Decision trees
Bagging

Vladislav Goncharneko

girafe ai

ITMO, fall 2024

Recap

Lecture 4:
SVM, PCA
kNN indexes

1. Support Vector Machine (SVM)
   - Hinge loss
   - Kernel trick
2. Dimensionality reduction and PCA
   - Problem statement
   - Singular Value Decomposition
   - Eckart–Young theorem
   - Equivalent definitions
   - Data normalization
3. k Nearest Neighbors
   - kNN indexes
   - HNSW

Outline

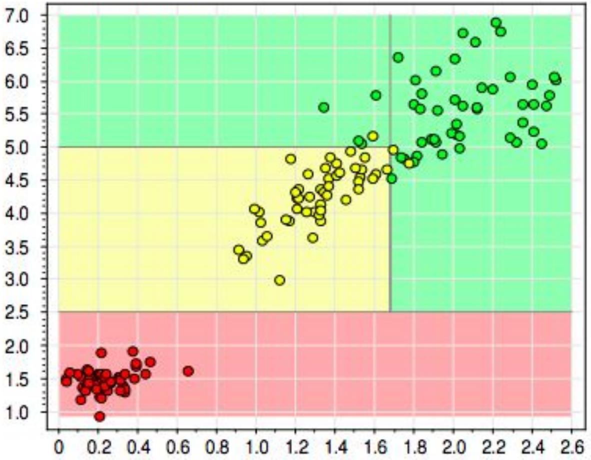
1. Intuition
2. Construction procedure
3. Information criteria
4. Special highlights
    - Dealing with missing data
    - Binarization
    - Decision tree as linear model
    - Standards
    - Hyperparameters
5. Bootstrap and Bagging
6. Random Forest

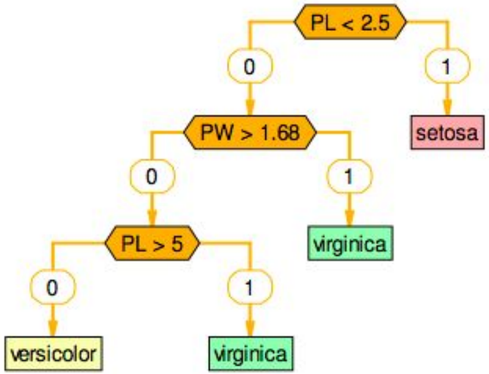
Intuition

girafe
ai

01

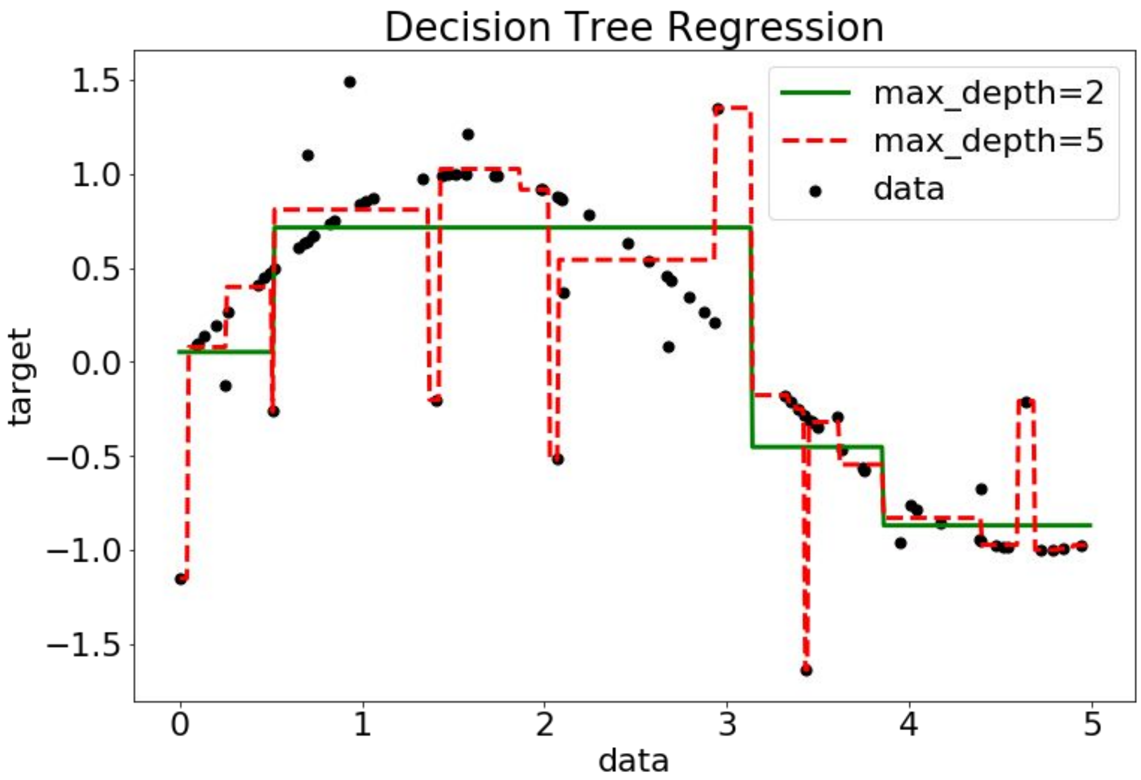
Decision tree for Iris data set





setosa	$r_1(x) = [PL \leqslant 2.5]$
virginica	$r_2(x) = [PL > 2.5] \wedge [PW > 1.68]$
virginica	$r_3(x) = [PL > 5] \wedge [PW \leqslant 1.68]$
versicolor	$r_4(x) = [PL > 2.5] \wedge [PL \leqslant 5] \wedge [PW < 1.68]$

Decision tree in regression



Green - decision tree of depth 2

Red - decision tree of depth 5

Every leaf corresponds to some constant.

