classification: introduction

carlos soares

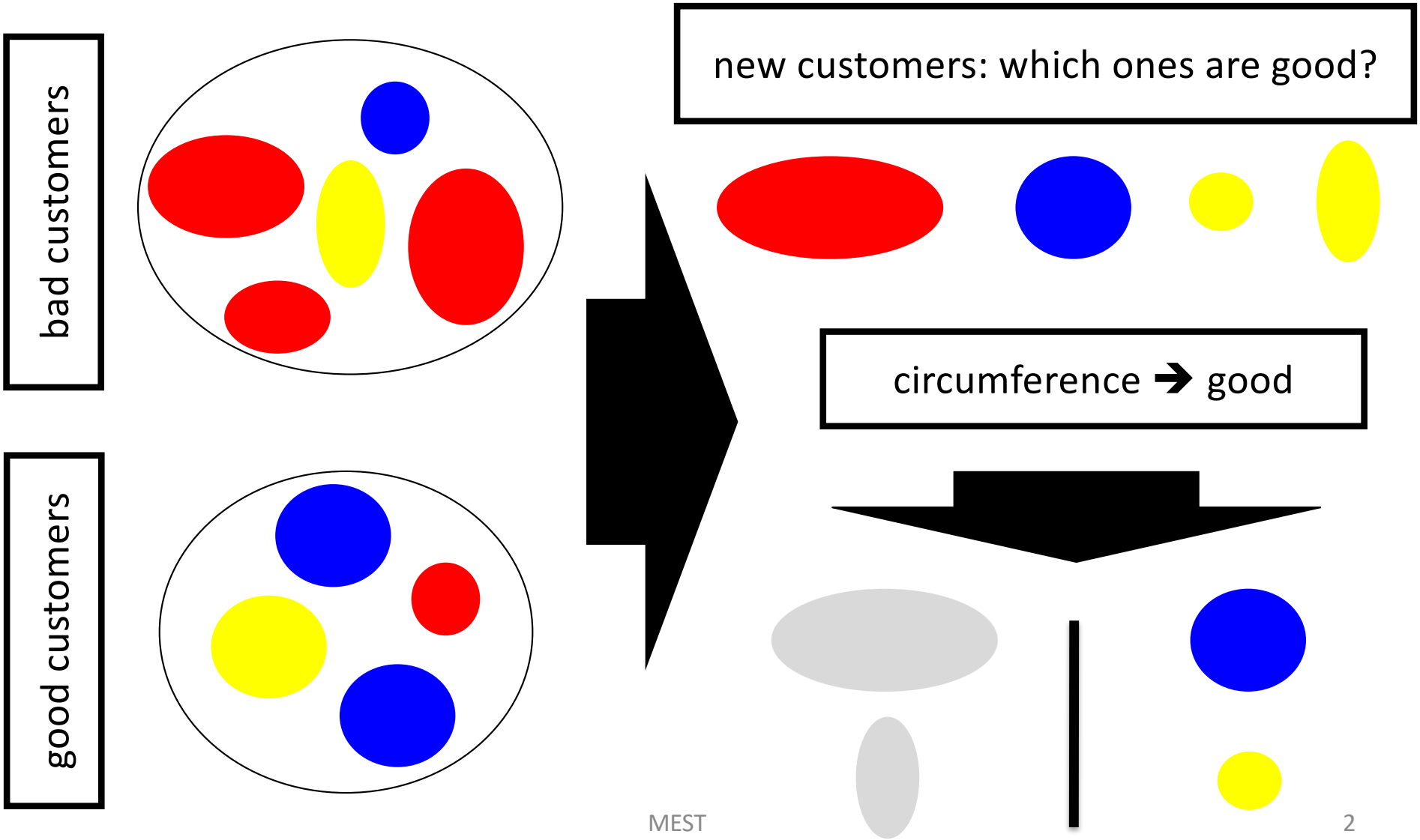
[including materials kindly provided by
A. Jorge, E. Keogh and J.M. Moreira]

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

http://www.flickr.com/photos/upv
cfrontdoor/4642103837/

predictive: classification for targeting

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



plan & goals

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- introduction to classification
  - what and what for
  - data
  - an example
- … classification algorithms
  - naive Bayes
- … and evaluation of classification models
  - measures
  - methodology

- identify problems where classification is useful
- understand the basic concepts of the naive Bayes algorithm
- know the simplest measures for evaluating classification models
  - error/accuracy
- understand the need to use different sets of data for modelling and for evaluation
- address classification problems using RapidMiner

- **understand the importance of basic statistics concepts for data science and artificial intelligence**

classification for campaign optimization

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

fonte: ferrari

- campaign to promote new vehicle
  - (large) list of prospects
  - invitations for test-drive
  - gifts
  - free phone line (800) for enquiries/reservations
- goal
  - reduce costs
  - maximize returns

- strategy
  - analyse response to previous campaigns
    - stored in a database
  - build customer relating customer characteristics and response
  - apply model to prospects
  - invite prospects selected by the model
    - [who bought last car more than 4 years ago]

data for classification

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- prospects
  - customers who didn't buy a car in the last 4 years
- … unlabelled data
  - new data

would like to predict

- results from previous campaigns
  - customers who were contacted and their response
- … labelled data
  - old data

already known

target (or dependent) variable

independent variable (or attribute)

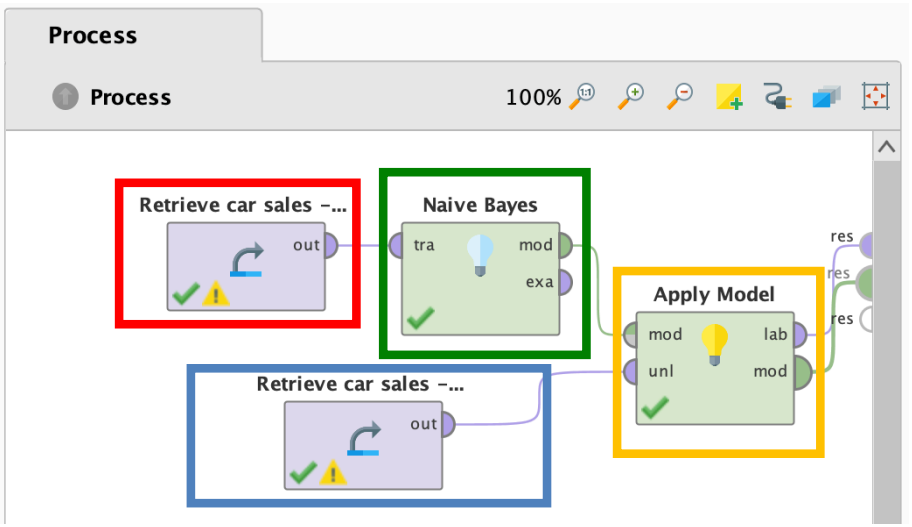
Comprou	Idade	Rendimento	Ag.fam	Vendas anteriores	Última Venda
	41	50000	2	1	0
	39	68000	2	0	30000
	58	61000	4	0	0
	26	25000	3	0	0
	21	50000	1	1	20000
	38	43000	2	0	0
	44	43000	4	1	47000
	27	47000	2	1	21000
	70	23000	2	0	25000

