classification, decision trees and scoring: a reminder

carlos soares

(csoares@fe.up.pt)

[including materials kindly provided by Alípio Jorge and adapted from David Sontag, Luke Zettlemoyer, Carlos Guestrin and Andrew Moore as well as from Eammon Keogh]

http://www.flickr.com/photos/upvcfrontdoor/4642103837/

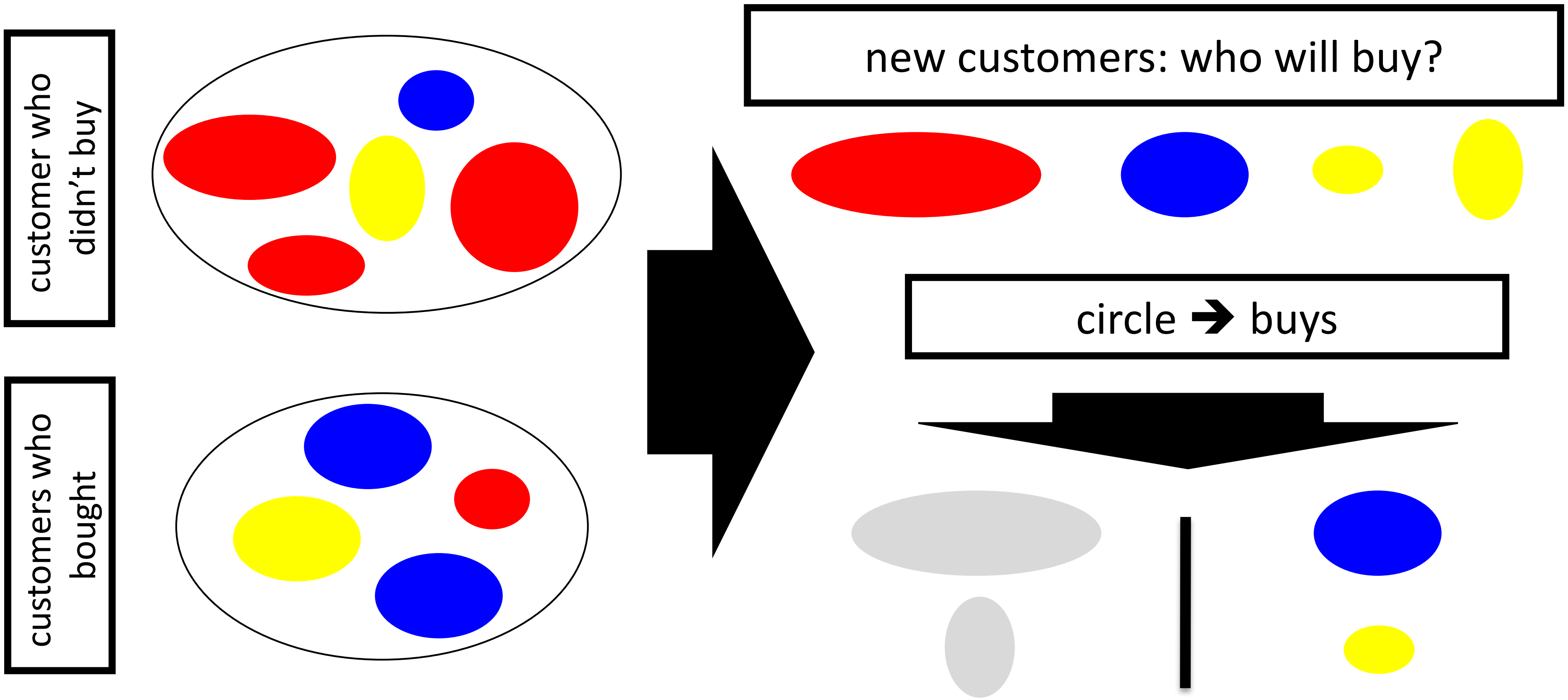
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

reference materials

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

- JMM et al. ch. 4+7

predictive: classification for targeting



data for classification

target (or dependent) variable

independent variable (or attribute)

- prospects
  - customers who didn't buy a car in the last 4 years

would like to predict

Comprou	Idade	Rendimento	Ag.fam	Vendas anteriores	Última Venda
	41	50000	2	1	0
	39	68000	2	0	30000
	58	61000	4	0	0
	26	25000	3	0	0
	21	50000	1	1	20000
	38	43000	2	0	0
	44	43000	4	1	47000
	27	47000	2	1	21000
	70	23000	2	0	25000

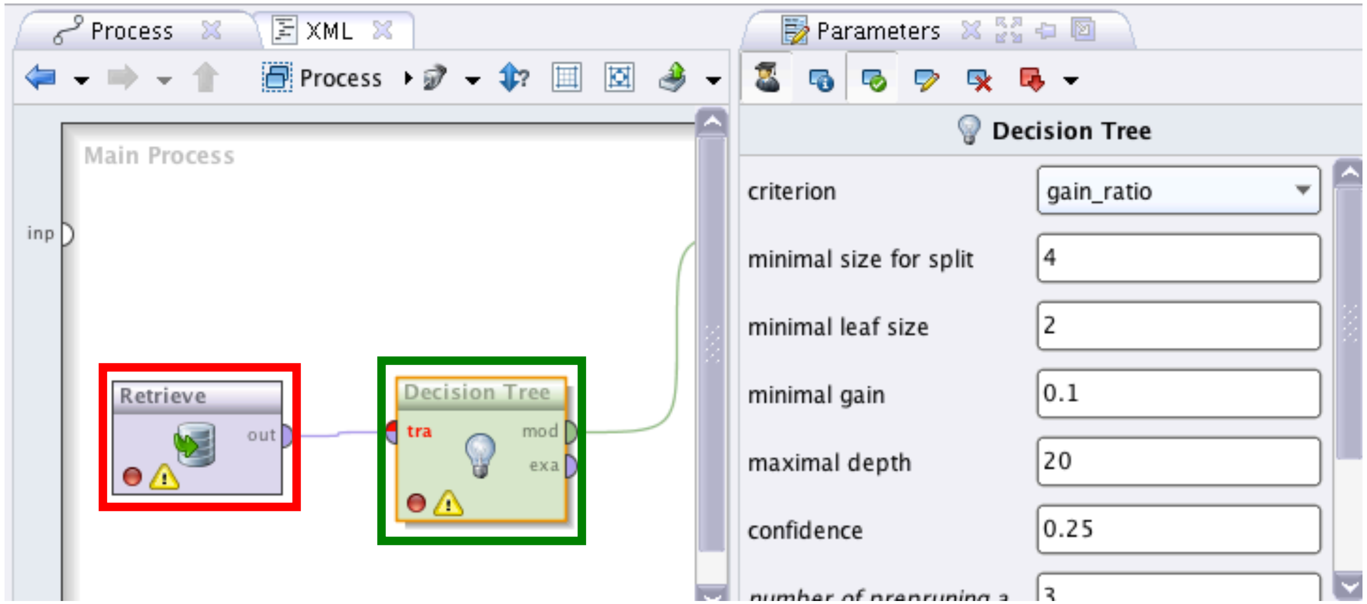
- results from previous campaigns
  - customers who were contacted and their response

