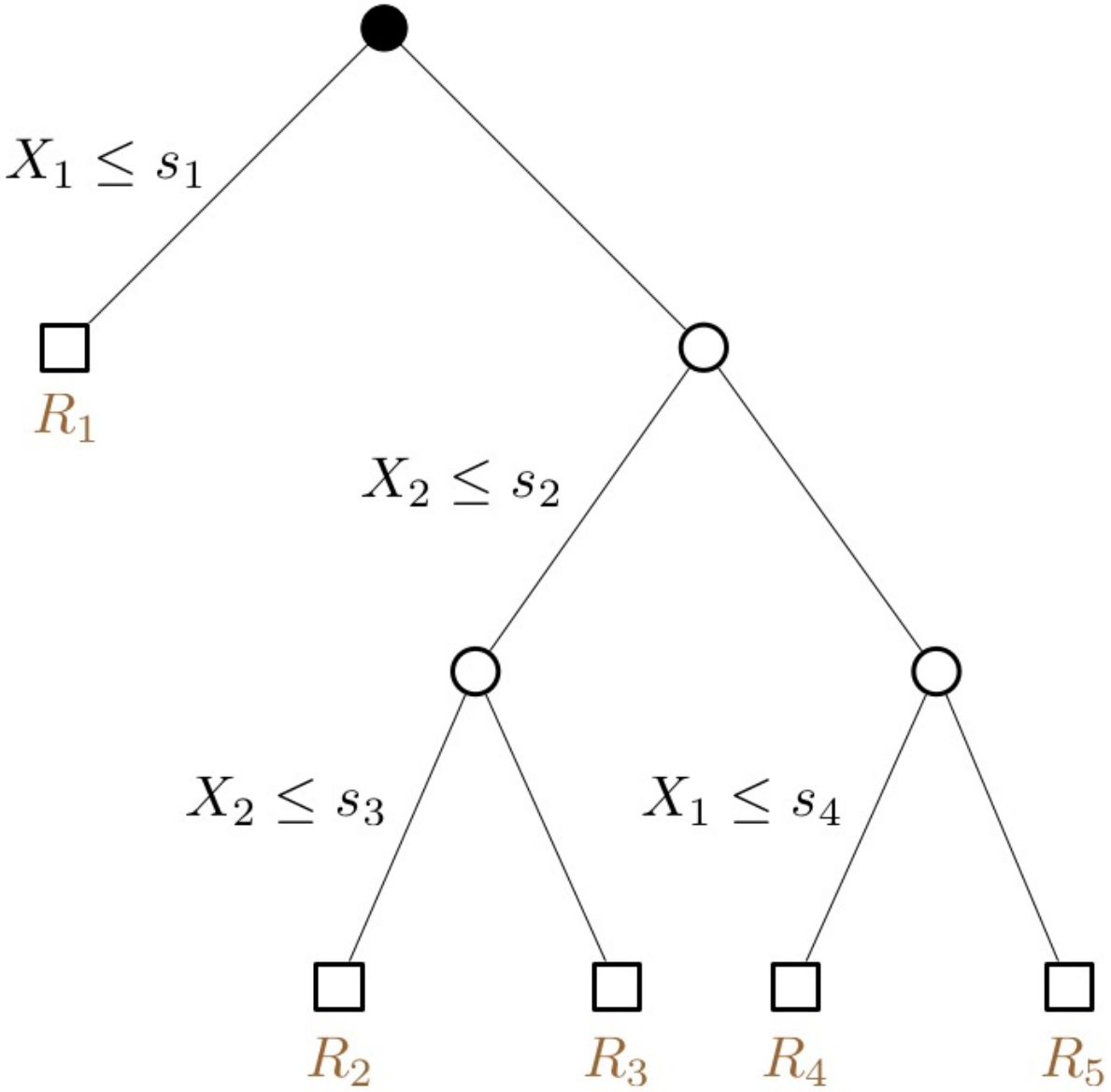
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 13/15