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

classification model in rapid miner

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- load old and new data into repository
  - target variable is a *label*!!
  - ... even for the new data
- load labelled data from repository into workspace
- apply naive bayes algorithm
  - operator: Naive Bayes
- load unlabelled data from repository into workspace
- apply naive bayes model to the new data
  - operator: Apply Model



predictions

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- predictions made by the model
- … and probability of each class

Row No.	Bought?	prediction(...	confidence(...	confidence(...	Age	Income	Family size	Cars boug...	Value of la...
1	?	nao	0.642	0.358	41	50000	2	1	0
2	?	sim	0.283	0.717	39	68000	2	0	30000
3	?	nao	0.524	0.476	58	61000	4	0	0
4	?	nao	0.935	0.065	26	25000	3	0	0
5	?	nao	0.869	0.131	21	50000	1	1	20000
6	?	nao	0.758	0.242	38	43000	2	0	0
7	?	sim	0.067	0.933	44	43000	4	1	47000
8	?	nao	0.704	0.296	27	47000	2	1	21000
9	?	nao	0.847	0.153	70	23000	2	0	25000

what was learned

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- probability distributions of the values of each attribute for each class

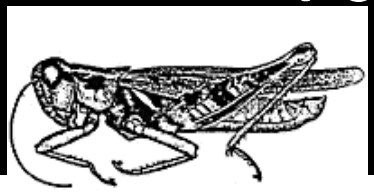
Attribute	Parameter	nao	sim
Age	mean	39.320	42.760
Age	standard deviation	10.061	7.862
Income	mean	52320	59040
Income	standard deviation	20241.706	18107.344
Family size	mean	2.040	2.260
Family size	standard deviation	0.978	1.046
Cars bought previously	mean	1.200	1.880
Cars bought previously	standard deviation	1.354	1.507
Value of last purchase	mean	16200	28100
Value of last purchase	standard deviation	14309.088	15110.815

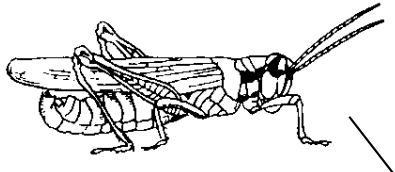
gps

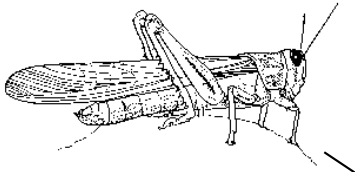
- introduction to classification
- … classification algorithms
  - naive Bayes
    - geometric intuition
    - probabilistic classification based on single variable
    - … and multiple variables
- … and evaluation of classification models