already known

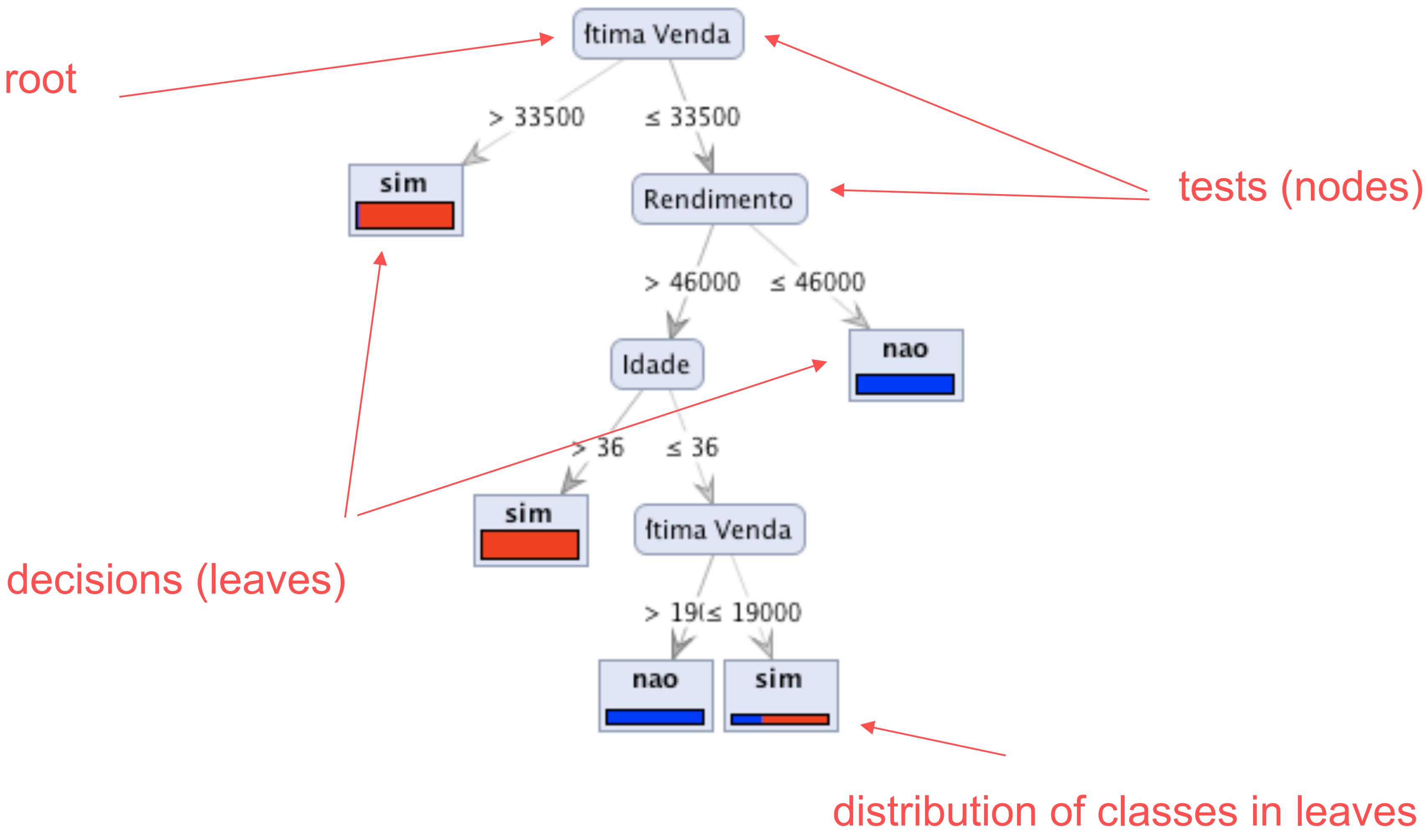
Comprou	Idade	Rendimento	Ag.fam	Vendas anteriores	Última Venda
não	37	49000	2	1	42000
sim	43	68000	3	0	0
sim	42	61000	4	0	0
sim	26	52000	2	0	0
sim	40	64000	1	1	21000
sim	38	52000	1	0	0
sim	45	43000	4	1	47000
sim	35	45000	2	1	34000
não	39	43000	2	0	0
sim	31	55000	3	1	46000
sim	34	57000	3	1	52000
não	38	44000	4	0	0
não	34	68000	2	1	33000
[illegible]	[illegible]	[illegible]	[illegible]	[illegible]	[illegible]

learning a model from data

- load data
- apply decision tree algorithm



model: classification tree (or decision tree)

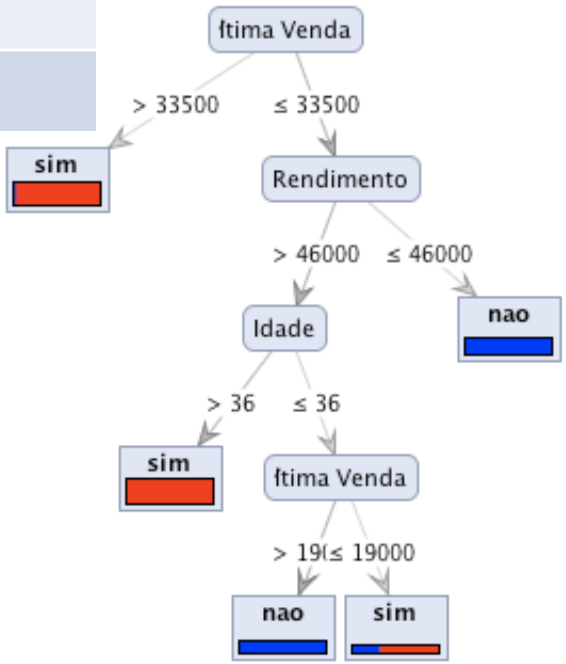


classify new examples

- prospects list

age	income	family size	previous sales	last sale
28	39.000	2	0	0
39	52.000	4	1	17.000
29	42.000	4	1	40.000

- class of the leaf each example is assigned by the tree?
  - i.e. predict…



classification: are we there yet?

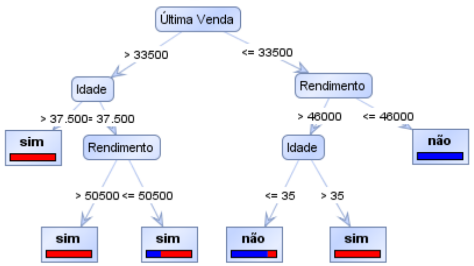
U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

responses to previous campaigns

known responses (class)

Comprou	Idade	Rendimento	Ag.fam	Vendas anteriores	Última Venda
não	37	49000	2	1	42000
sim	43	68000	3	0	0
sim	42	61000	4	0	0
sim	26	52000	2	0	0
sim	40	64000	1	1	21000
sim	38	52000	1	0	0
sim	45	43000	4	1	47000
sim	35	45000	2	1	34000
não	39	43000	2	0	0

1



prospects

	A	B	C	D	
1	Comprou	Idade	Rendimento	Ag.fam	Vendas
2		41	50000	2	
3		39	68000	2	
4		58	61000	4	
5		26	25000	3	
6		21	50000	1	
7		38	43000	2	
8		44	43000	4	
9		27	47000	2	
10		70	23000	2	

unknown responses (class)