Construction procedure

girafe
ai

02

Constructing decision trees

threshold value

j feature

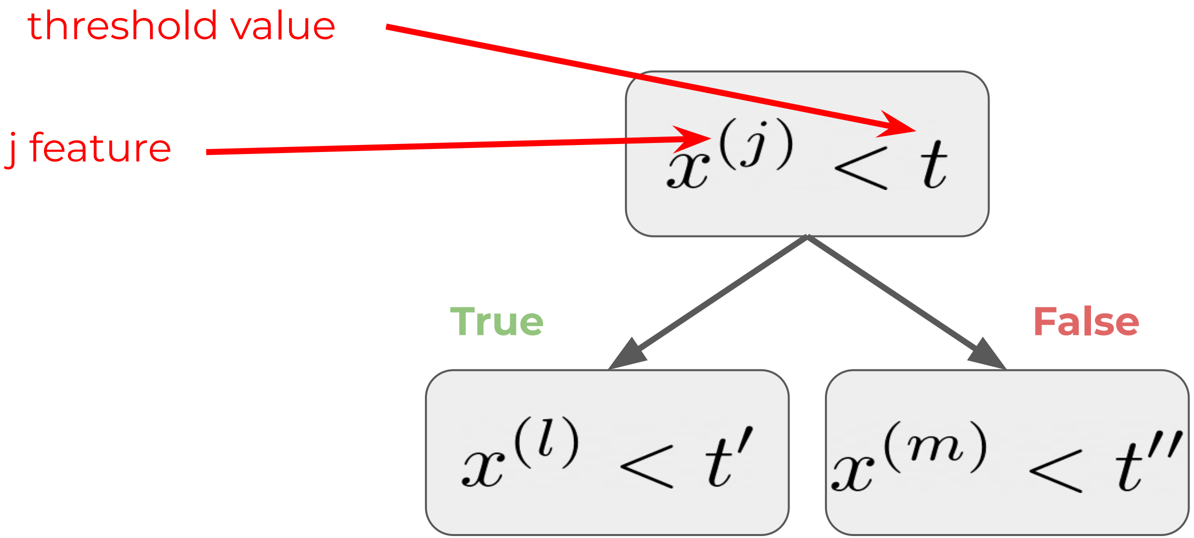
$$x^{(j)} < t$$

True

False

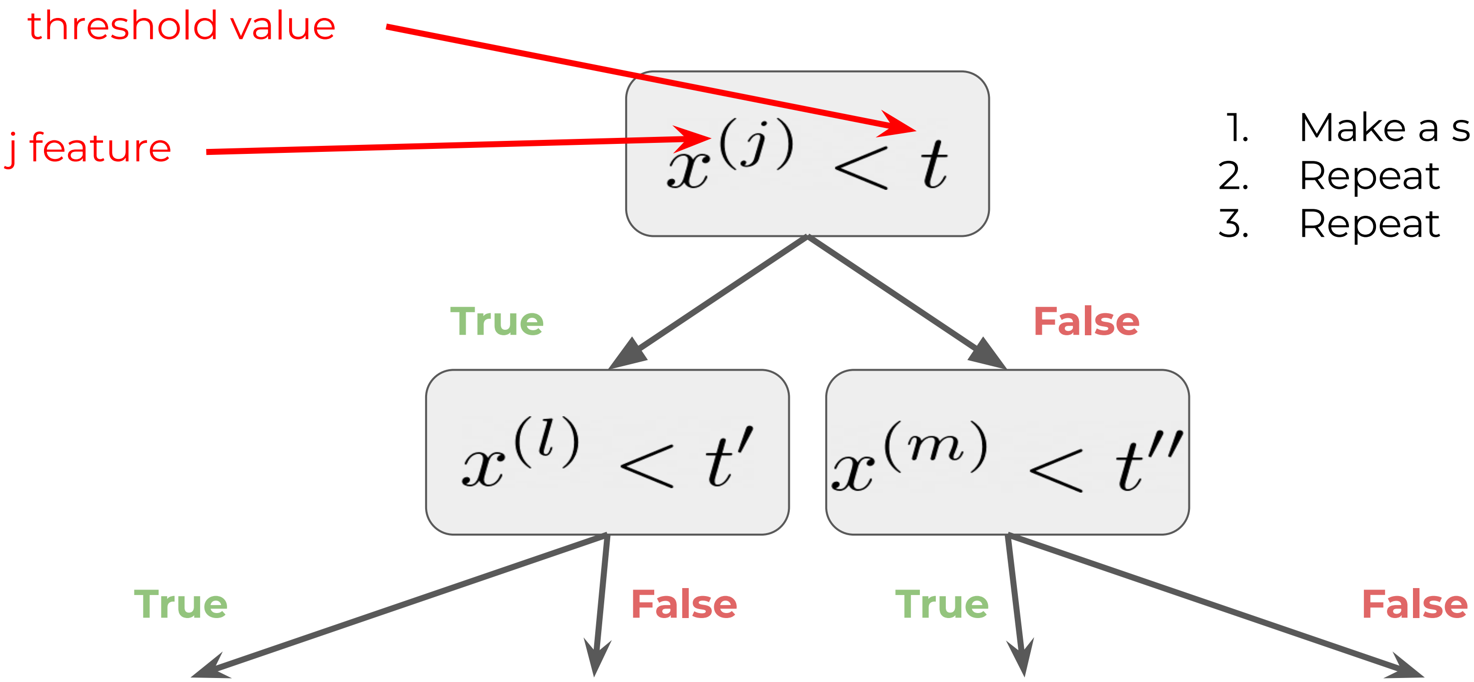
1. Make a split

Constructing decision trees



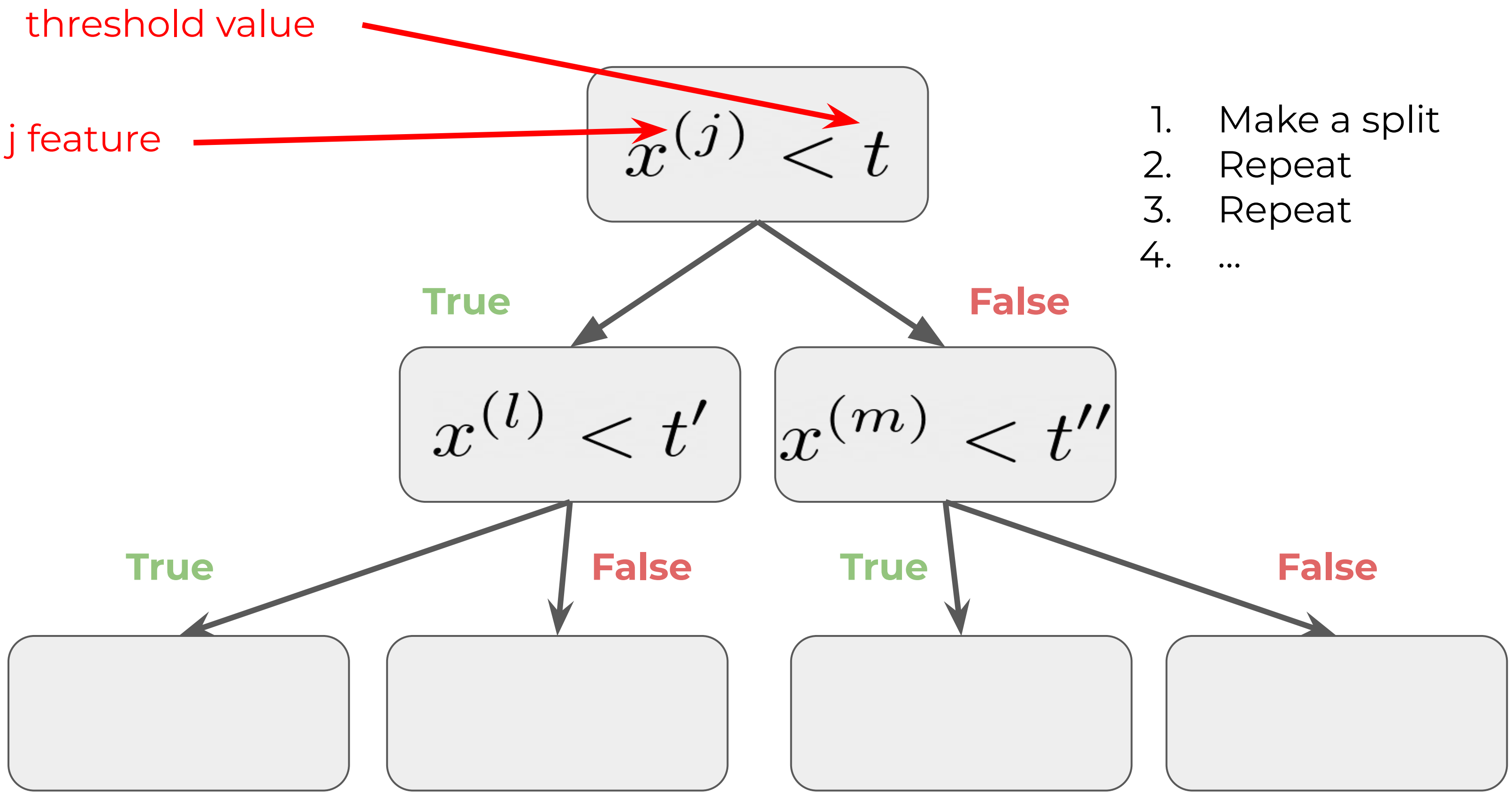
1. Make a split
2. Repeat

Constructing decision trees



1. Make a split
2. Repeat
3. Repeat

Constructing decision trees

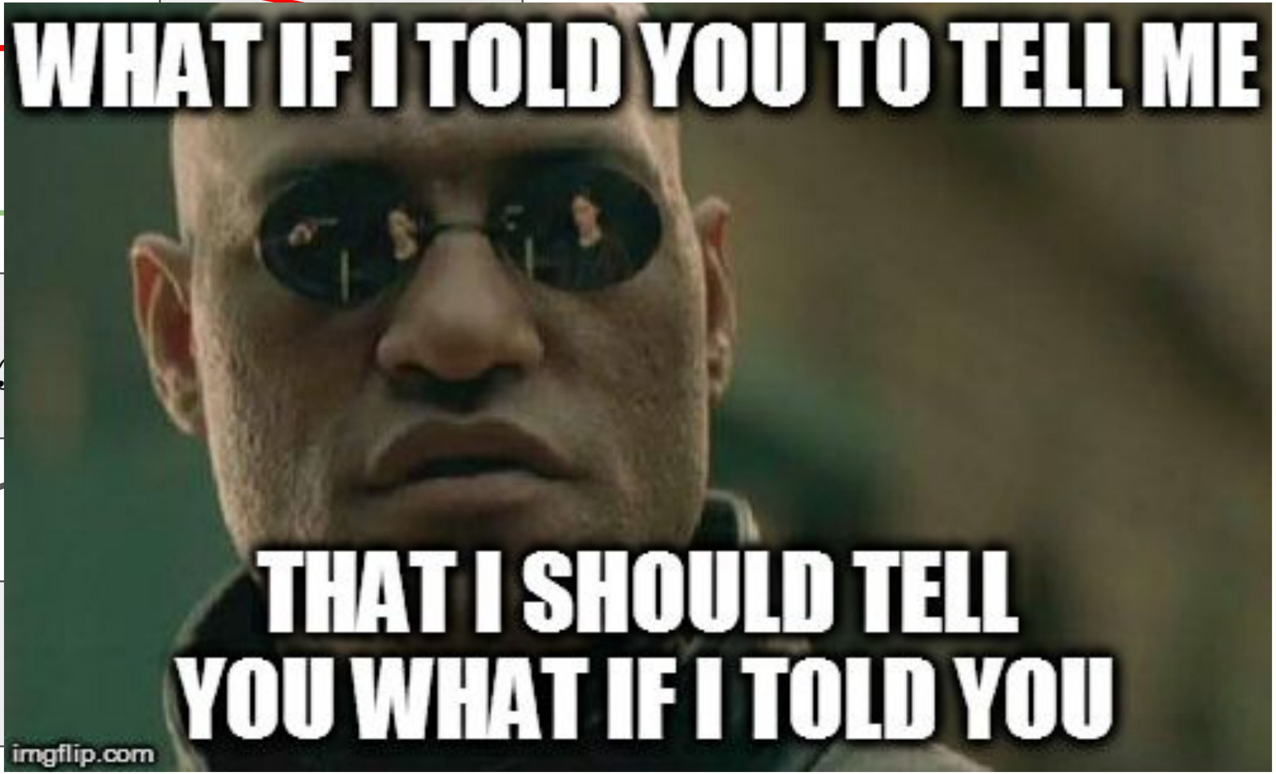


1. Make a split
2. Repeat
3. Repeat
4. …

threshold value

j feature

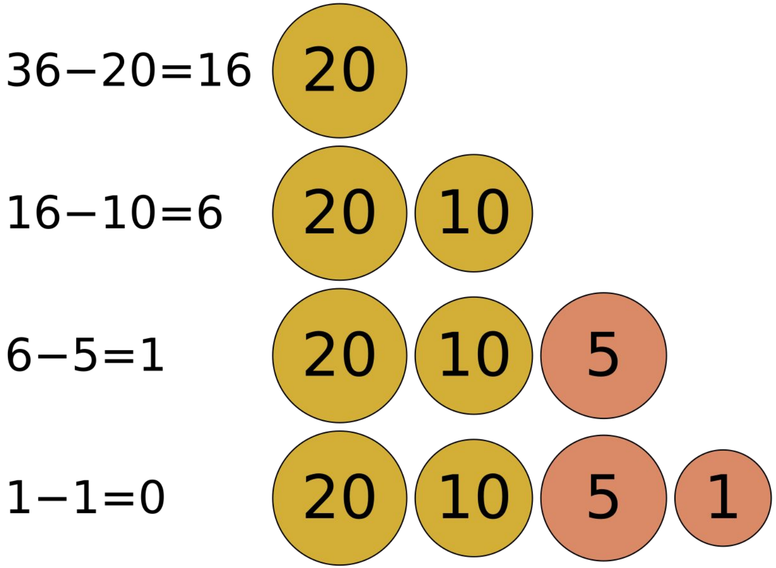
True



Greedy algorithm

A greedy algorithm is any algorithm that follows the problem-solving heuristic of making the locally optimal choice at each stage.