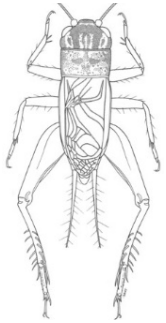
Thomas Bayes
1702 - 1761

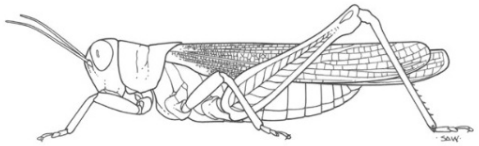
running (hopping?) example

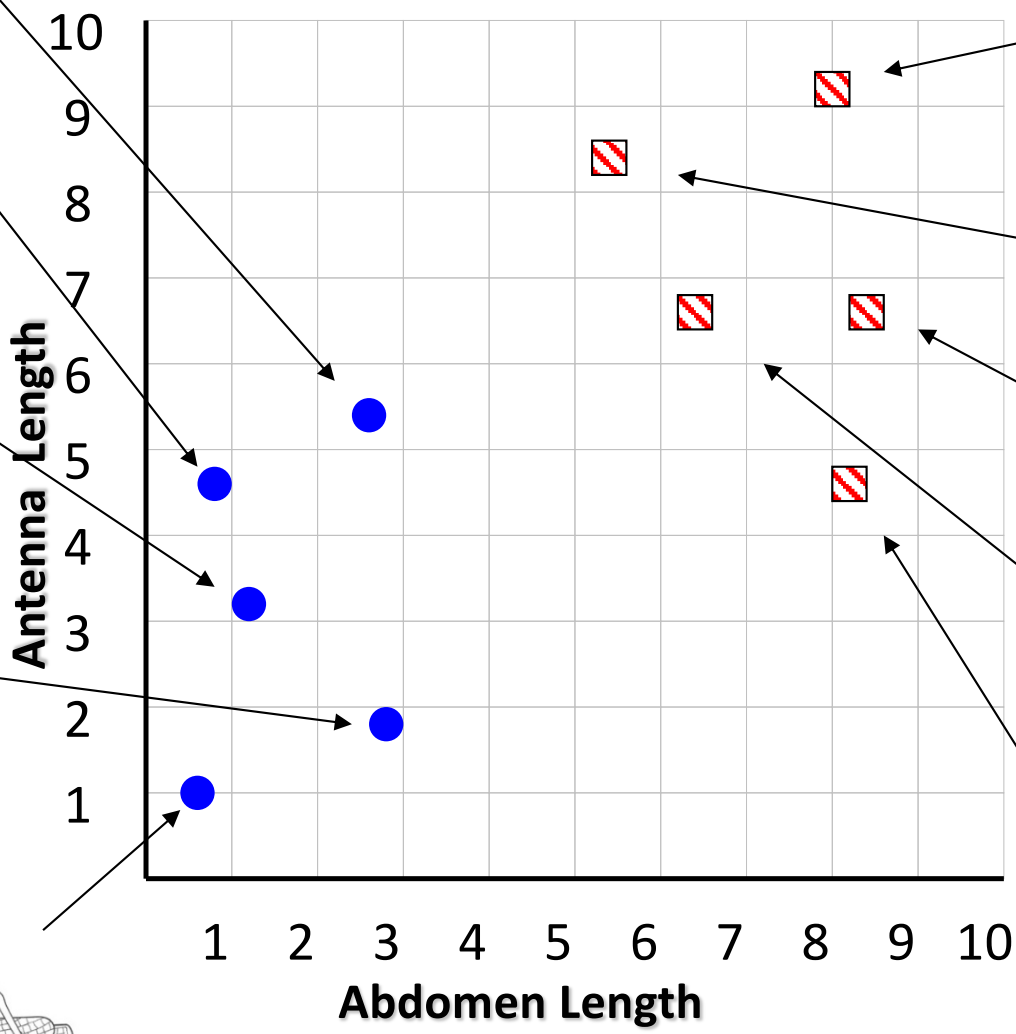


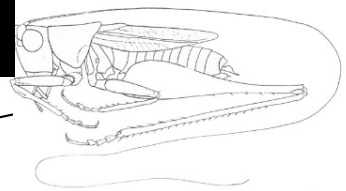


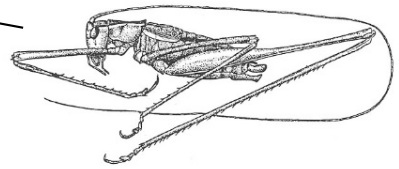


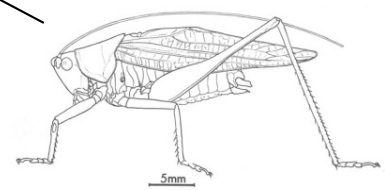


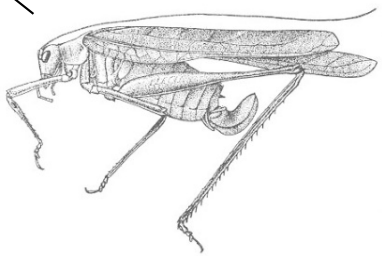












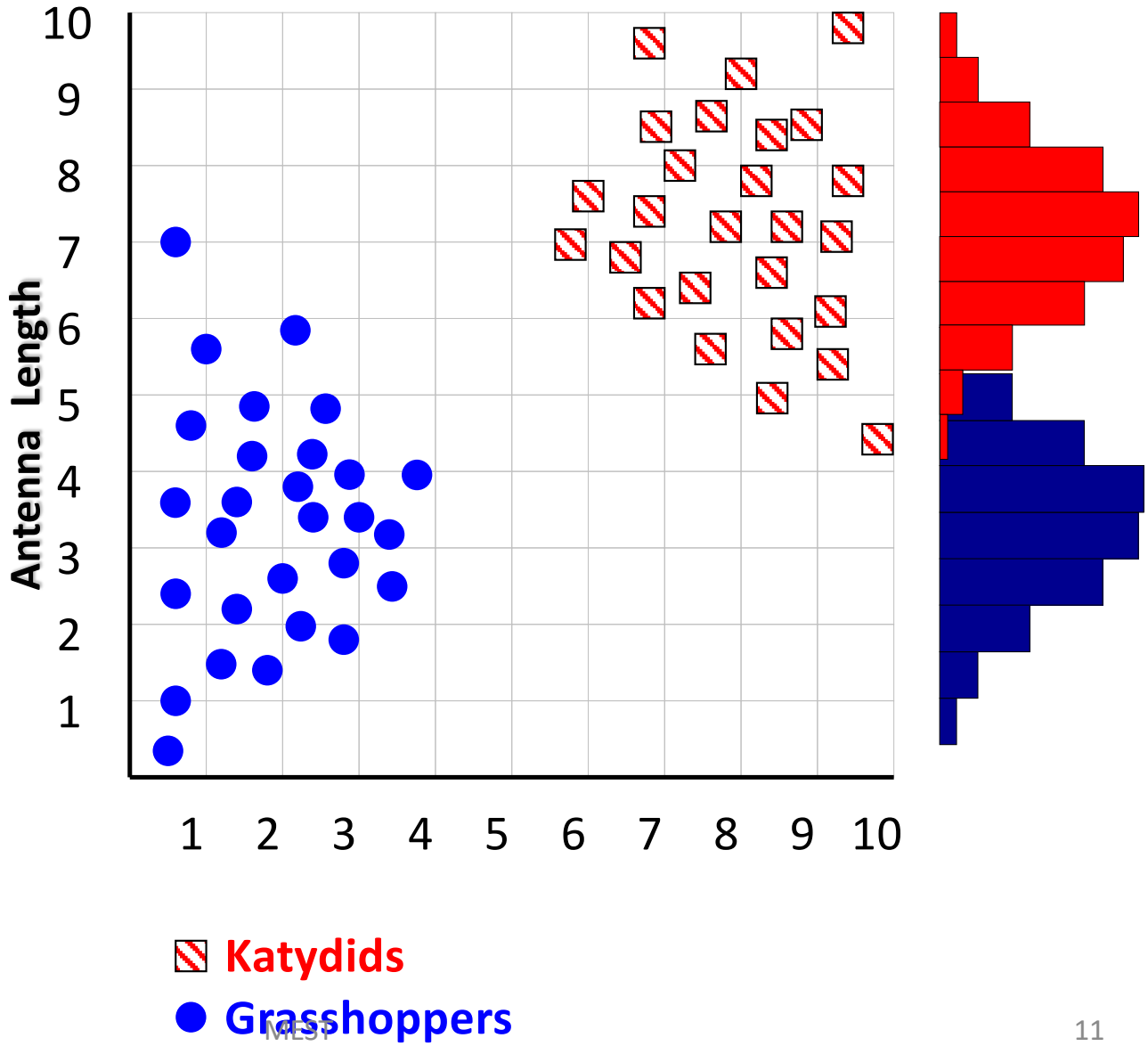
Grasshoppers

Katydids

The Plagues…

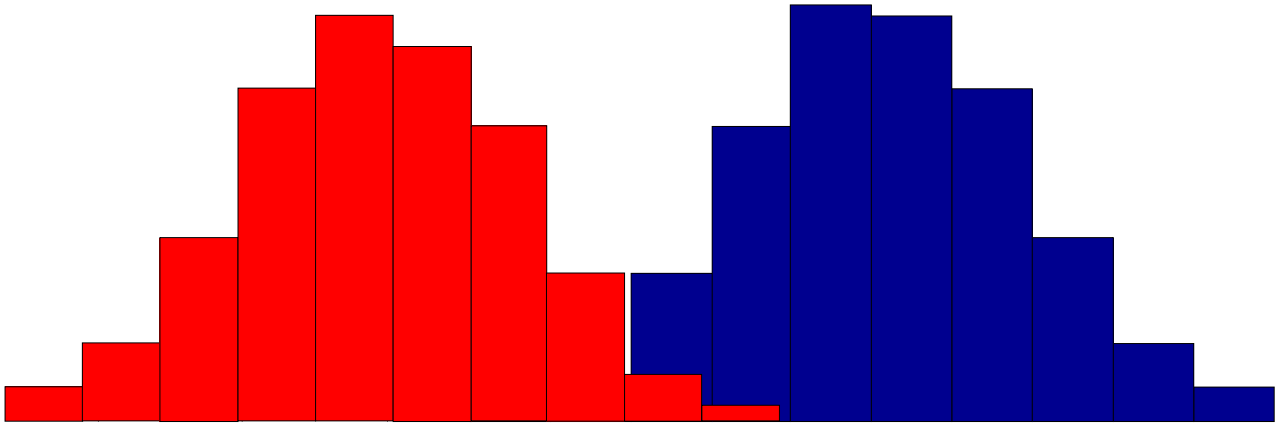
U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- Class histogram
  - eg “Antenna Length”