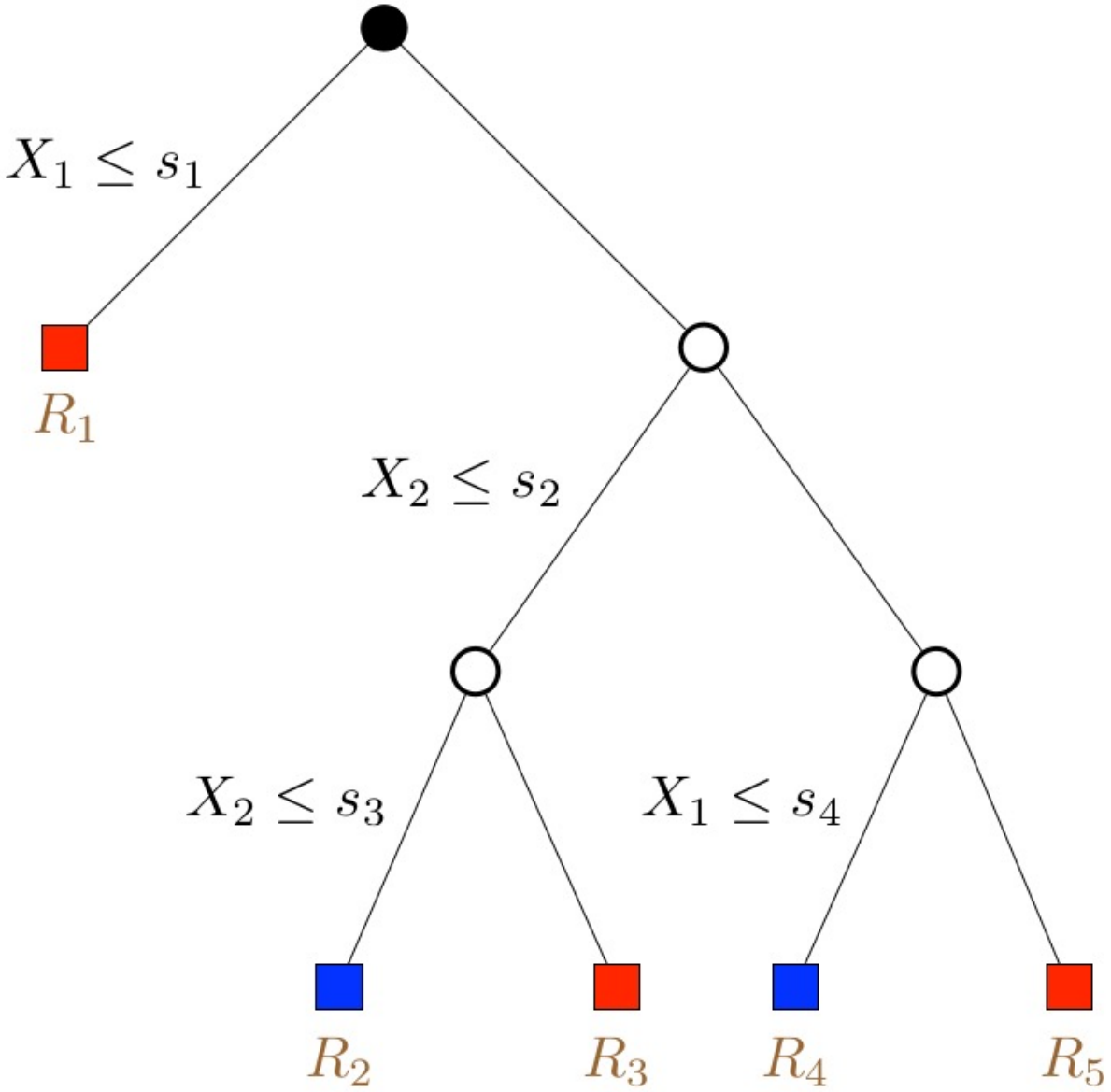
Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 14/15



Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Example (Classification) 15/15



Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Many possible splits

- For discrete and continuous variables: $\ell - 1$ where $\ell \leq n$ is the number of unique values.
- For ordinal variables: $K - 1$ where K is the number of labels.
- For nominal variables: $2^K - 1$ where K is the number of labels.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Warning for categorical variables

The algorithm tends to favor variables with many categories.