In many problems, a greedy strategy does not produce an optimal solution, but a greedy heuristic can yield locally optimal solutions that approximate a globally optimal solution in a reasonable amount of time.



How to answer in leaf?

Classification:

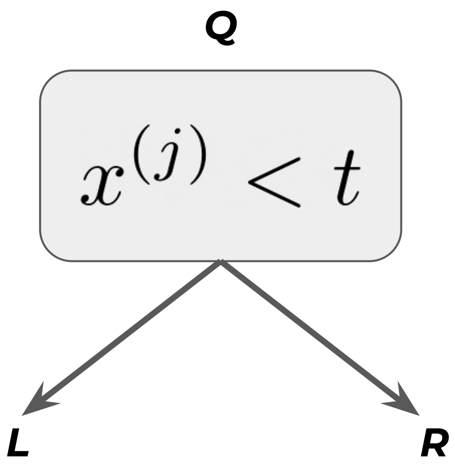
- most popular
- sample with frequencies of classes

Regression:

Depends on loss function!

- for MSE
  - average in node
- for MAE
  - median in node

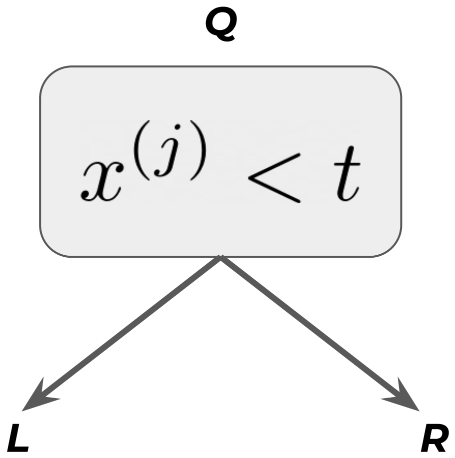
How to split data properly?



We can not use gradient this time because solution set is discrete.

So let’s apply discrete optimization!

How to split data properly?



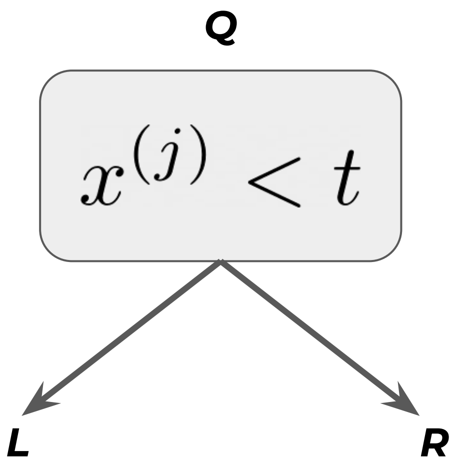
$$\frac{|L|}{|Q|}H(L) + \frac{|R|}{|Q|}H(R) \longrightarrow \min_{j,t}$$

How to choose concrete split?

Brute force algorithm will take too much time.

Random splits are chosen and compared.

How to split data properly?



What is H?

$$\frac{|L|}{|Q|}H(L) + \frac{|R|}{|Q|}H(R) \longrightarrow \min_{j,t}$$

Information criteria

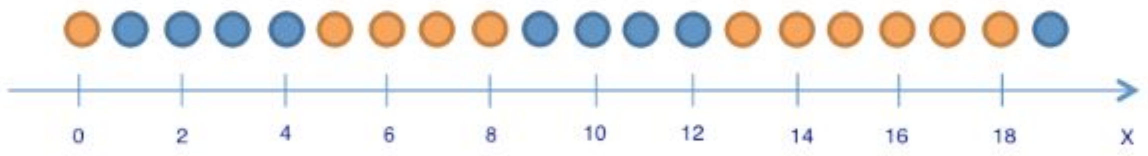
girafe ai

03

Information criteria

H(R) is measure of “heterogeneity” of our data.

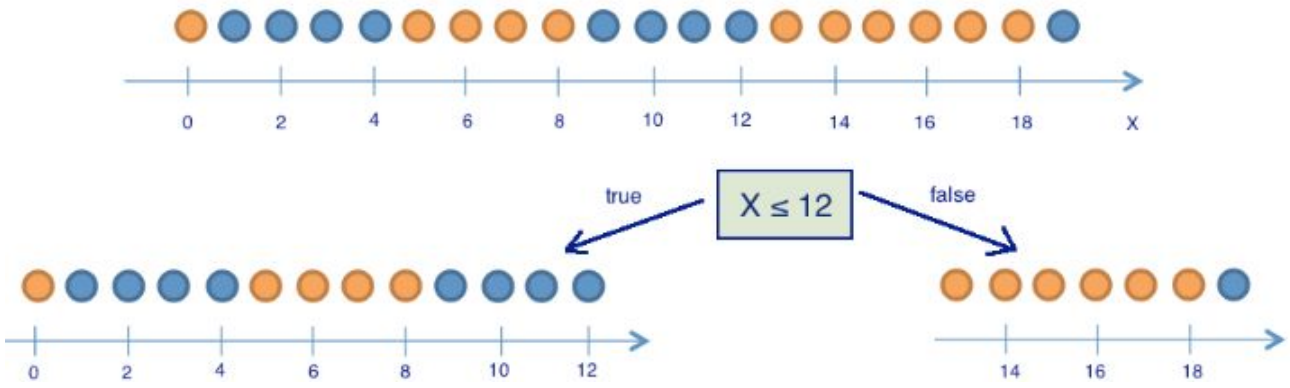
Consider binary classification problem:



Information criteria

H(R) is measure of “heterogeneity” of our data.

Consider binary classification problem:



Information criteria

H(R) is measure of “heterogeneity” of our data.

Consider binary classification problem:

Obvious way:
Misclassification criteria:

$$H(R) = 1 - \max\{p_0, p_1\}$$

1. Entropy criteria: $$H(R) = -p_0 \log p_0 - p_1 \log p_1$$

2. Gini impurity: $$H(R) = 1 - p_0^2 - p_1^2 = 2p_0p_1$$

Information criteria

H(R) is measure of “heterogeneity” of our data.

Consider multiclass classification problem:

Obvious way:
Misclassification criteria:

$$H(R) = 1 - \max_k \{p_k\}$$

1. Entropy criteria:

$$H(R) = -\sum_{k=0}^{K} p_k \log p_k$$

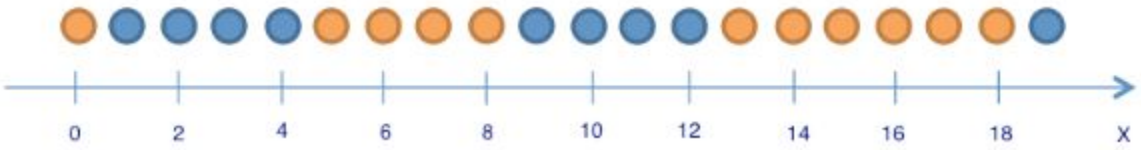
2. Gini impurity:

$$H(R) = 1 - \sum_k (p_k)^2$$

Information criteria

H(R) is measure of “heterogeneity” of our data.

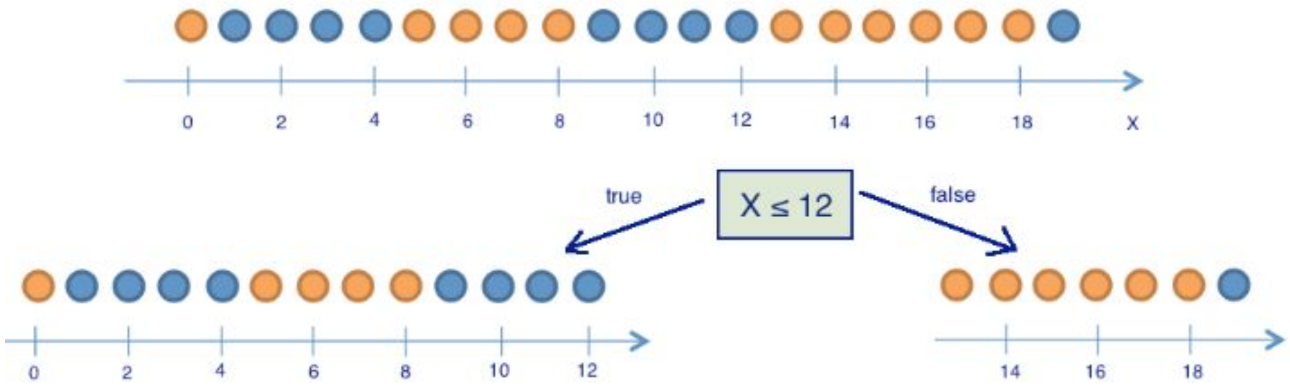
Consider binary classification problem:



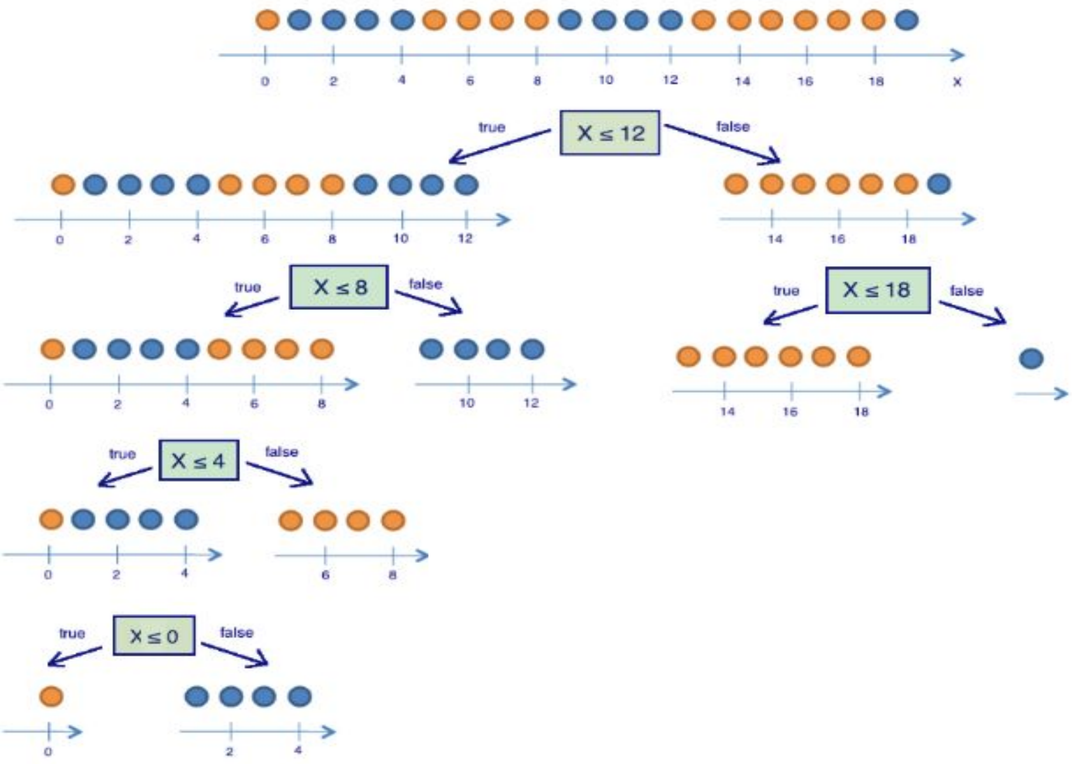
Information criteria

H(R) is measure of “heterogeneity” of our data.

Consider binary classification problem:



Information criteria: Entropy



$$S = -M\sum_{k=0}^{K} p_k \log p_k$$

In binary case N = 2

$$S = -p_+ \log_2 p_+ - p_- \log_2 p_- = -p_+ \log_2 p_+ - (1 - p_+) \log_2 (1 - p_+)$$

Information criteria: Gini impurity

$$G = 1 - \sum_k (p_k)^2$$

In binary case N = 2

$$G = 1 - p_+^2 - p_-^2 = 1 - p_+^2 - (1 - p_+)^2 = 2p_+(1 - p_+)$$

Information criteria

H(R) is measure of "heterogeneity" of our data.

Consider multiclass classification problem:

Obvious way: Misclassification criteria:

$$H(R) = 1 - \max_k \{p_k\}$$

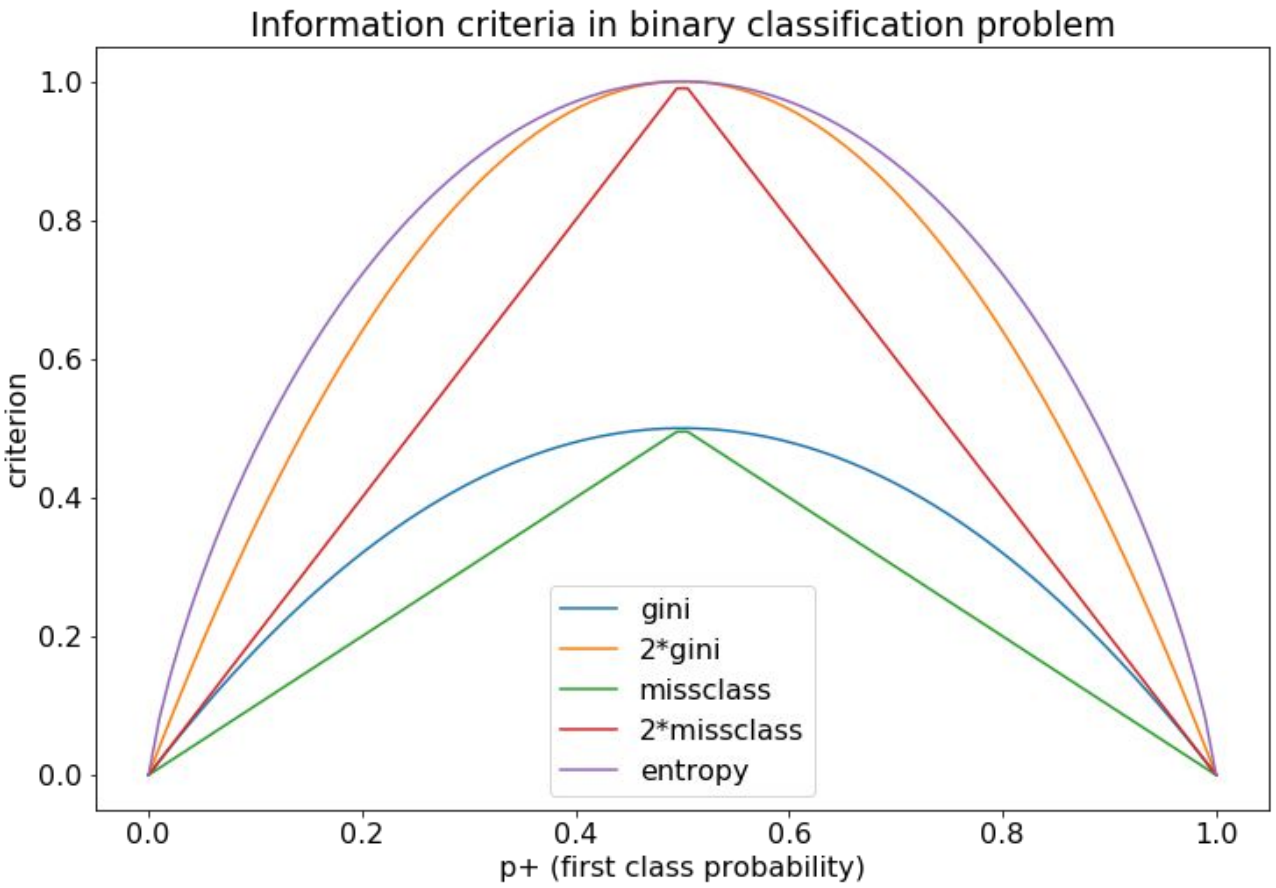
1. Entropy criteria:

$$H(R) = -\sum_k p_k \log_2 p_k$$

2. Gini impurity:

$$H(R) = 1 - \sum_k (p_k)^2$$

Information criteria



Information criteria

H(R) is measure of “heterogeneity” of our data.

Consider regression problem:

1. Mean squared error

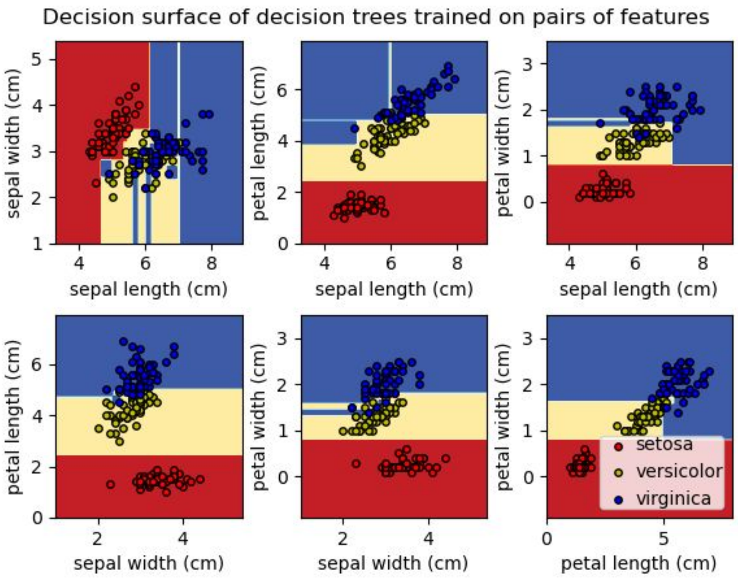
$$H(R) = \min_c \frac{1}{|R|} \sum_{(x_i, y_i) \in R} (y_i - c)^2$$

What is the constant?

$$c^* = \frac{1}{|R|} \sum_{y_i \in R} y_i$$

Hyperparameters

- max_depth: min 1
- min_samples_split: min 2
- min_samples_leaf: min 1
- min_impurity_decrease



Minor

- criterion:
  - gini, entropy, log_loss for classification
  - MSE or MAE for regression
- splitter: best, random
- max_features: sqrt, log2

As of [sklearn implementation]

Standards

- ID-3
  - Entropy criteria; Stops when no more gain available
- C4.5
  - Normalised entropy criteria; Stops depending on leaf size; Incorporates pruning
- C5.0
  - Some updates on C4.5
- CART
  - Gini criteria; Cost-complexity Pruning; Surrogate predicates for missing data;
- etc.

Read more

Special highlights

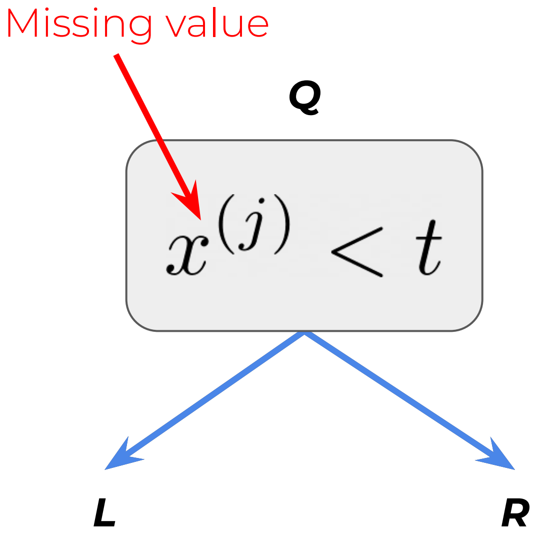
girafe
ai

04

Missing values in Decision Trees

If the value is missing, one might use both sub-trees and average their predictions.