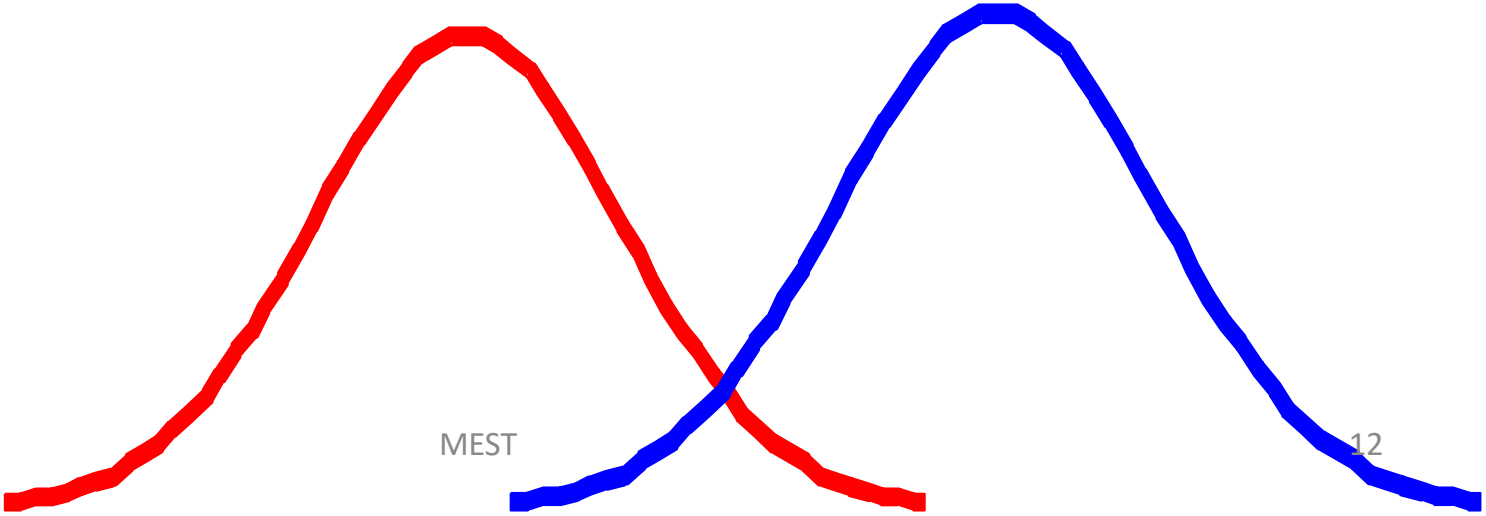


The Distribution of Plagues

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

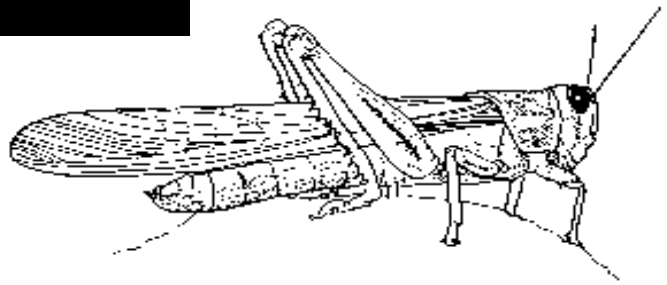


histograms can be summarized with two normal distributions



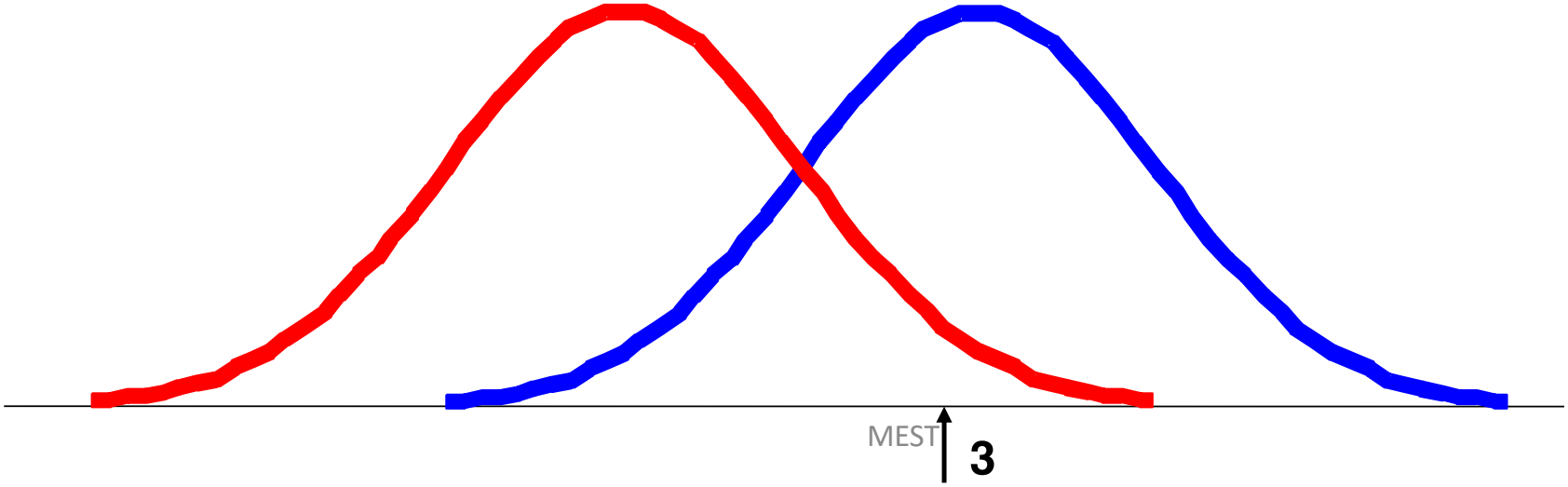
Bayes Classifies Bugs Based on Their Distribution (1/4)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO



- new insect, with are 3 units long antennae
- more *probably* a **Grasshopper** or a **Katydid**?
  - given the distributions of antennae lengths we have seen