2

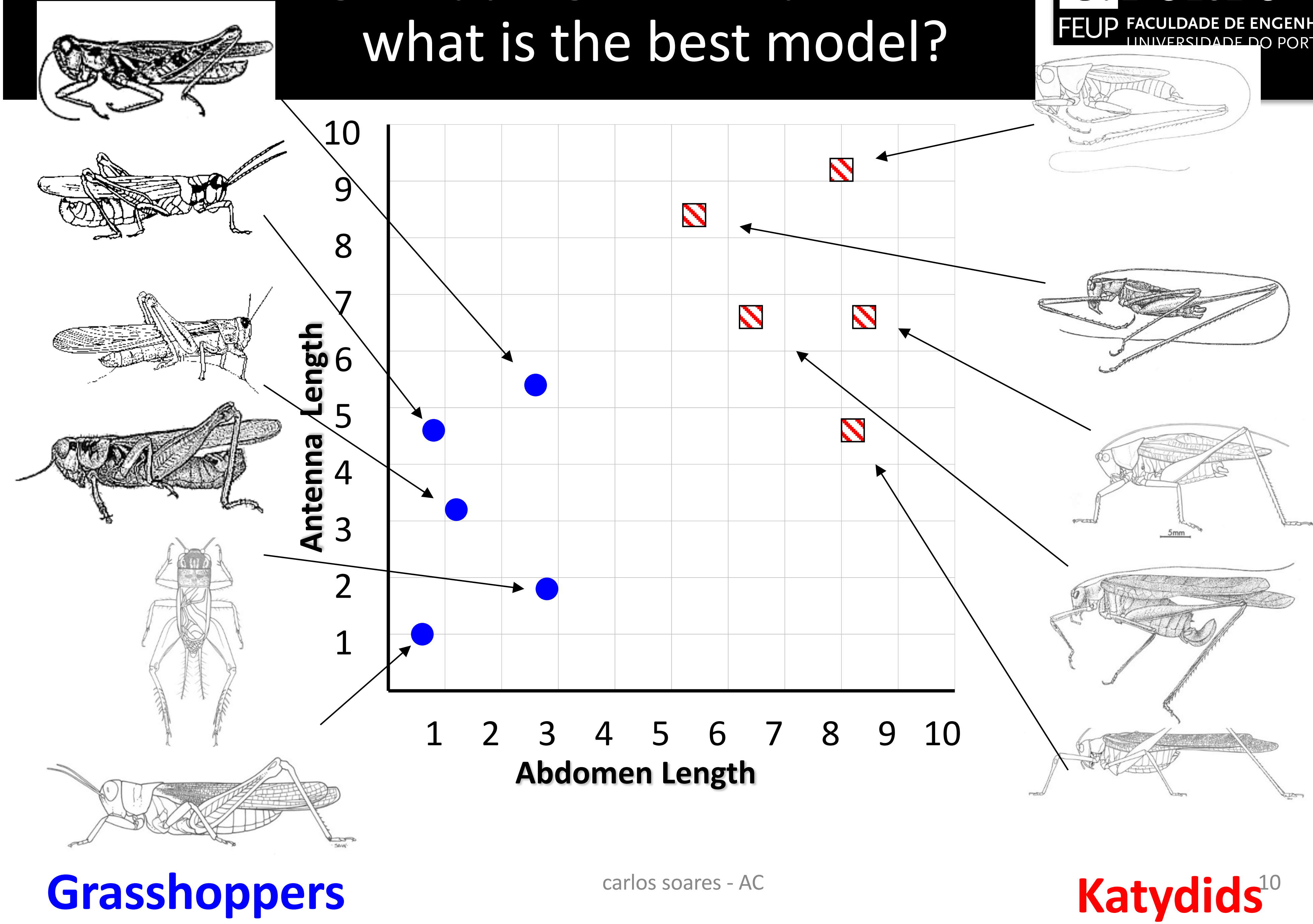
3

row no.	Comprou	prediction(...	confidence(...	confidence(...	Ida
1	?	sim	0	1	41
2	?	sim	0	1	39
3	?	sim	0	1	58
4	?	não	1	0	26
5	?	não	0.818	0.182	21
6	?	não	1	0	38
7	?	sim	0	1	44
8	?	não	0.818	0.182	27
9	?	não	1	0	70

predictions by the model

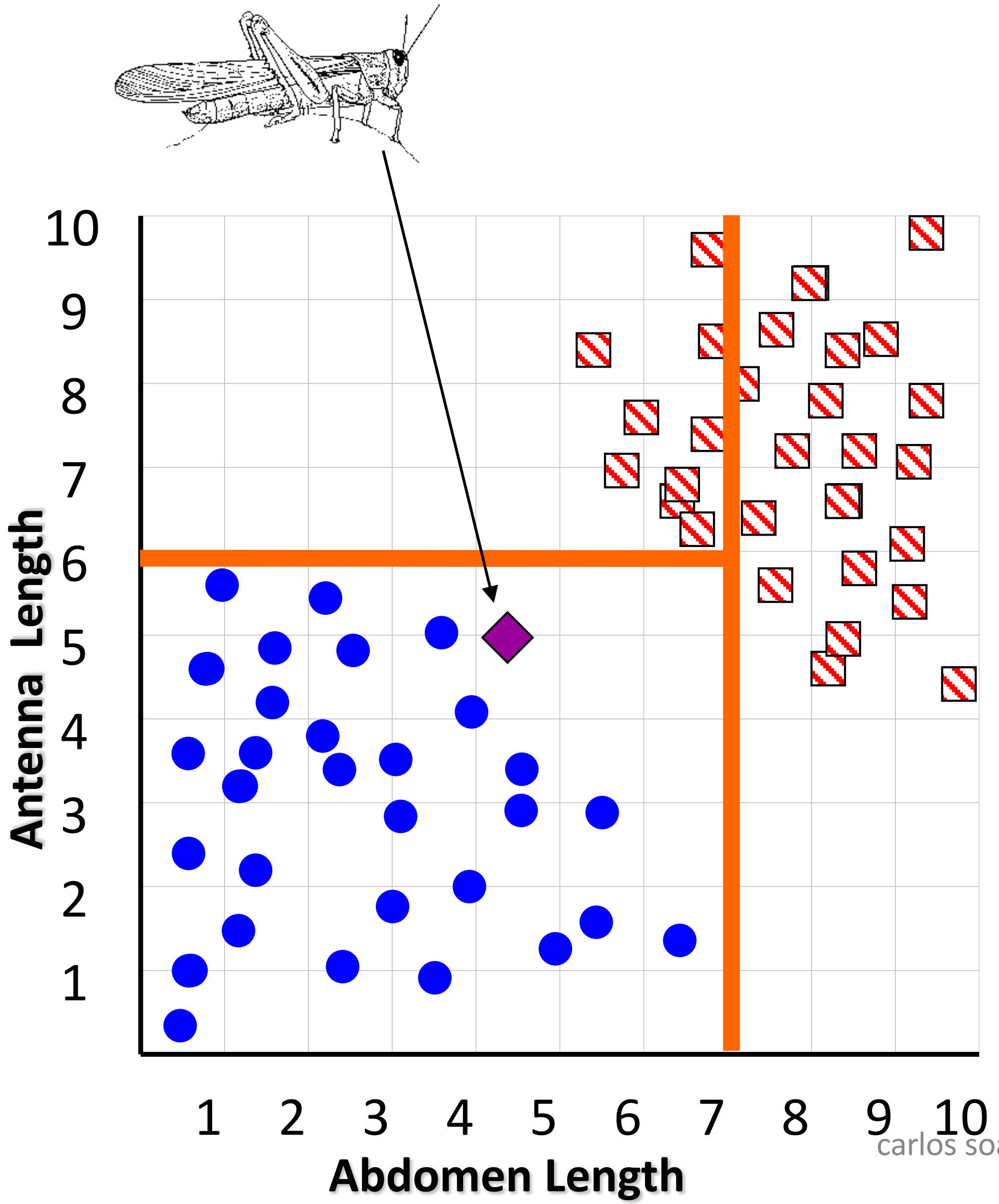
would you use the decisions proposed by this model?

running (hopping?) example: what is the best model?