But this will negatively affect model computational performance.



$$\hat{y} = \frac{|L|}{|Q|}\hat{y}_L + \frac{|R|}{|Q|}\hat{y}_R$$

Missing values in Catboost

Forbidden: Missing values are not supported, their presence is interpreted as an error

Min: Missing values are processed as the minimum value (less than all other values) for the feature. It is guaranteed that a split that separates missing values from all other values is considered when selecting trees.

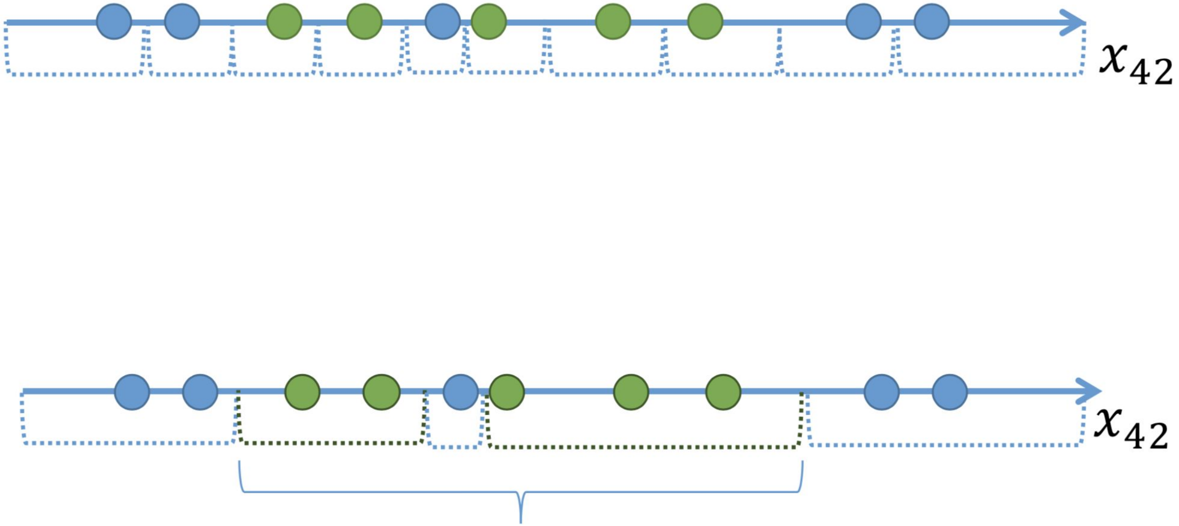
Max: Missing values are processed as the maximum value (greater than all other values) for the feature. It is guaranteed that a split that separates missing values from all other values is considered when selecting trees.

The **default** processing mode **is Min**

Documentation

Binarization

Idea: instead selecting one threshold define several for one feature.



e.g. Border count hyperparameter in Catboost (defaults to 254)

Decision Trees as Linear models

Let J be the subspace of the original feature space, corresponding to the leaf of the tree.

Prediction takes form

$$\hat{y} = \sum_j w_j [x \in J_j]$$

Bootstrap and Bagging

girafe
ai

06

Bootstrap

Consider dataset X containing m objects.

Pick m objects with return from X and repeat in N times to get N datasets.

Error of model trained on Xj: $$\varepsilon_j(x) = b_j(x) - y(x), \qquad j = 1, \dots, N,$$

Then $$\mathbb{E}_x(b_j(x) - y(x))^2 = \mathbb{E}_x \varepsilon_j^2(x).$$

The mean error of N models: $$E_1 = \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}_x \varepsilon_j^2(x).$$

Bootstrap

Consider the errors unbiased and uncorrelated:

$$\mathbb{E}_x \varepsilon_j(x) = 0;$$
$$\mathbb{E}_x \varepsilon_i(x) \varepsilon_j(x) = 0, \quad i \neq j.$$

The final model averages all predictions:

$$a(x) = \frac{1}{N} \sum_{j=1}^{N} b_j(x).$$

$$E_N = \mathbb{E}_x \left(\frac{1}{N} \sum_{j=1}^{n} b_j(x) - y(x) \right)^2 =$$
$$= \mathbb{E}_x \left(\frac{1}{N} \sum_{j=1}^{N} \varepsilon_j(x) \right)^2 =$$
$$= \frac{1}{N^2} \mathbb{E}_x \left(\sum_{j=1}^{N} \varepsilon_j^2(x) + \underbrace{\sum_{i \neq j} \varepsilon_i(x) \varepsilon_j(x)}_{=0} \right) =$$
$$\boxed{= \frac{1}{N} E_1.}$$

Error decreased by N times!

Bootstrap

Consider the errors ~~unbiased and uncorrelated~~:

This is a lie

$$\mathbb{E}_x \varepsilon_j(x) = 0;$$
$$\mathbb{E}_x \varepsilon_i(x)\varepsilon_j(x) = 0, \quad i \neq j.$$

The final model averages all predictions:

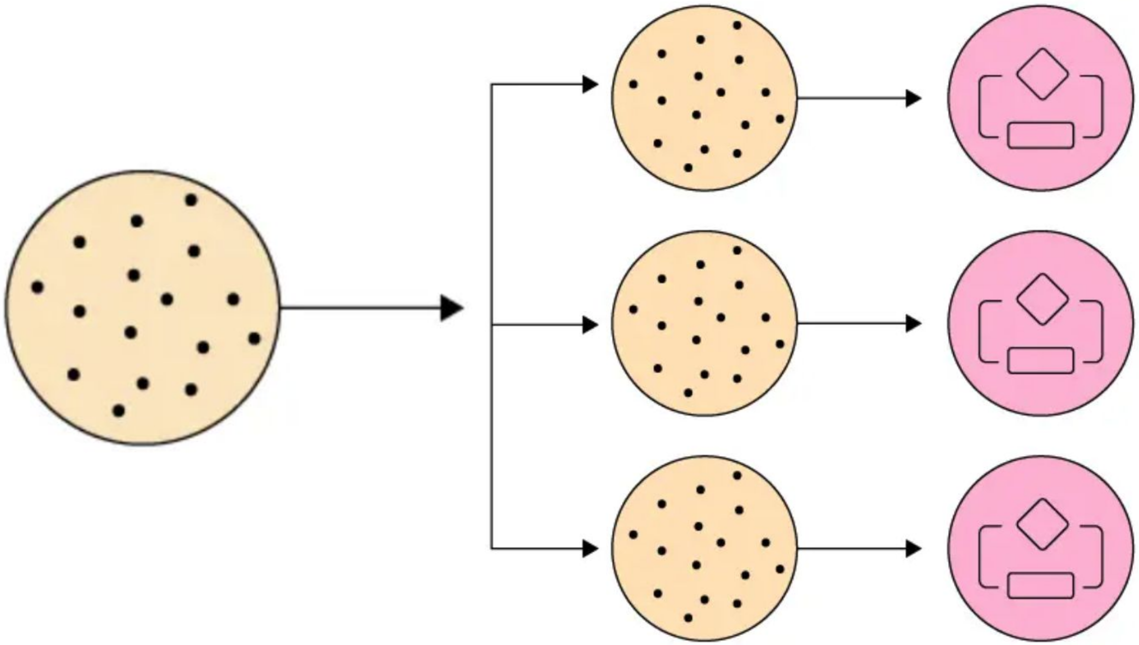
$$a(x) = \frac{1}{N}\sum_{j=1}^{N} b_j(x).$$

$$E_N = \mathbb{E}_x\left(\frac{1}{N}\sum_{j=1}^{n} b_j(x) - y(x)\right)^2 =$$
$$= \mathbb{E}_x\left(\frac{1}{N}\sum_{j=1}^{N} \varepsilon_j(x)\right)^2 =$$
$$= \frac{1}{N^2}\mathbb{E}_x\left(\sum_{j=1}^{N} \varepsilon_j^2(x) + \underbrace{\sum_{i\neq j} \varepsilon_i(x)\varepsilon_j(x)}_{=0}\right) =$$
$$\boxed{= \frac{1}{N}E_1.}$$

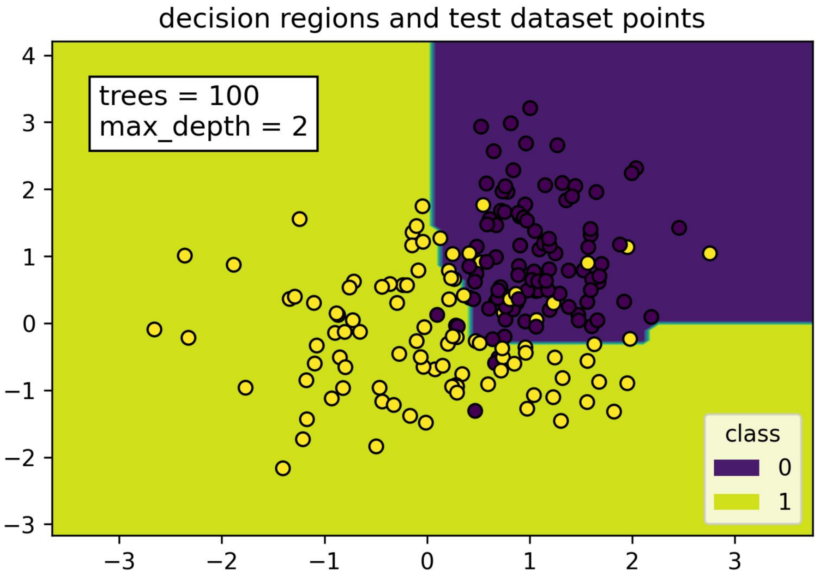
Error decreased by N times!