- "the most *probable* classification" can be discussed formally…

$p(c_j \mid d)$ = probability of class c_j, *given* that we have observed d



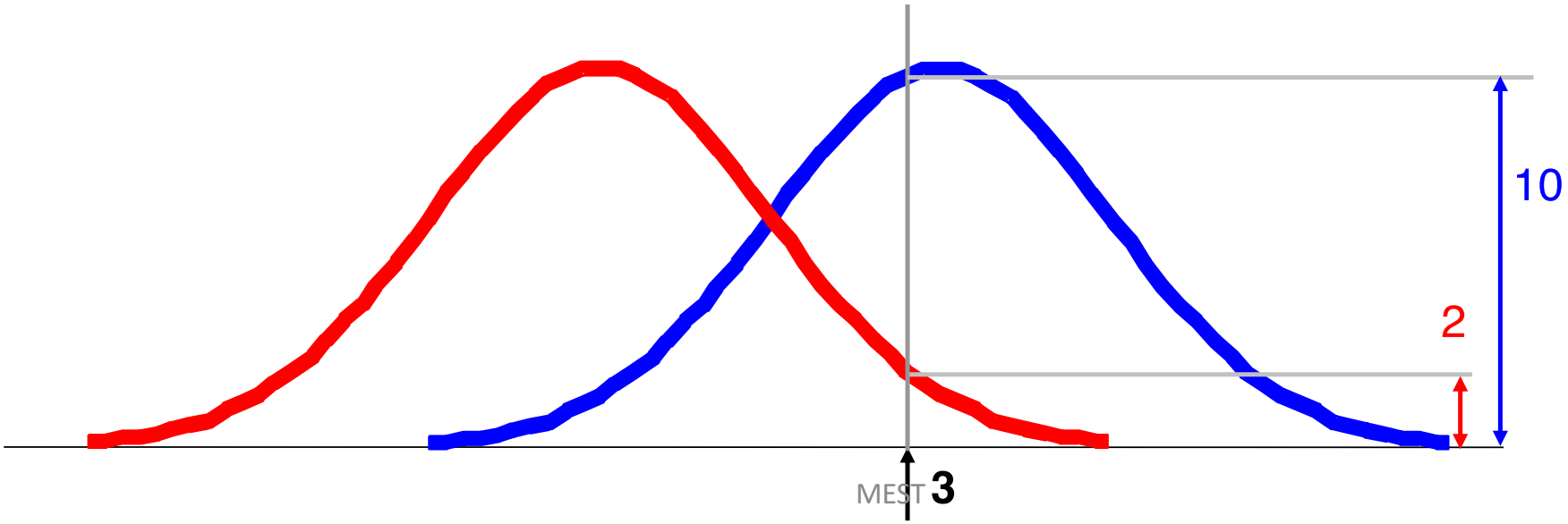
Bayes Classifies Bugs Based on Their Distribution (2/4)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **3**) = 10 / (10 + 2) = 0.833

P(**Katydid** | **3**) = 2 / (10 + 2) = 0.166



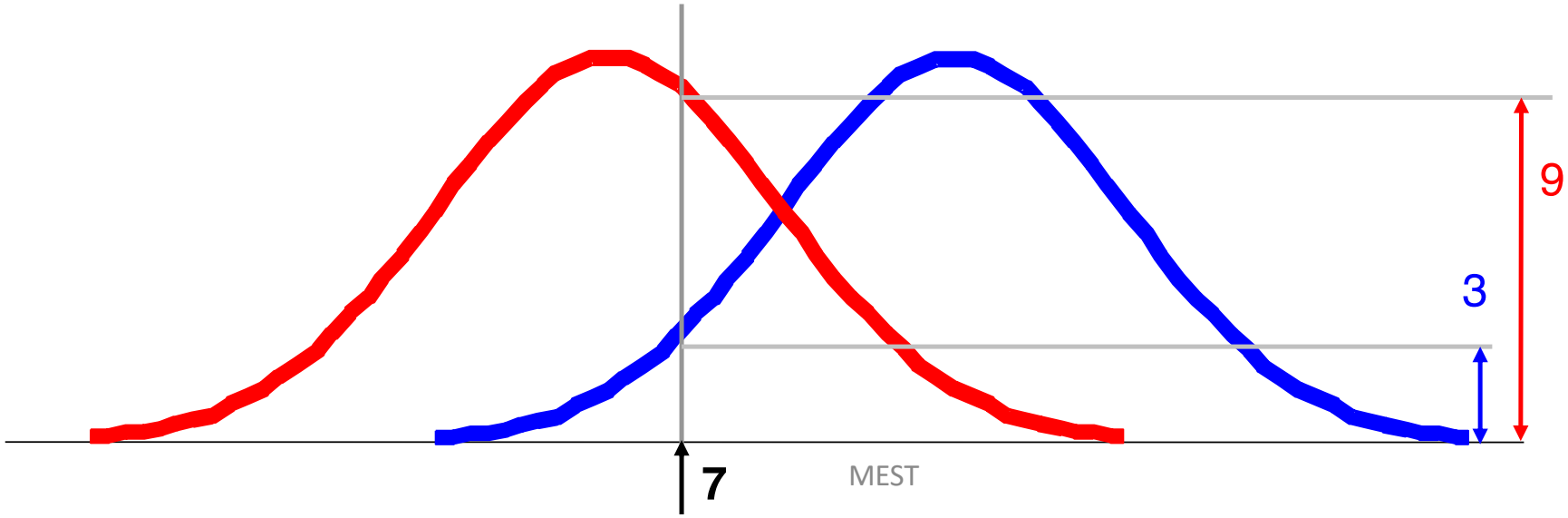
Bayes Classifies Bugs Based on Their Distribution (3/4)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **7**) = 3 / (3 + 9) = 0.250