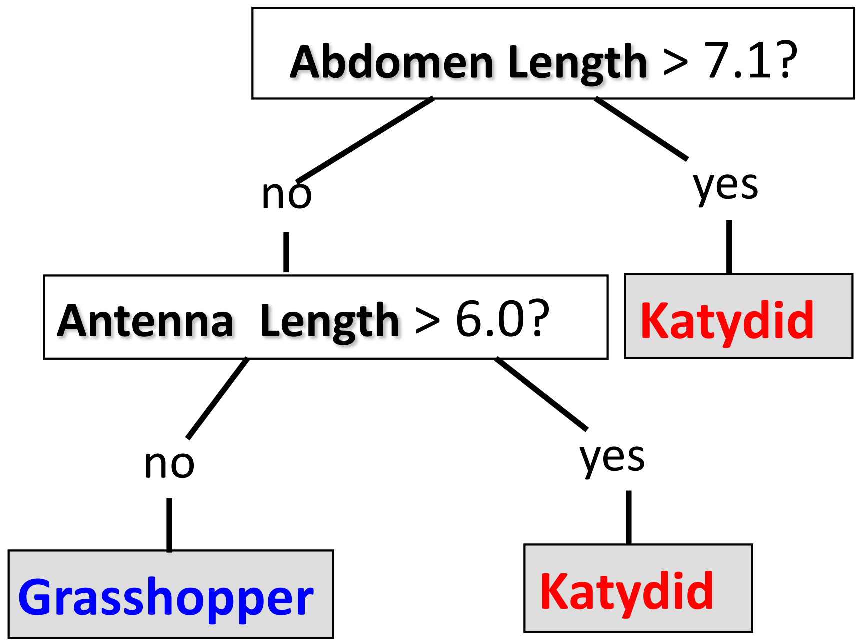


how decision tree algorithms see data…

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



Ross Quinlan

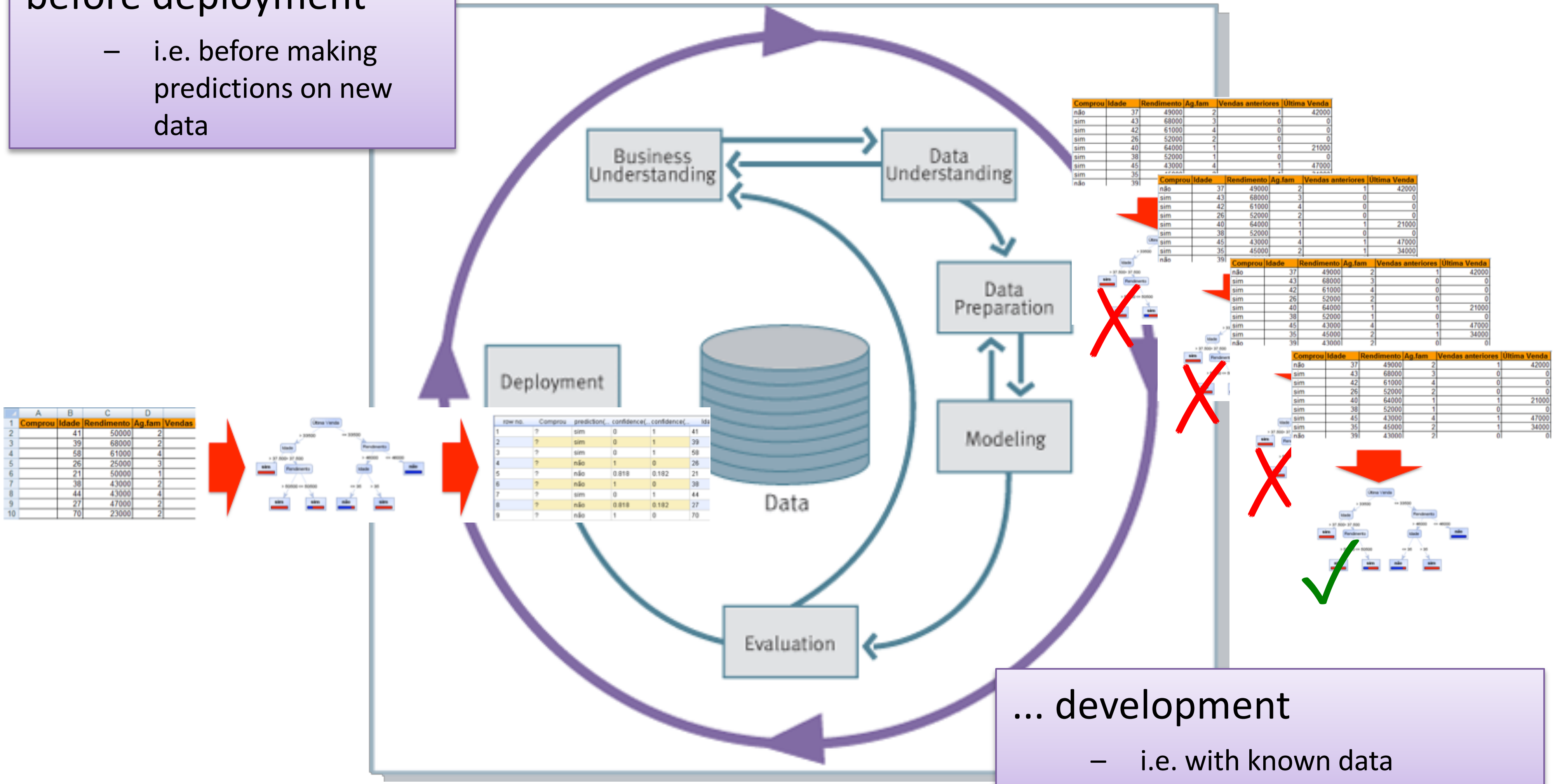


model development

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

before deployment

- i.e. before making predictions on new data



... development

- i.e. with known data

how good are the predictions?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- confusion matrix
  - prediction vs reality
  - number of right answers on the main diagonal
  - sum of the array is the total number of examples
- error rate
  - percentage/proportion of cases where the model misses
  - e.g. (2 + 1)/(5 + 1 + 2 + 29) = 8.1%

	truth: no	truth: yes
prediction: no	5	1
prediction: yes	**2**	**29**

not all errors are born equal...

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- e.g. targetting campaign
  - is contacting a non-buyer or missing a buyer equally important?
- the focus is on one class
  - ... usually
- the positive class
  - ... positive not necessarily good
- a new perpective on the confusion matrix

	truth: no	truth: yes
prediction: no	TN	FN
prediction: yes	FP	TP

evaluation measures

- multiple measures can be computed from the confusion matrix, including...

	truth: no	truth: yes
prediction: no	TN	FN
prediction: yes	**FP**	**TP**

$\frac{FP}{FP+TN}$	False positive rate (FPR) = 1-TNR	$\boxed{\frac{TP}{TP+FP}}$	Positive predictive value (PPV), also known as precision
$\frac{FN}{TP+FN}$	False negative rate (FNR) = 1-TPR	$\boxed{\frac{TN}{TN+FN}}$	Negative predictive value (NPV)
$\boxed{\frac{TP}{TP+FN}}$	True positive rate (TPR), also known as recall or sensitivity	$\frac{TP+TN}{TP+TN+FP+FN}$	Accuracy
$\boxed{\frac{TN}{TN+FP}}$	True negative rate (TNR), also known as specificity	$\frac{2}{1/\,precision+1/\,recall}$	F1-measure

is the model any good at all?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

	truth: no	truth: yes
prediction: no	5	1
prediction: yes	2	29

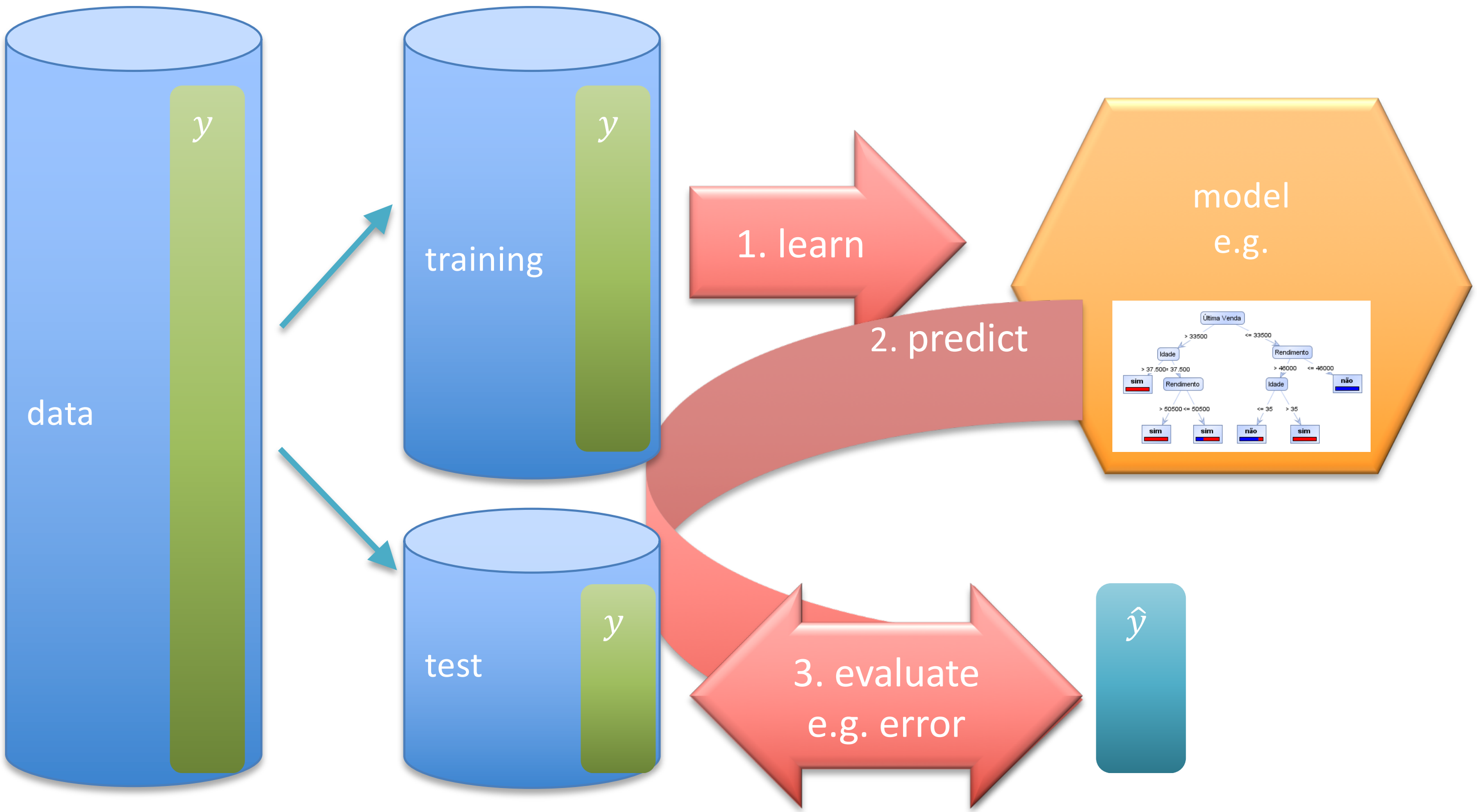
- model error: 3/37 = 8.1%
- **baseline**
  - simplest model that can be obtained from the data