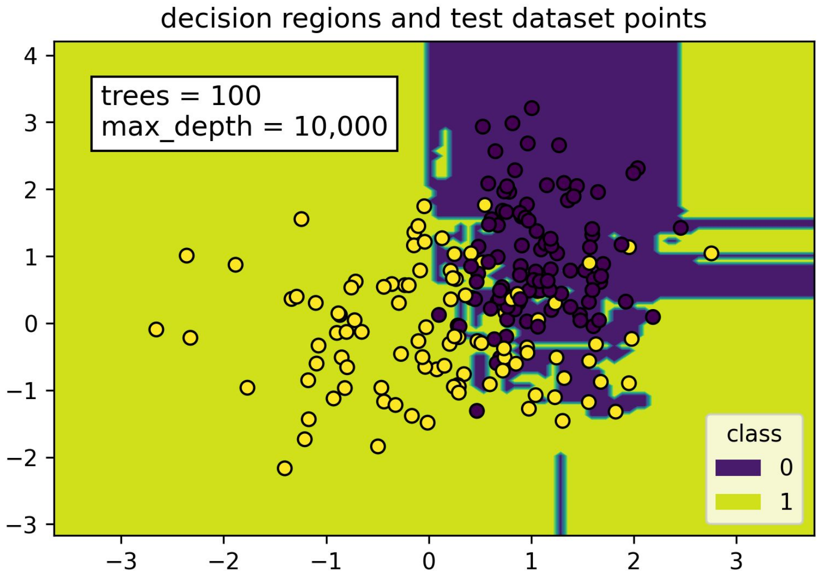
Bagging = Bootstrap aggregating

Decreases the variance if the basic algorithms are not correlated



Bagging overfitting





https://www.tommykhoo.com/machine-learning/rf_overfit/

Random Forest

07

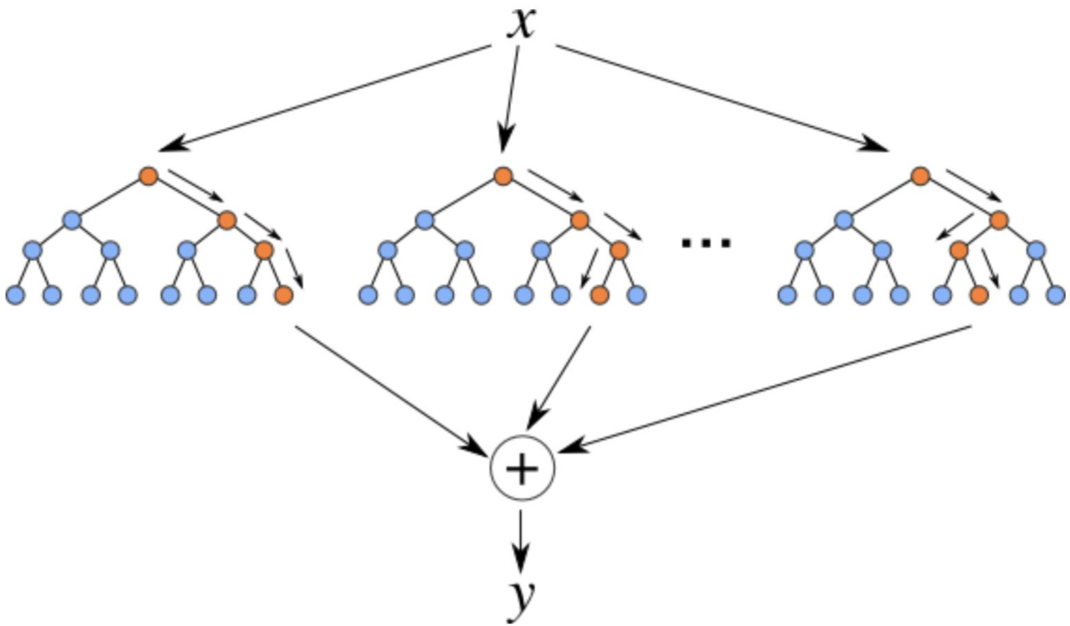
RSM - Random Subspace Method

Same approach, but with features.

Just subsample of features for each bootstraped dataset

Random Forest

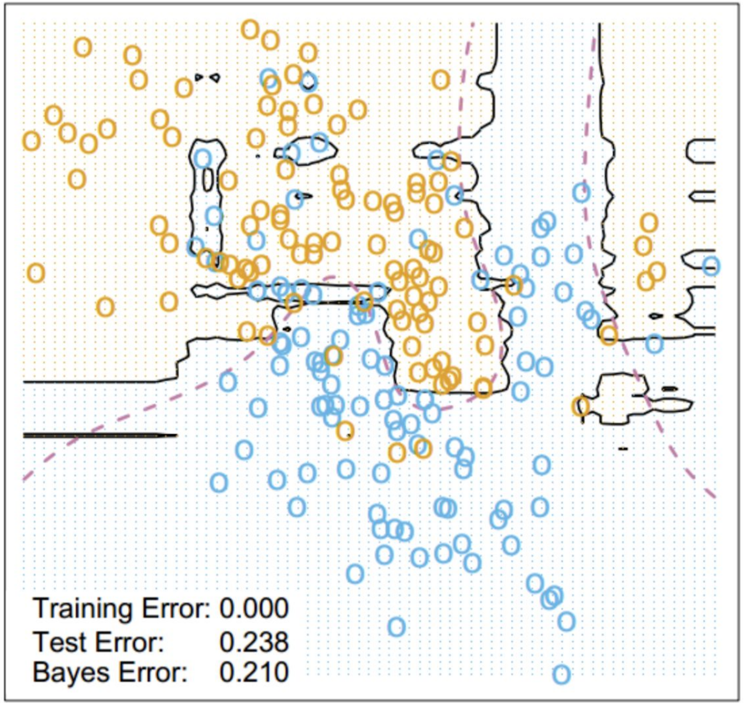
Bagging + RSM = Random Forest



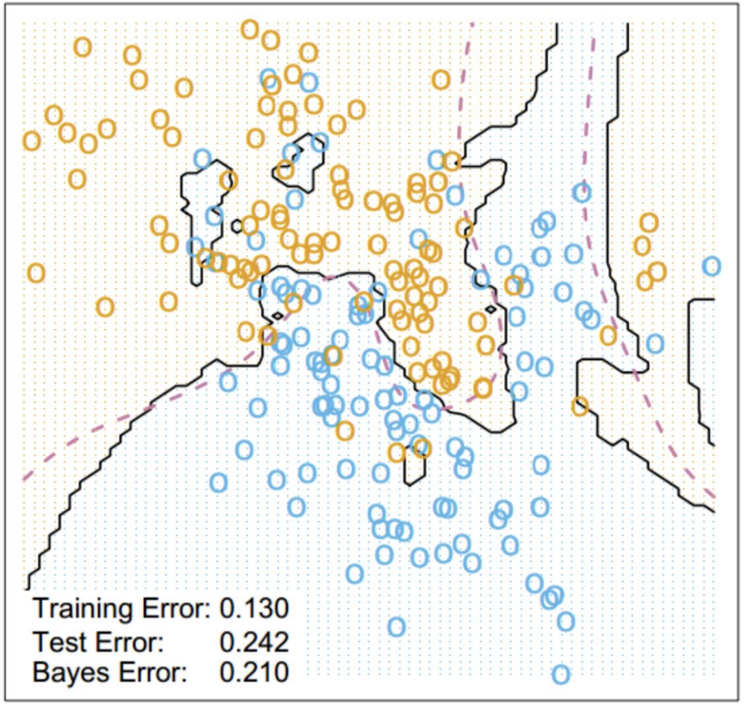
Random Forest

- One of the greatest “universal” models
- There are some modifications: Extremely Randomized Trees, Isolation Forest, etc

Random Forest Classifier



3−Nearest Neighbors

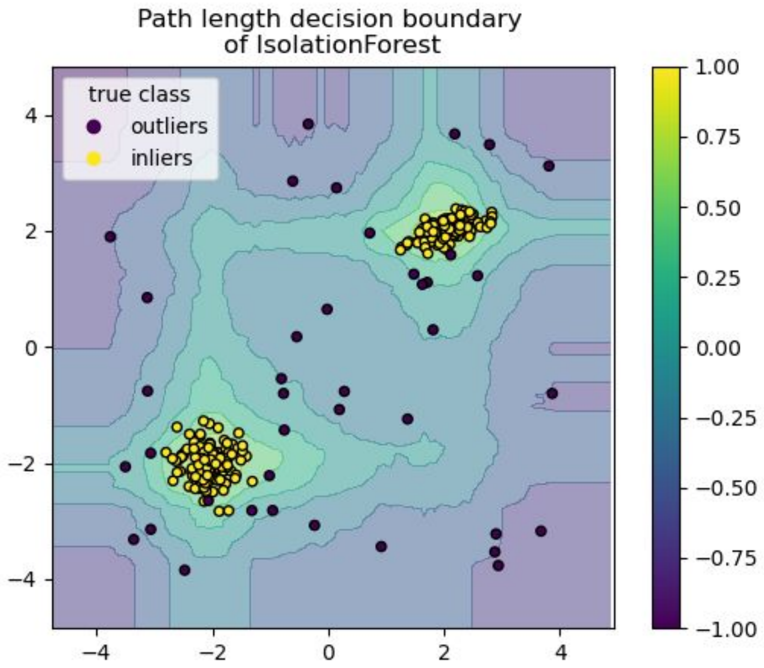


Isolation forest

girafe
ai

08

Method to search for anomalies



Method to search for anomalies

Isolation Forest ‘isolates’ observations by randomly selecting a feature and then randomly selecting a split value between the maximum and minimum values of the selected feature.

This path length, averaged over a forest of such random trees, is a measure of normality and our decision function.