P(**Katydid** | **7**) = 9 / (3 + 9) = 0.750



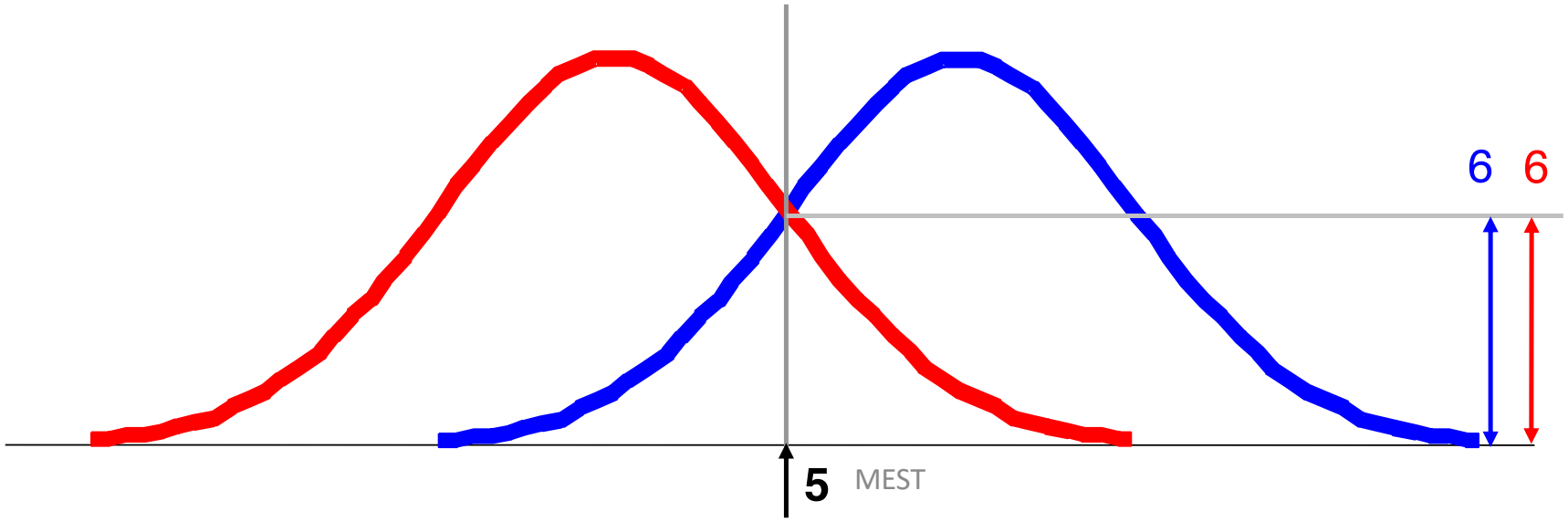
Bayes Classifies Bugs Based on Their Distribution (4/4)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

$p(c_j \mid d)$ = probability of class c_j, given that we have observed d

P(**Grasshopper** | **5**) = 6 / (6 + 6) = 0.500

P(**Katydid** | **5**) = 6 / (6 + 6) = 0.500



Bayes Classifiers

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- Bayesian classifiers use **Bayes theorem**, which says

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

  - $p(c_j \mid d)$ = probability of instance d being in class c_j,

This is what we are trying to compute

  - $p(d \mid c_j)$ = probability of generating instance d given class c_j,

We can imagine that being in class c_j, causes you to have feature d with some probability

  - $p(c_j)$ = probability of occurrence of class c_j,

This is just how frequent the class c_j, is in our database

  - $p(d)$ = probability of instance d occurring

This can actually be ignored, since it is the same for all classes

Assume that we have two classes

c_1 = male, and c_2 = female.

We have a person whose sex we do not know, say "*drew*" or *d*.

Classifying *drew* as male or female is equivalent to asking is it more probable that *drew* is male or female, I.e which is greater *p*(male | *drew*) or *p*(female | *drew*)

(Note: "Drew can be a male or female name")

Drew Barrymore

Drew Carey

What is the probability of being called "*drew*" given that you are a male?

$$p(\text{male} \mid drew) = \frac{p(drew \mid \text{male})\, p(\text{male})}{p(drew)}$$

What is the probability of being a male?

What is the probability of being named "*drew*"? (actually irrelevant, since it is that same for all classes)

Is Officer Drew a Male or Female? (1/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Officer Drew
(arrested E.K. – in 1997)

Luckily, we have a small database with names and sex.