	truth: no	truth: yes
class distribution	7	30

most "popular" choice

  - ... with error: 7/37 = 18,9%
- so, should we use the model?

evaluation methodology: do not forget!



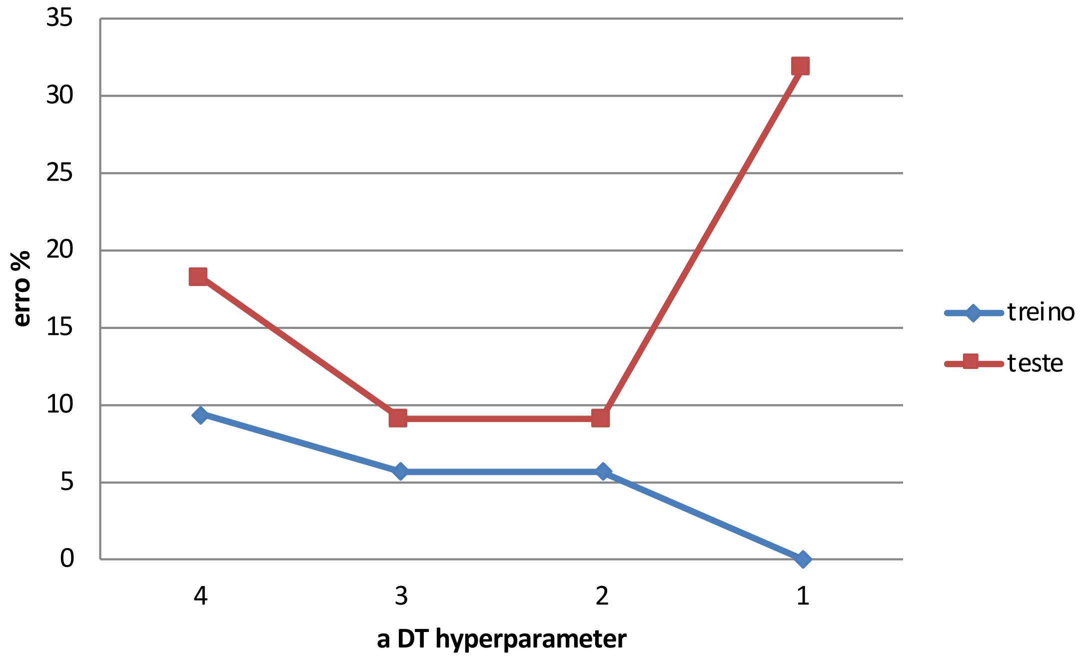
DECISION TREES

gps

- overfitting
- how the algorithm for induction of decision trees works

overfitting

- algorithms can be controlled to adjust to the data more or less



- good reason to
  - design your experimental setup carefully
  - understand how algorithms work

learning a decision tree (1 and 2/5)

x1	x2	class
a	p	yes
a	q	no
a	p	yes
b	q	no
c	p	no
a	p	yes
c	p	yes
b	p	no

yes= 4
no= 4

1. we have a set of labelled examples
   – the target variable indicates the class of each case (e.g. yes, no)
   – on the root knot we have all the cases
2. if all the examples are of the same class, we stop

learning a decision tree (3/5)

x1	x2	class
a	p	yes
a	q	no
a	p	yes
b	q	no
c	p	no
a	p	yes
c	p	yes
b	p	no

yes= 4
no= 4

test	true (yes/no)	false (yes/no)
x1=a	3/1	1/3

3. otherwise, we will divide (split) the examples in the root
   - so that the classes are well separated
   - each test is of variable type = value, or variable > value

learning a decision tree (3/5 – cont’d)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

x1	x2	class
a	p	yes
a	q	no
a	p	yes
b	q	no
c	p	no
a	p	yes
c	p	yes
b	p	no

yes= 4
no= 4

test	true (yes/no)	false (yes/no)
x1=a	3/1	1/3
x1=b	0/2	4/2
x1=c	1/1	3/3
x2=p	4/2	0/2
x2=q		

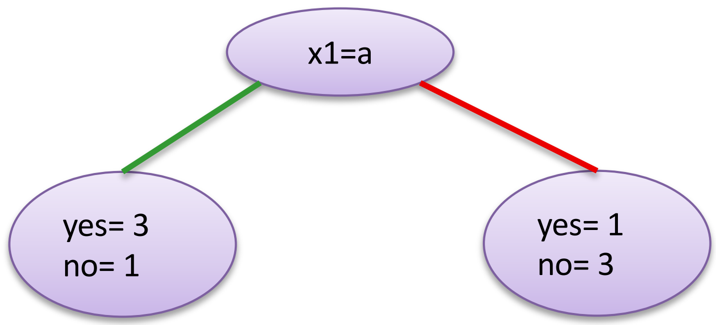
3. otherwise, we will divide (split) the examples in the root
   - so that the classes are well separated
   - each test is of variable type = value, or variable > value

unnecessary

learning a decision tree (4/5)

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

x1	x2	class
a	p	yes
a	q	no
a	p	yes
b	q	no
c	p	no
a	p	yes
c	p	yes
b	p	no

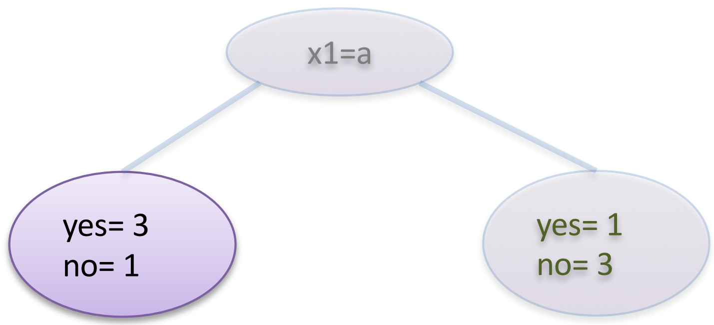


4. create to descendant nodes according to the selected test

test	true (yes/no)	false (yes/no)
x1=a	3/1	1/3

learning a decision tree (5/5)

x1	x2	class
a	p	yes
a	q	no
a	p	yes
b	q	no
c	p	no
a	p	yes
c	p	yes
b	p	no



test	true (yes/no)	false (yes/no)
x1=a	3/1	1/3
x2=p	3/0	0/1

unnecessary

5. We repeat the process for the set of examples in each of these descendant nodes

splits: the good, the bad and the ugly

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

x1	x2	class
a	p	yes
a	q	no
a	p	yes
b	q	no
c	p	no
a	p	yes
c	p	yes
b	p	no

too good to be true?

good or bad?...