It is recommended to reduce this number by merging certain categories.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Technical issues

- Choose the best split for each variable.
- Choose the best variable to split.
- Decide that a node is a leaf.
- Fit a simple model in each rectangle.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Overfitting

- A very large tree might overfit the data.
- A too small tree might not capture the underlying structure.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Data

Let $(x_i, y_i)_{i \in \{1,\ldots,n\}}$ be n observations of (X, Y), where:

$$X = (X_1, \ldots, X_p)^\top \ ,$$
$$x_i = (x_{i1}, \ldots, x_{ip})^\top \ .$$

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Table of contents

Preamble: performance of a binary classifier

Principles

Classification tree

Regression tree

Stop growing a tree

Advantages and disadvantages of trees

Models averaging

Goal

- Classify a categorical variable Y with K classes:

$$\{1, \ldots, K\} \ .$$

- Based on p predictors:

$$(X_1, \ldots, X_p) \ .$$

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Region forecast

Let consider a partition into M regions $\{R_1, \ldots, R_M\}$.

Let C_m be the class of the m-th region.
For $k \in \{1, \ldots, K\}$, we estimate $\mathbb{P}(C_m = k)$ by :

$$\widehat{p}_k^m = \frac{1}{\text{Card}\{x_i \in R_m\}} \sum_{x_i \in R_m} \mathbb{1}_{y_i = k} \ .$$

The predicted class for the m-th region is the most important class among points in the region:

$$\widehat{c}_m = \arg \max_{k \in \{1, \ldots, K\}} \widehat{p}_k^m \ .$$

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Splitting criterion

To find the best binary partition with a sum of squares criterion, we use a "greedy" algorithm.

Let consider a binary partition for the j-th predictor and a split point s:

$$R_1(j,s) = \{X / X_j \leq s\} \ ,$$
$$R_2(j,s) = \{X / X_j > s\} \ ,$$

We choose the splitting variable j and the split point s that solve (minimization of the missclassification errors):

$$\min_{j \in \{1,\ldots,p\}} \min_{s} \left[\left(1 - \widehat{p}^m_{\widehat{c}_1}\right) + \left(1 - \widehat{p}^m_{\widehat{c}_2}\right) \right] \ .$$

Once we have found this optimal split, we repeat the splitting step on the two regions obtained, and so on...

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Impurity function

We have used an impurity function i: the missclassification error.

More generally, we consider functions such that:

- i is minimal, and equal to 0, for configurations with only one class:

$$\begin{gathered}(1,0,0,\ldots,0)\\(0,1,0,\ldots,0)\\\vdots\\(1,0,0,\ldots,0)\ .\end{gathered}$$

- i is maximal for the configuration:

$$\forall i \in \{1,\ldots,K\} : p_i = \frac{1}{K}\ .$$

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Examples of impurity functions

For a region R_m:

- Misclassification error:

$$i(R_m) = 1 - \widehat{p}^m_{\widehat{c}_m} \ .$$

- Gini index:

$$i(R_m) = \sum_{k=1}^{K} \widehat{p}^m_k \left(1 - \widehat{p}^m_k\right) \ .$$

- Cross-entropy:

$$i(R_m) = -\sum_{k=1}^{K} \widehat{p}^m_k \ \ln\left(\widehat{p}^m_k\right) \ .$$

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Table of contents

Preamble: performance of a binary classifier

Principles

Classification tree

Regression tree

Stop growing a tree

Advantages and disadvantages of trees

Models averaging

Region forecast

Let consider a partition into M regions $\{R_1, \ldots, R_M\}$.
Consider the regression function (a constant in each region):

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

If we consider a sum of squares criterion, the predicted regression function is:

$$\widehat{m}(x) = \sum_{m=1}^{M} \widehat{c}_m \, \mathbb{1}_{x \in R_m} \ .$$

where:

$$\widehat{c}_m = \frac{1}{\text{Card}\,\{x_i \in R_m\}} \sum_{x_i \in R_m} y_i \ .$$

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Splitting criterion

To find the best binary partition with a sum of squares criterion, we use a "greedy" algorithm.