Random partitioning produces noticeably shorter paths for anomalies. Hence, when a forest of random trees collectively produce shorter path lengths for particular samples, they are highly likely to be anomalies.

https://scikit-learn.org/stable/modules/outlier_detection.html#isolation-forest

https://alexanderdyakonov.wordpress.com/2017/04/19/%D0%BF%D0%BE%D0%B8%D1%81%D0%BA-%D0%B0%D0%BD%D0%BE%D0%BC%D0%B0%D0%BB%D0%B8%D0%B9-anomaly-detection/

Revise

1. Intuition
2. Construction procedure
3. Information criteria
4. Decision trees special highlights
    - Decision tree as linear model
    - Dealing with missing data
    - Categorical features
5. Bootstrap and Bagging
6. Random Forest

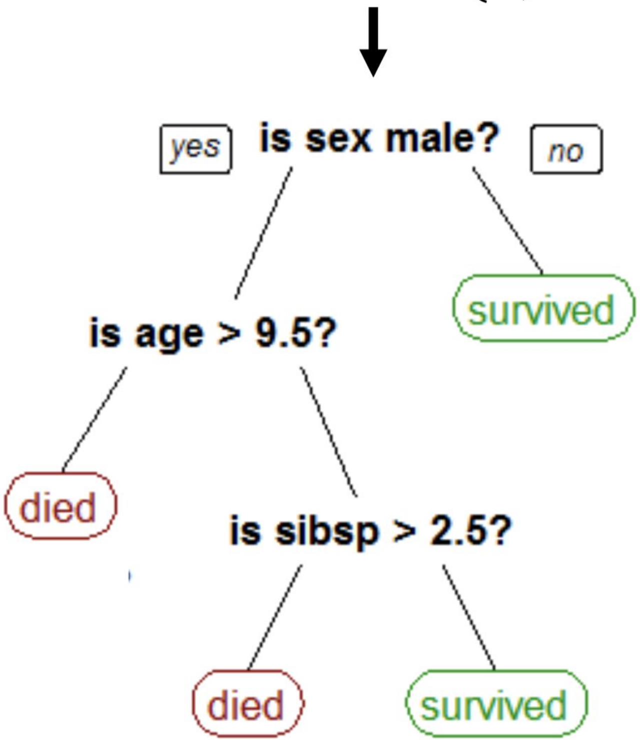
Thanks for attention!

Questions?

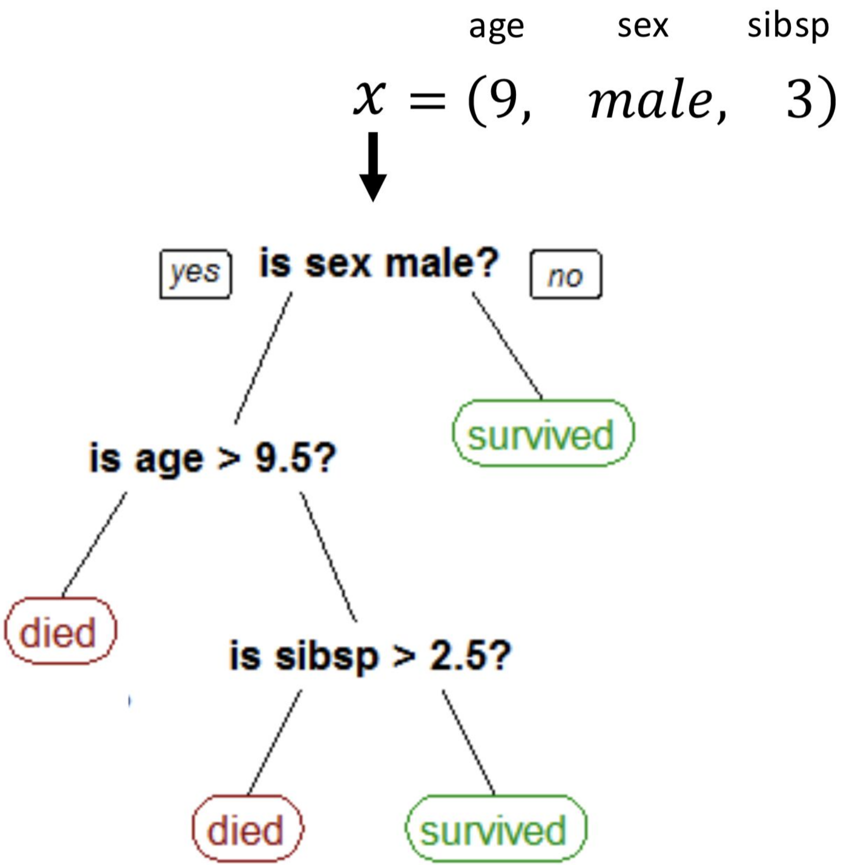
girafe
ai

Decision tree

$$x = (9, \quad male, \quad 3)$$

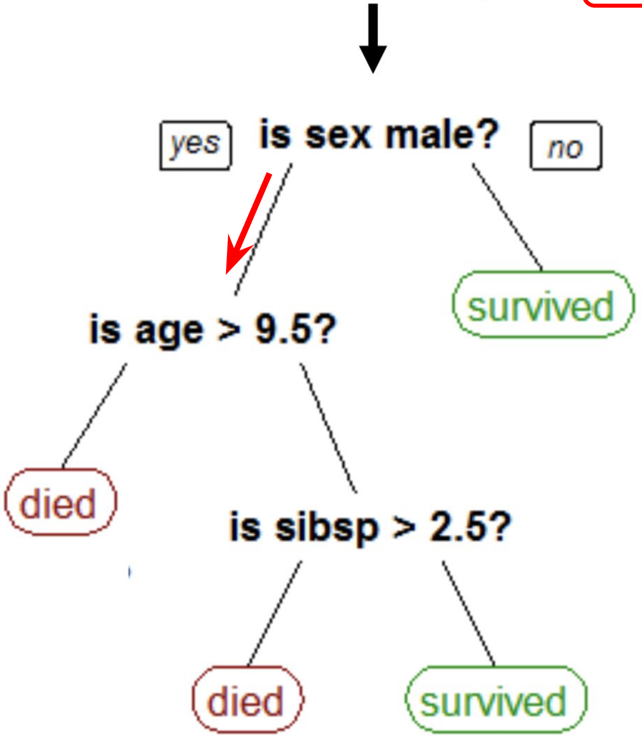


Decision tree

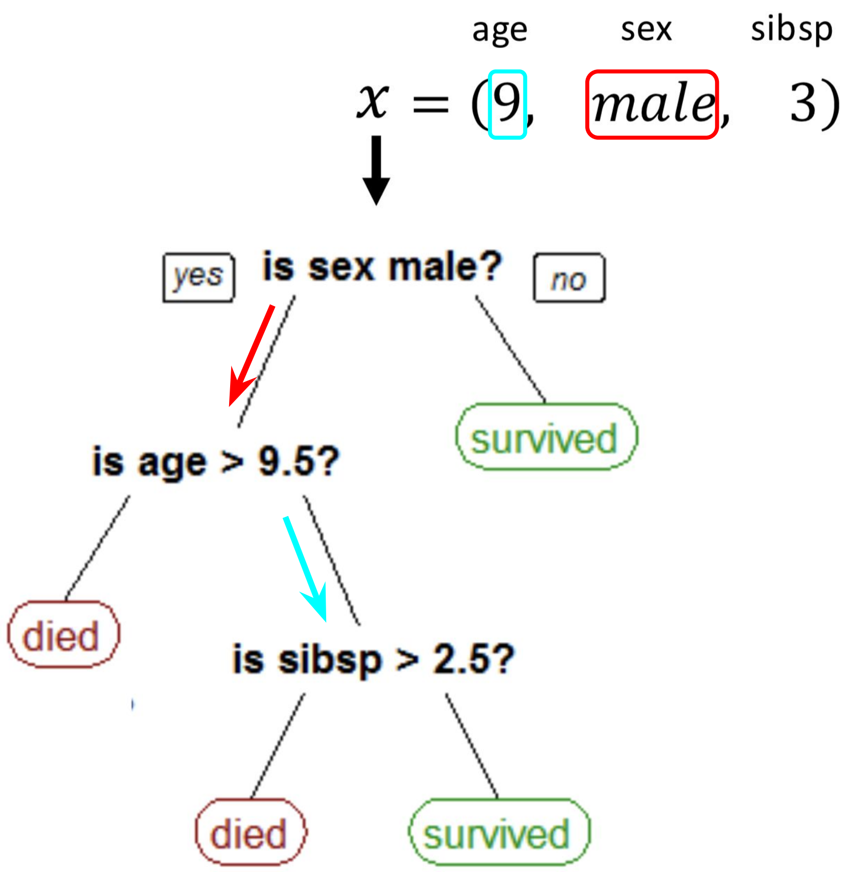


Decision tree

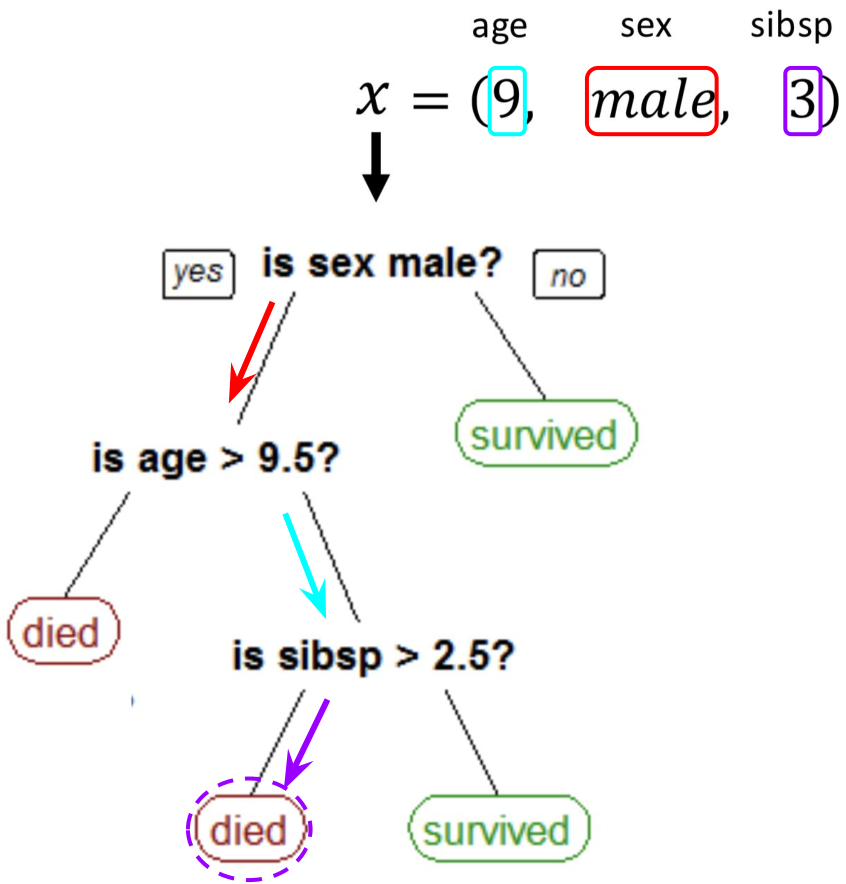
$$
\begin{array}{cccc}
 & \text{age} & \text{sex} & \text{sibsp} \\
x = (& 9, & \boxed{male}, & 3)
\end{array}
$$