We can use it to apply Bayes rule…

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Is Officer Drew a Male or Female? (2/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Officer Drew
(arrested E.K. – in 199

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

$$p(\text{male} \mid drew) = \frac{1/3 * 3/8}{3/8} = \frac{0.125}{3/8}$$

$$p(\text{female} \mid drew) = \frac{2/5 * 5/8}{3/8} = \frac{0.250}{3/8}$$

Name	Sex
Drew	Male
Claudia	Female
Drew	Female
Drew	Female
Alberto	Male
Karin	Female
Nina	Female
Sergio	Male

Officer Drew is more likely to be a Female

Is Officer Drew a Male or Female? (3/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

Officer Drew
(arrested E.K. – in 199

Officer Drew IS a female!

$$p(\text{male} \mid \textit{drew}) = \frac{1/3 * 3/8}{3/8} = \frac{0.125}{3/8}$$

$$p(\text{female} \mid \textit{drew}) = \frac{2/5 * 5/8}{3/8} = \frac{0.250}{3/8}$$

Officer Drew is more likely to be a Female

NB with Many Features (1/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

How do we use all the features?

$$p(c_j \mid d) = \frac{p(d \mid c_j)\, p(c_j)}{p(d)}$$

Name	Over 170cm	Eye	Hair length	Sex
Drew	No	Blue	Short	Male
Claudia	Yes	Brown	Long	Female
Drew	No	Blue	Long	Female
Drew	No	Blue	Long	Female
Alberto	Yes	Brown	Short	Male
Karin	No	Blue	Long	Female
Nina	Yes	Brown	Short	Female
Sergio	Yes	Blue	Long	Male

NB with Many Features (2/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- To simplify the task, **naive Bayesian classifiers** assume attributes have independent distributions, and thereby estimate

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * ....* p(d_n|c_j)$$

The probability of class c_j generating instance d, equals….

The probability of class c_j generating the observed value for feature 1, multiplied by..

The probability of class c_j generating the observed value for feature 2, multiplied by..

NB with Many Features (3/3)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- To simplify the task, **naive Bayesian classifiers** assume attributes have independent distributions, and thereby estimate

$$p(d|c_j) = p(d_1|c_j) * p(d_2|c_j) * ....* p(d_n|c_j)$$

$p(\text{officer drew}|c_j) = p(\text{over\_170}_{cm} = \text{yes}|c_j) * p(\text{eye} = blue|c_j) * ....$

Officer Drew is blue-eyed, over 170_{cm} tall, and has long hair

$p(\text{officer drew}|\ \text{Female}) = 2/5 * 3/5 * ....$

$p(\text{officer drew}|\ \text{Male}) = 2/3 * 2/3 * ....$

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

CLASSIFIER EVALUATION

classification: applying a model to new cases

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

Attribute	Parameter	nao	sim
Age	mean	39.320	42.760
Age	standard deviation	10.061	7.862
Income	mean	52320	59040
Income	standard deviation	20241.706	18107.344
Family size	mean	2.040	2.260
Family size	standard deviation	0.978	1.046
Cars bought previously	mean	1.200	1.880
Cars bought previously	standard deviation	1.354	1.507
Value of last purchase	mean	16200	28100
Value of last purchase	standard deviation	14309.088	15110.815

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

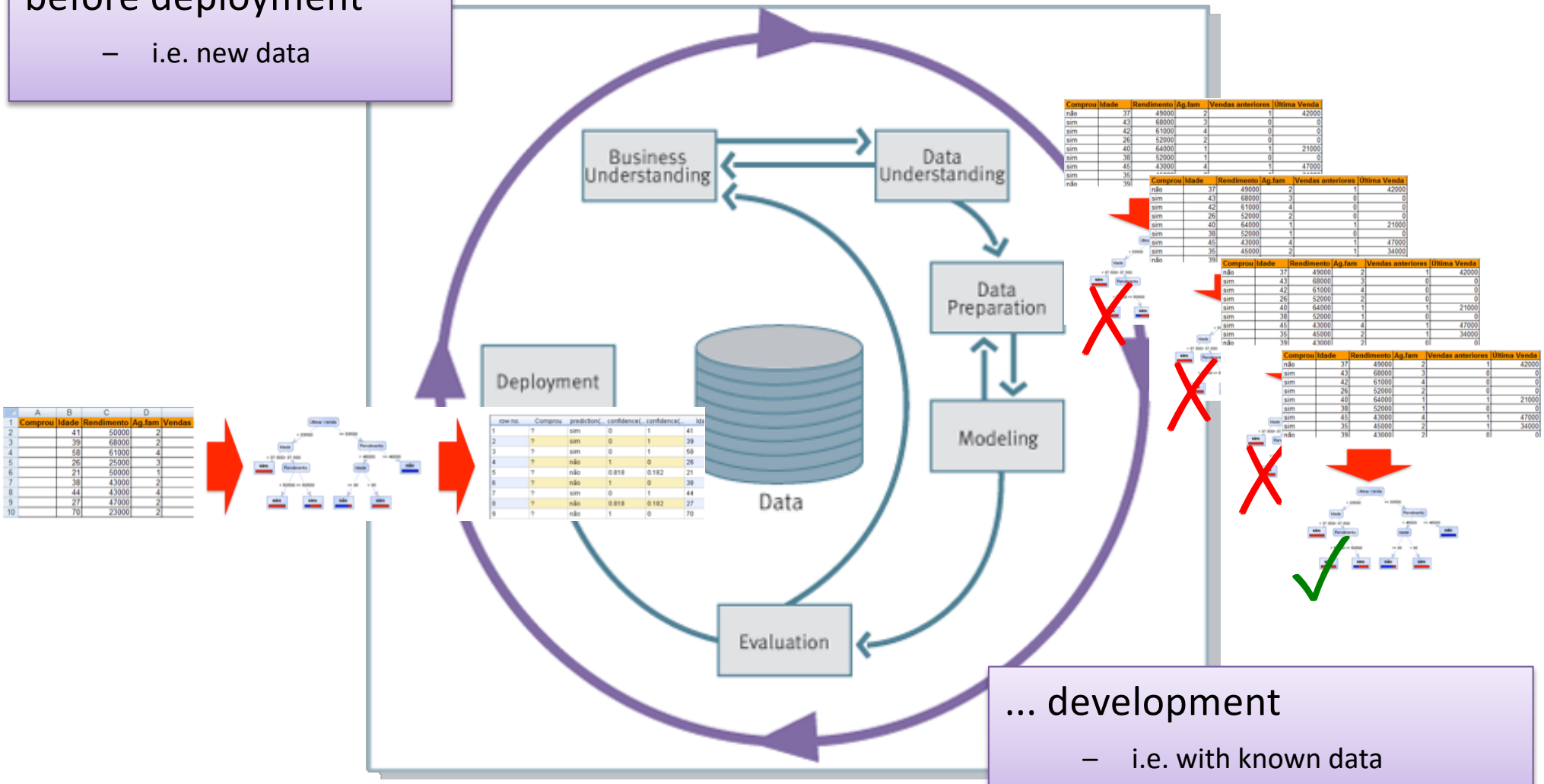
would you use the decisions proposed by this model?