Let consider a binary partition for the j-th predictor and a split point s:

$$R_1(j,s) = \{X / X_j \leq s\} \ ,$$
$$R_2(j,s) = \{X / X_j > s\} \ ,$$

We choose the splitting variable j and the split point s that solve:

$$\min_{j \in \{1,\ldots,p\}} \min_{s} \left(\sum_{x_i \in R_1(j,s)} (y_i - \widehat{c}_1)^2 + \sum_{x_i \in R_2(j,s)} (y_i - \widehat{c}_2)^2 \right) \ .$$

Once we have found this optimal split, we repeat the splitting step on the two regions obtained, and so on...

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Impurity function

We have used here the following impurity function:

$$i\left(R_m\right) = \sum_{x_i \in R_m(j,s)} \left(y_i - \widehat{c}_m\right)^2$$

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Table of contents

Preamble: performance of a binary classifier

Principles

Classification tree

Regression tree

Stop growing a tree

Advantages and disadvantages of trees

Models averaging

Som criteria

- Reduction of impurity smaller than a threshold.
- Number of points lower than a threshold (e.g 5).
- Hypothesis testing.
- Fully grow and prune.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Pruning

(Breiman et al., 1984) proposed to:

- Grow a very large tree using forward selection.
  At each step, find the best split and stop when all terminal nodes have less than a fixed number of points (commonly between 1 and 5) or have almost the same outcomes.

- Create a nested sequence of trees, with decreasing complexity.

- Prune back unnecessary splits (find the optimal tree).

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Complexity cost

For a tree T, with a number of leaves $|T|$, we define:

$$C(T) = \sum_{i=1}^{|T|} i\,(R_i) + \alpha\,|T| \ .$$

Let T_0 be the maximum tree.

To seek the tree $T \subset T_0$ that minimizes C, we recursively prune the weakest leaf (in the sense of i) to find this optimal tree.

In CART, the tuning parameter α is chosen by cross validation (5-fold or 10-fold).

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

p-fold cross-validation

- Divide the data set at random into p regions $(R_1, \ldots, R_p)$.
- For $i \in \{1, \ldots, p\}$, fit the predictor on $\bigcup_{j=1, j \neq i}^{p} R_j$ and predict R_p.
- Estimate the accuracy (calculate the cross validation criterion) using the p predictions.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Table of contents

Preamble: performance of a binary classifier

Principles

Classification tree

Regression tree

Stop growing a tree

Advantages and disadvantages of trees

Models averaging

Advantages

- No distribution assumption.
- Easy to implement.
- Nice graphical representation of a set of rules
- Automatic variable selection.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Disadvantages

- Need large data sets.
- Only horizontal or vertical splits.
- No interactions between variables.
- Instability (a small change in the data set can provide a different tree).

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Table of contents

Preamble: performance of a binary classifier

Principles

Classification tree

Regression tree

Stop growing a tree

Advantages and disadvantages of trees

Models averaging

Different ideas

- Bagging: (Breiman, 1996)
  Fit many trees to bootstrap-resampled versions of the training data set, and "combine" them (average or majority vote).

- Boosting: (Schapire and Freund, 2012)
  Fit many large or small trees to reweighted versions of the training data set, and "combine" them (with weights).

- Random Forests (Breiman, 2001)
  Fit many de-correlated trees, and "combine" them.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

R libraries

- ElemStatLearn: data sets used in (Hastie et al., 2009).
- rpart: classification and regression trees.
- randomForest: random forests.

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

Web references

Leo Breiman and Adele Cutler built a software tool (*Random Forests*):
http://www.stat.berkeley.edu/~breiman/RandomForests/

See also the nice website of Andrej Karpathy, including a random forest demo in Javascript:
http://cs.stanford.edu/people/karpathy/svmjs/demo/demoforest.html

Preamble: performance of a binary classifier
Principles
Classification tree
Regression tree
Stop growing a tree
Advantages and disadvantages of trees
Models averaging
References

References

Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24:123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, 45:5–32.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. (1984). *Classification And Regression Trees*. Taylor & Francis.

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *The elements of statistical learning. Data Mining, inference, and prediction*. Springer Series in Statistics. Springer, 2 edition.

Schapire, R. E. and Freund, Y. (2012). *Boosting. Foundations and algorithms*. MIT Press.

Preamble: performance of a binary classifier

Principles

Classification tree

Regression tree

Stop growing a tree

Advantages and disadvantages of trees

Models averaging

References