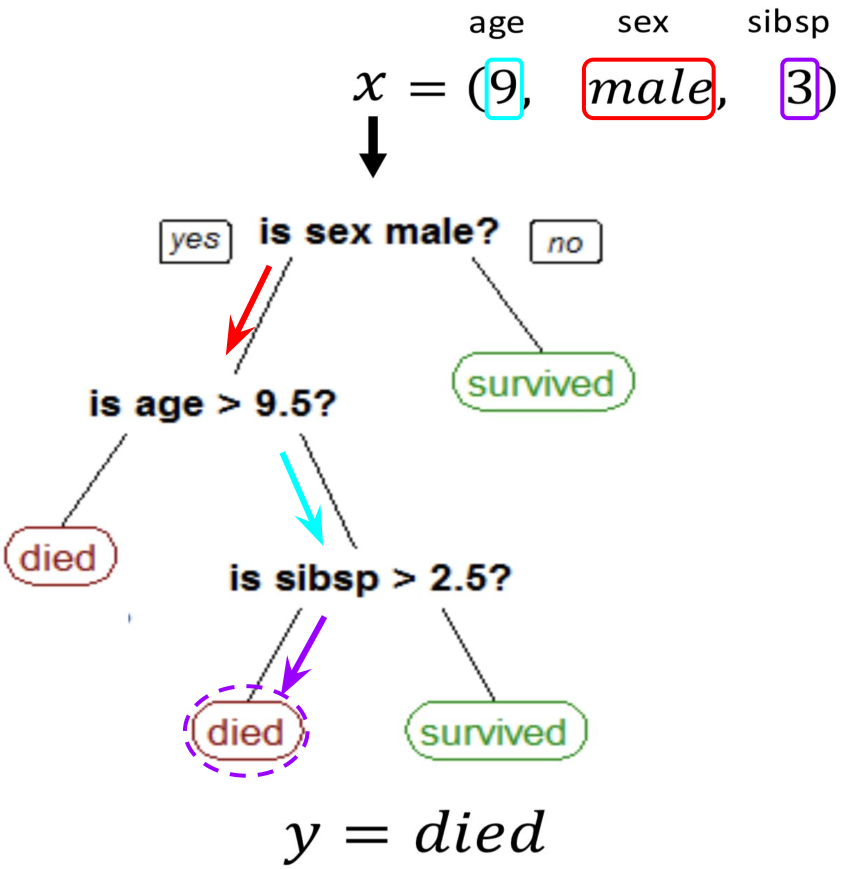
Decision tree



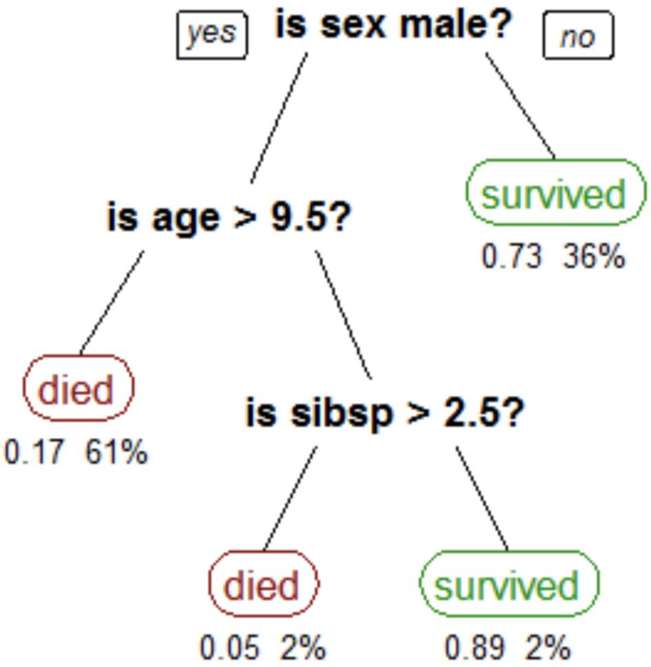
Decision tree



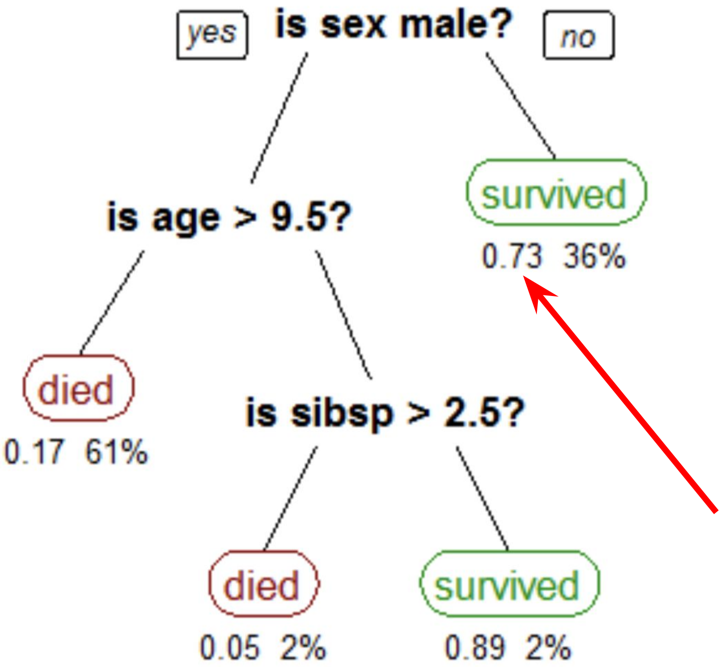
Decision tree



Decision tree in classification

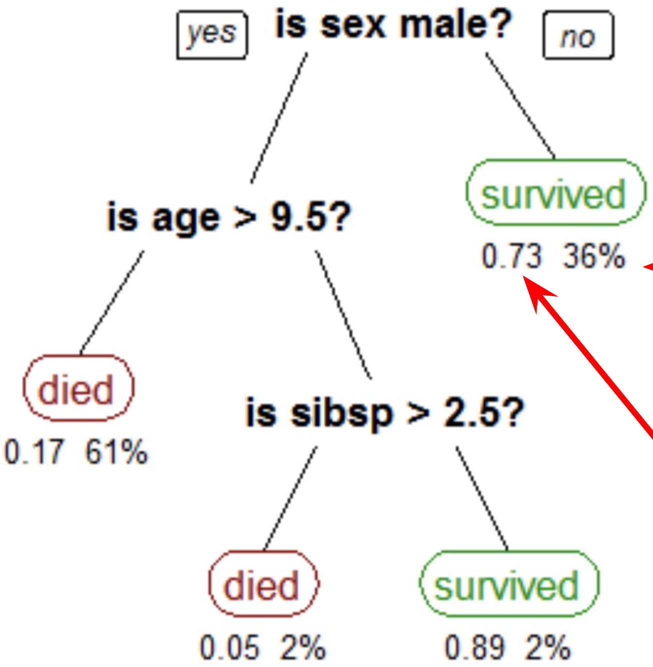


Decision tree in classification



Part of survived passengers
= part of objects of 1st class

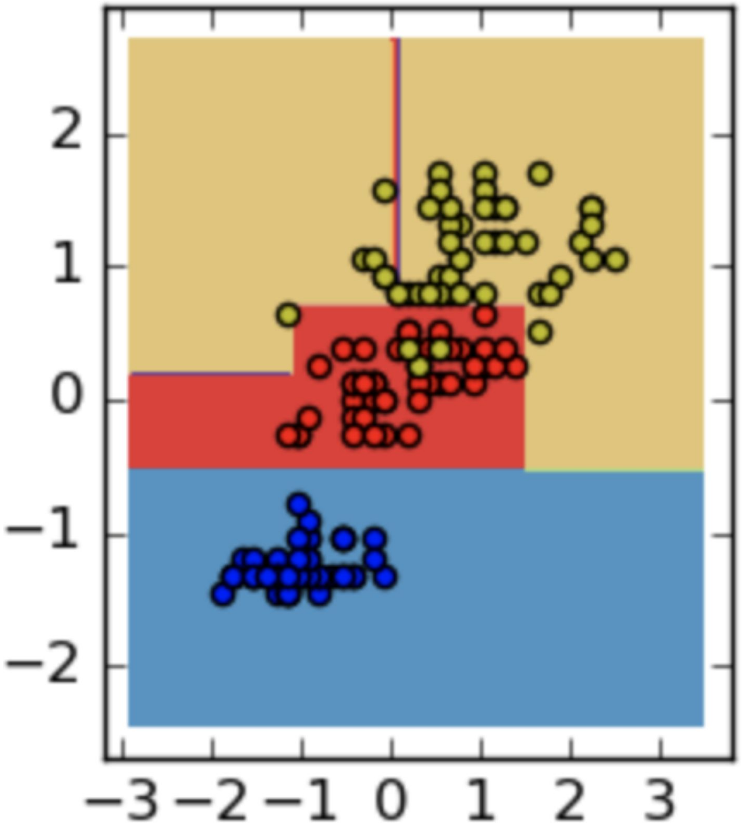
Decision tree in classification



Part of dataset in this leaf

Part of survived passengers = part of objects of 1st class

Decision tree in classification



Classification problem with 3 classes and 2 features.