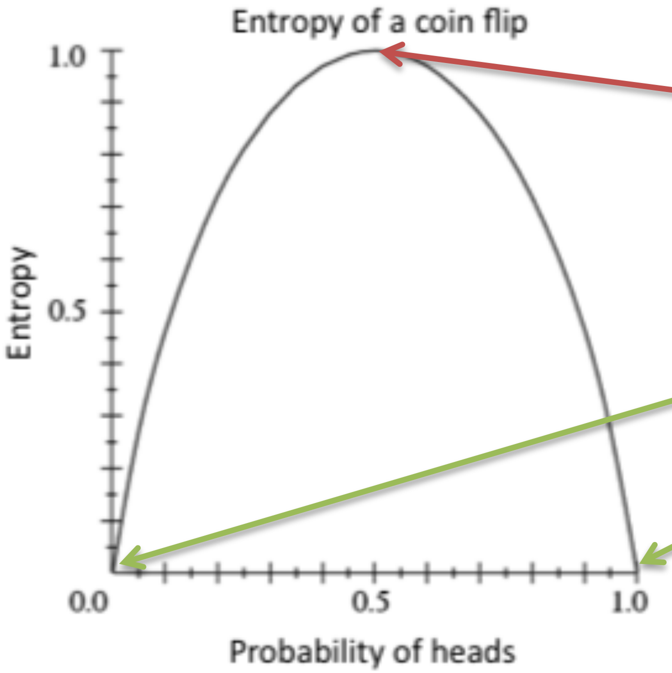
test	true (yes/no)	false (yes/no)
x1=a	3/1	1/3
x1=b	0/2	4/2
x1=c	1/1	3/3
x2=p	4/2	0/2
x2=q		

is this always bad?

entropy as diversity

- Entropy *H(Y)* of a random variable *Y*

test	true (yes/no)	false (yes/no)
x1=a	3/1	1/3
x1=b	0/2	4/2
x1=c	1/1	3/3
x2=p	4/2	0/2
x2=q		



$$H(Y) = -\sum_{i=1}^{k} P(Y = y_i) \log_2 P(Y = y_i)$$

... the bigger the decrease in diversity, the better

U.PORTO FEUP FACULDADE DE ENGENHARIA UNIVERSIDADE DO PORTO

x1	x2	class
a	p	yes
a	q	no
a	p	yes
b	q	no
c	p	no
a	p	yes
c	p	yes
b	p	no

test	true (yes/no)	false (yes/no)
x1=a	3/1	1/3

i.e. d(current set) >> d(left branch) + d(right branch)

e.g. information gain

$$IG(X) = H(Y) - H(Y \mid X)$$

numerical attributes

… x3	class
1	yes
3	no
2	yes
3	no
9	no
5	yes
5	yes
3	no

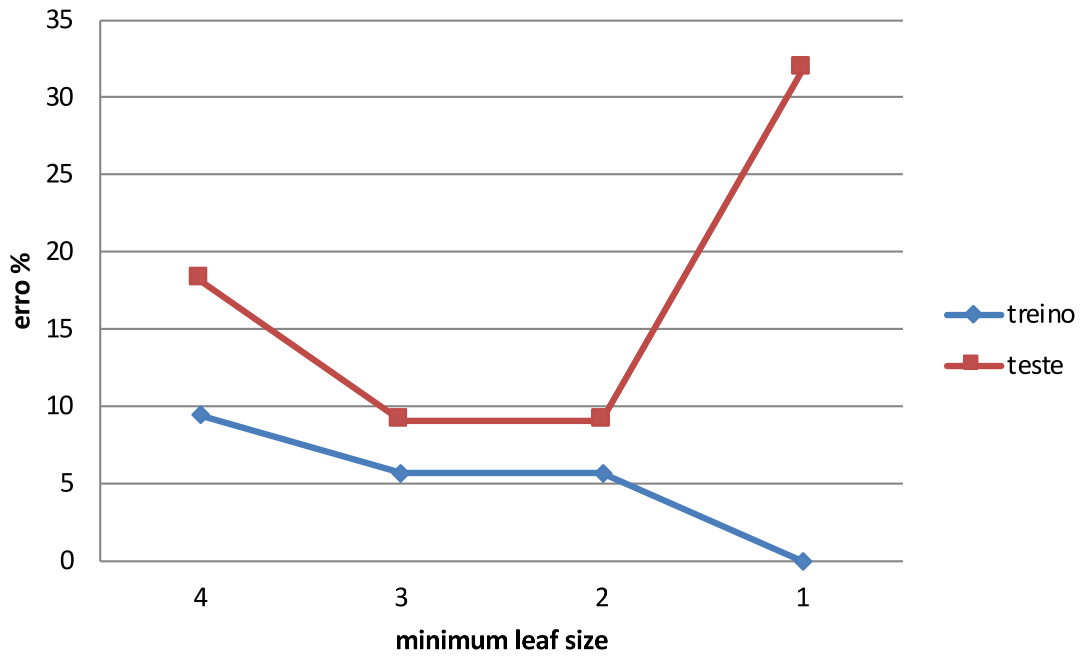
… x3	class
1	yes
2	yes
3	no
3	no
3	no
5	yes
5	yes
9	no

test	true (yes/no)	false (yes/no)
x3<1.5	1/0	3/4
x3<2.5	2/0	2/4
x3<4	2/3	2/1
x3<7	4/3	0/1

- only splits between examples of different classes should be considered
  - x3 < 1.5 cannot be better than x3 < 2.5

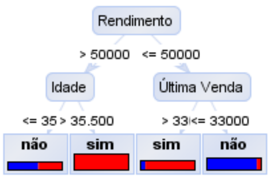
overfitting in DTs

- smaller leaves
  - eg. minimum leaf size hyperparameter

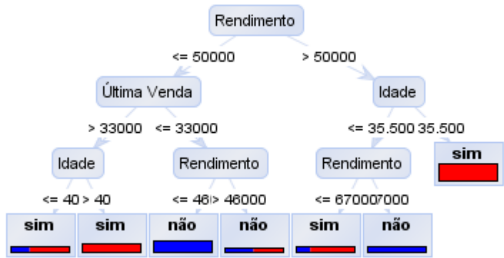


overfitting

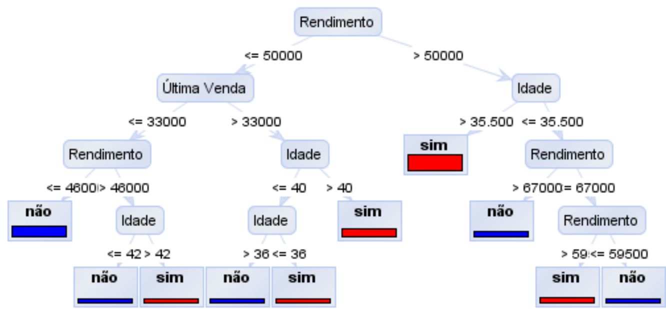
- trees obtained with different values of "minimum leaf size"
  - 4, 2 and 1



error (train)=18,18



error (train)=9,09%



error (train)=0,00%

CLASSIFICATION FOR SCORING

plan

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- binary classification
- evaluation in binary classification

classification: b&w or gray?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- direct application of the model splits examples into classes
  - eg. good and bad customers/buys or doesn't buy
- not suitable for all problems
  - list of 1000 prospects but send only to the 200 with the highest probability of buying
  - ... what if model selects only 30
  - ... or 300?
- *scoring*: use estimated probability of buying to order cases
  - select 200 with the highest probability ... or score
- score also provides information about (um)certainty of prediction

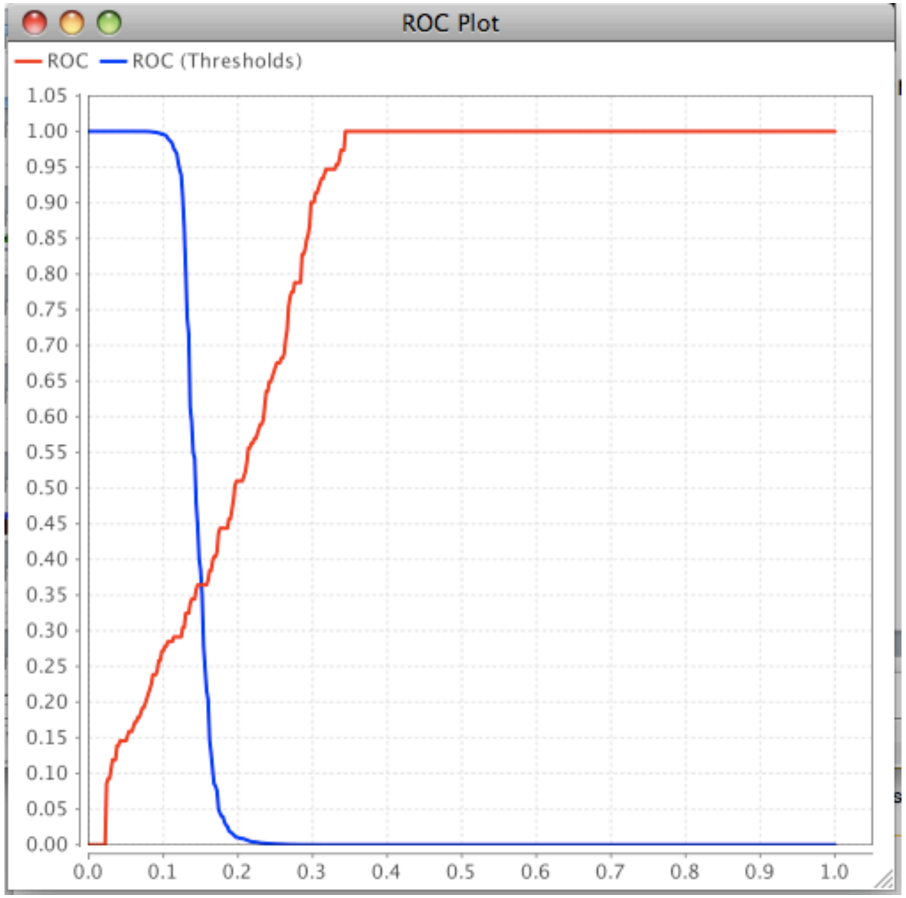
fonte: http://www.flickr.com/photos/backpackphotography/3354435787/

where to cut?

- imposed by available resources
  - *n* cases
- is *n* a suitable number?
  – maybe too many “bad” customers...
- arbitrary threshold
  - by default, class = yes if Prob(yes) > 0.5
- ... but which is the right value?
  - eg. important to find all “yes”
  - ... send if Prob(sim) > 0.3

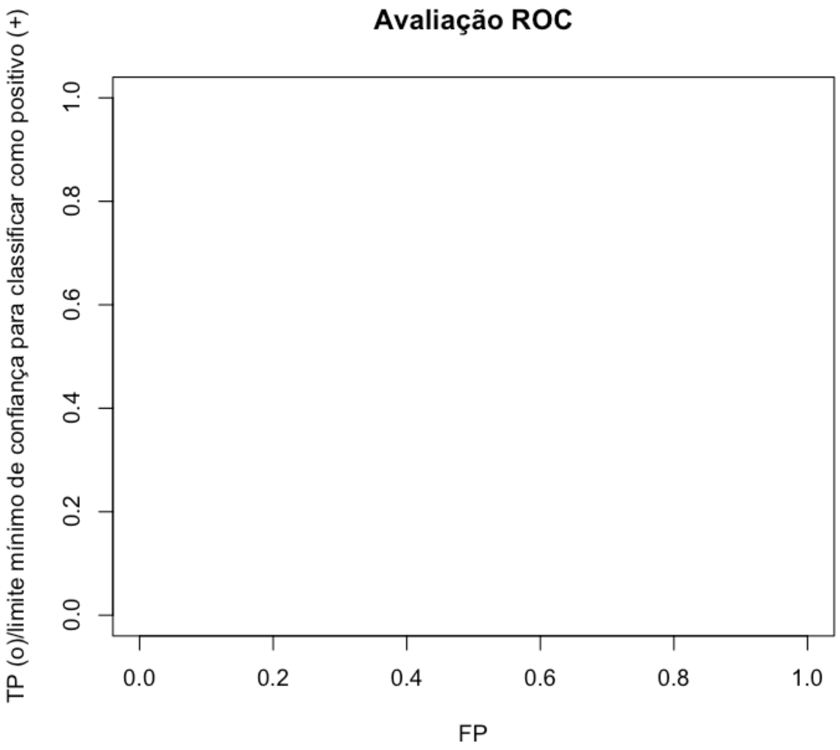
evaluate scoring models

- sort prediction by increasing order of belonging to positive class
  - P(sim | features)
- ROC analysis
  - *Receiver Operating Characteristic*
  - visualize proportion TP vs. FP
    - ... threshold
      - only rapidminer
  - ... to find best compromise



build ROC curve (1/5)

P(sim)	classe
1,00	sim
1,00	sim
0,90	sim
0,90	não
0,90	sim
0,85	não
0,83	sim
0,76	sim
0,75	não
0,73	sim
… (20 exemplos)	**… 10 × sim/10 × não**

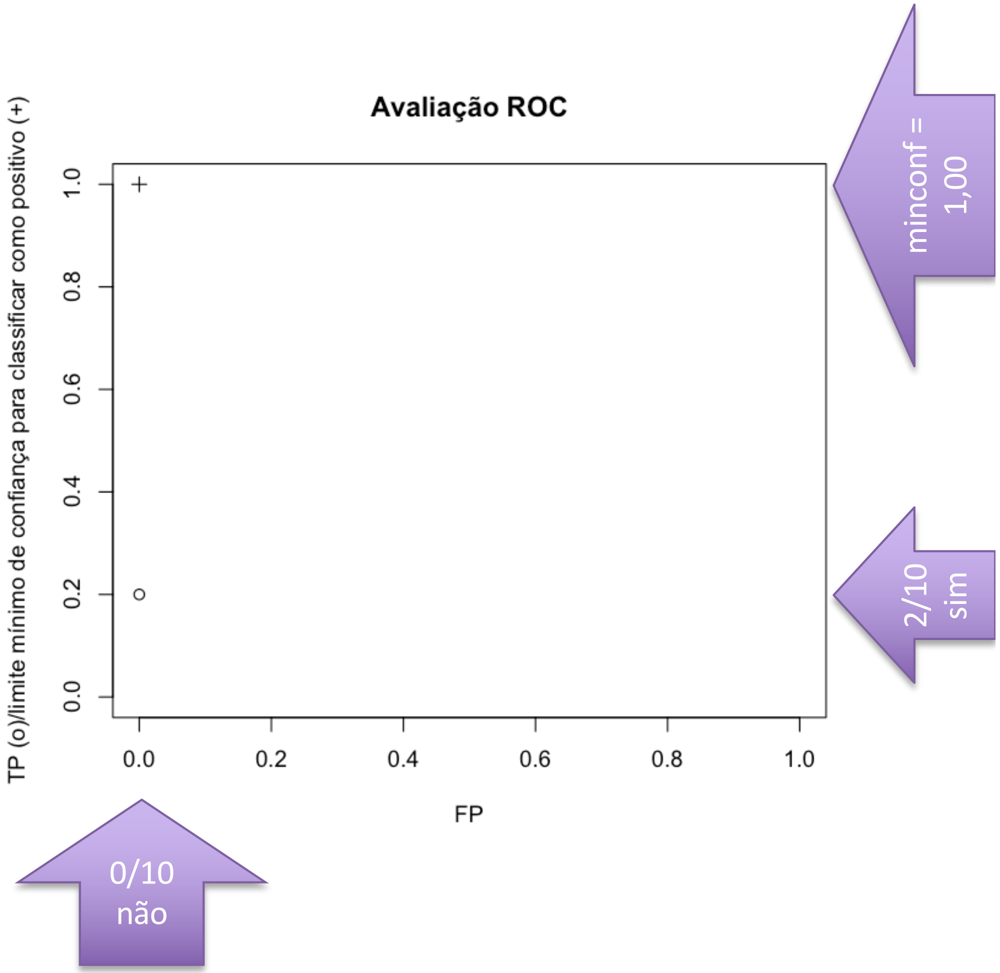


build ROC curve (2/5)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

min conf

P(sim)	classe
1,00	sim
1,00	sim
0,90	sim
0,90	não
0,90	sim
0,85	não
0,83	sim
0,76	sim
0,75	não
0,73	sim
… (20 exemplos)	**… 10 × sim/10 × não**

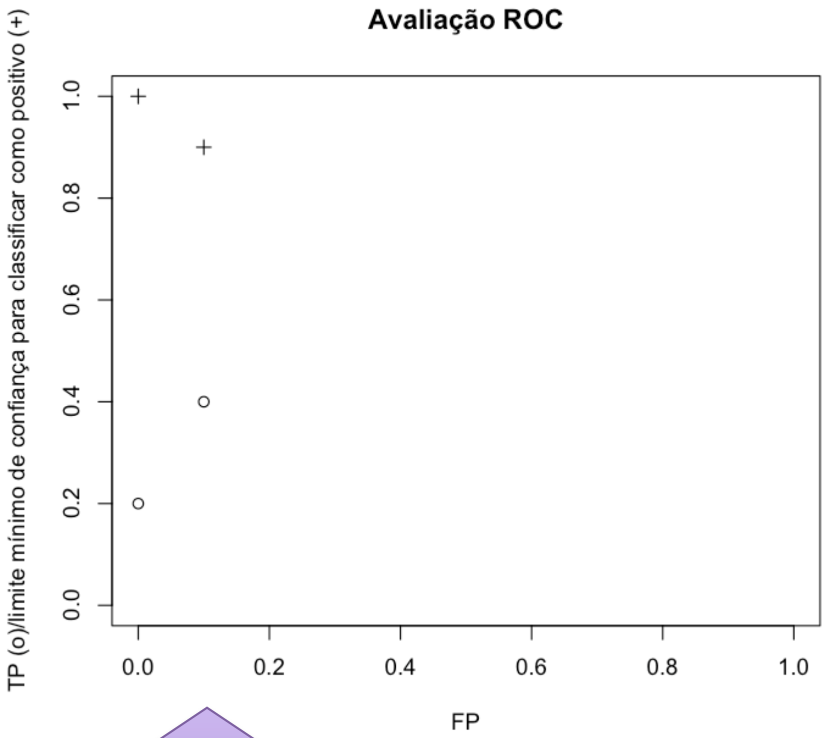


build ROC curve (3/5)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

min conf

P(sim)	classe
1,00	sim
1,00	sim
0,90	sim
0,90	não
0,90	sim
0,85	não
0,83	sim
0,76	sim
0,75	não
0,73	sim
… (20 exemplos)	**… 10 × sim/10 × não**



minconf = 0,90

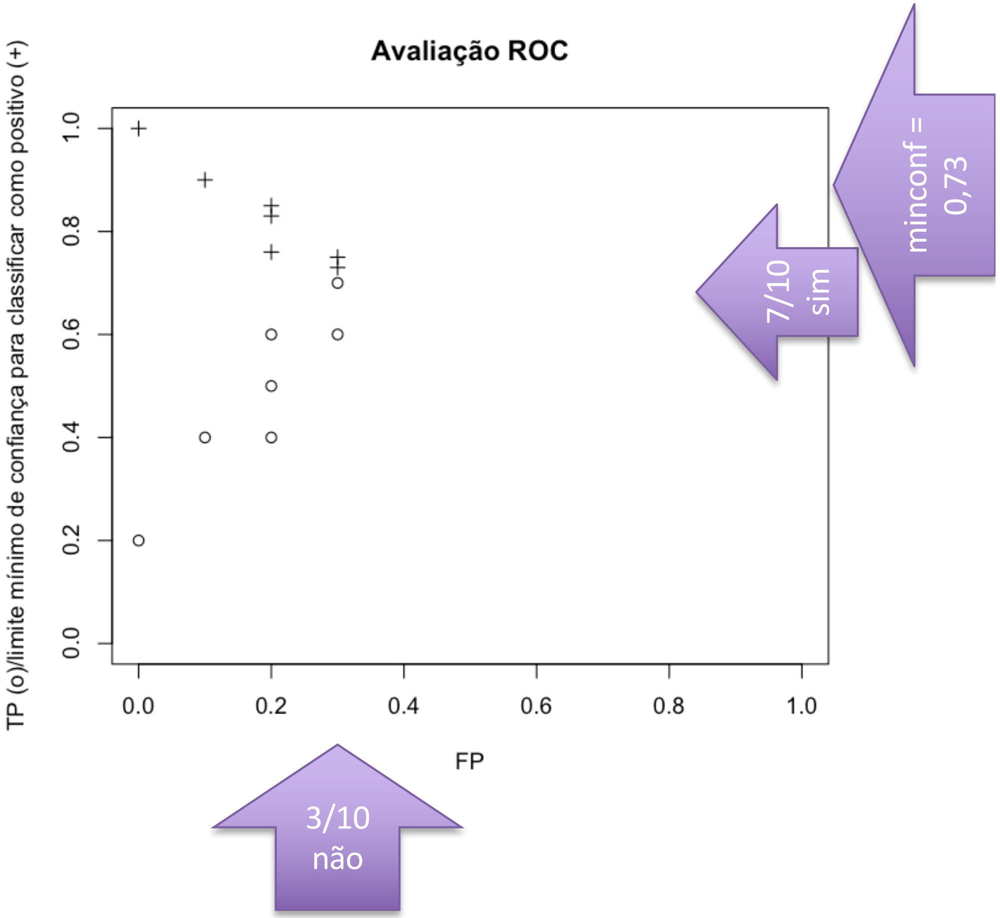
4/10 sim

1/10 não

build ROC curve (4/5)

P(sim)	classe
1,00	sim
1,00	sim
0,90	sim
0,90	não
0,90	sim
0,85	não
0,83	sim
0,76	sim
0,75	não
0,73	sim
... (20 exemplos)	**... 10 × sim/10 × não**

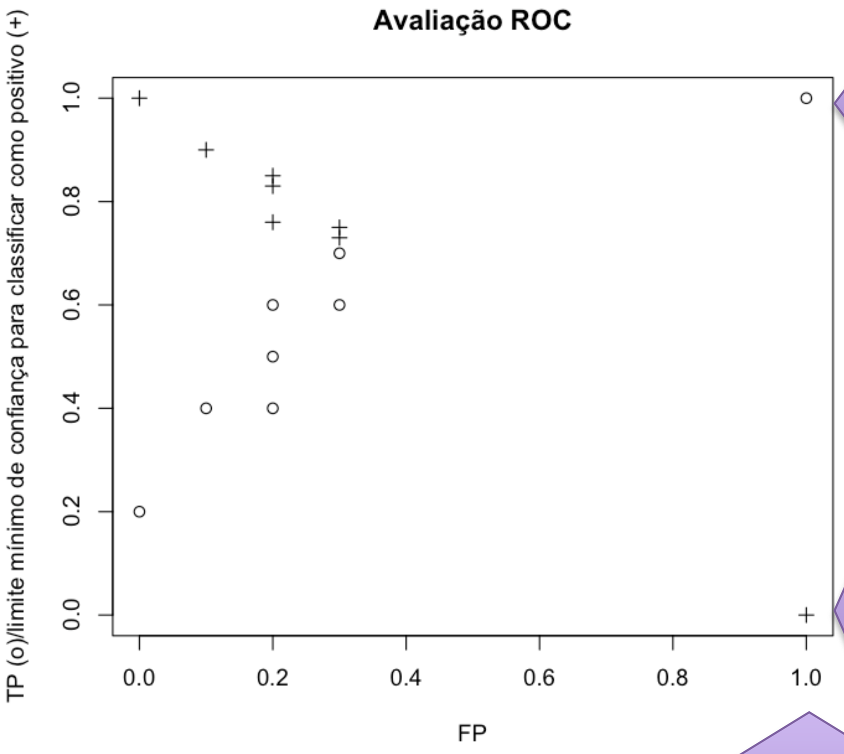
min conf



build ROC curve (5/5)

P(sim)	classe
1,00	sim
1,00	sim
0,90	sim
0,90	não
0,90	sim
0,85	não
0,83	sim
0,76	sim
0,75	não
0,73	sim
… (20 exemplos)	**… 10 × sim/10 × não**

min conf



10/10 sim

minconf = 0

10/10 não

are we there yet?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- identify problems where classification is useful
- identify relevant data
- know how to analyze and use a decision tree
- know the most common evaluation measures of classification models
- understand the need to use different data for modelling and evaluation
- understand how to evaluate the results of a classification model
- superficially understand the algorithm for induction of decision trees
- understand how to use and evaluate a classification model for scoring