model development

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

before deployment

- i.e. new data



... development

- i.e. with known data

gps

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- introduction to classification
- … classification algorithms
- … and evaluation of classification models
  - measures
  - methodology

what is the value of the model?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- what is the model for?
  - make predictions on new cases
    - ex. customers, campaigns, products
  - ... correctly
    - ex. contact good customers
    - reject bad customers
  - ... but failing sometimes
    - ex. reject good customers
    - or contact bad customers

- 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%
- applicable when there are > 2 classes

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

evaluation measures

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- many measures can be obtained from the confusion matrix

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

$\frac{FP}{FP+TN}$ False positive rate (FPR) = 1-TNR

$\frac{FN}{TP+FN}$ False negative rate (FNR) = 1-TPR

$\frac{TP}{TP+FN}$ True positive rate (TPR), also known as recall or sensitivity

$\frac{TN}{TN+FP}$ True negative rate (TNR), also known as specificity

$\frac{TP}{TP+FP}$ Positive predictive value (PPV), also known as precision

$\frac{TN}{TN+FN}$ Negative predictive value (NPV)

$\frac{TP+TN}{TP+TN+FP+FN}$ Accuracy

$\frac{2}{1/precision+1/recall}$ F1-measure

- … different perspectives concerning the value of the model

is the model of any use?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

	truth: no	truth: yes
class distribution	7	30

most “popular” class

- baseline model
  - simplest decision that can be obtained from the data
- baseline error: 7/37 = 18,9%

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

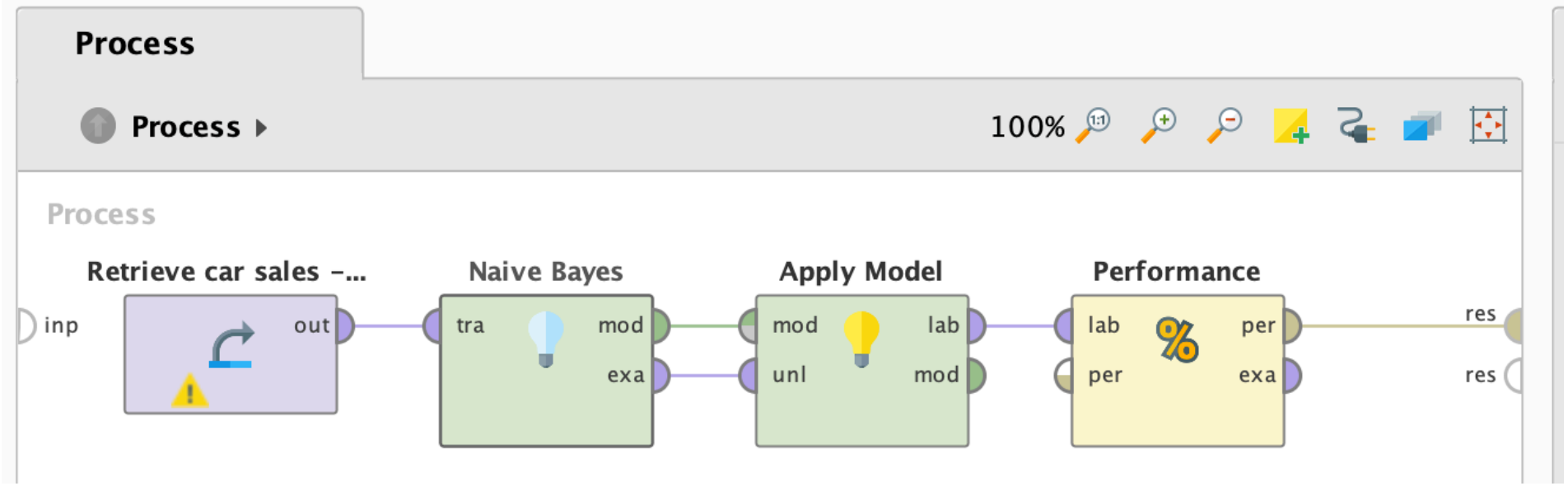
- model error: 3/37 = 8.1%

should we use the model?

exercise I: marketing campaign

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- evaluate decision tree
  - operator: performance (classification)

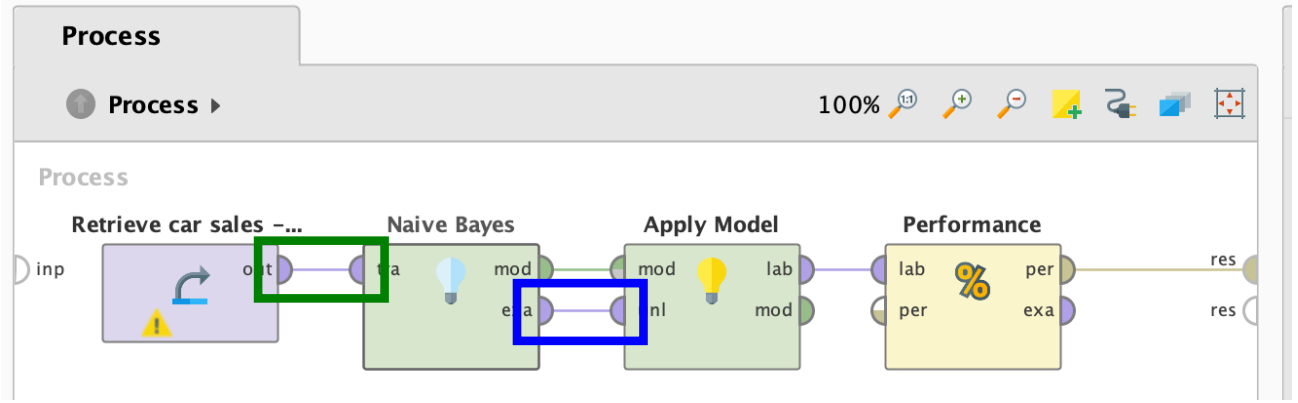


  - doesn’t this feel strange?

what is the value of the model?

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

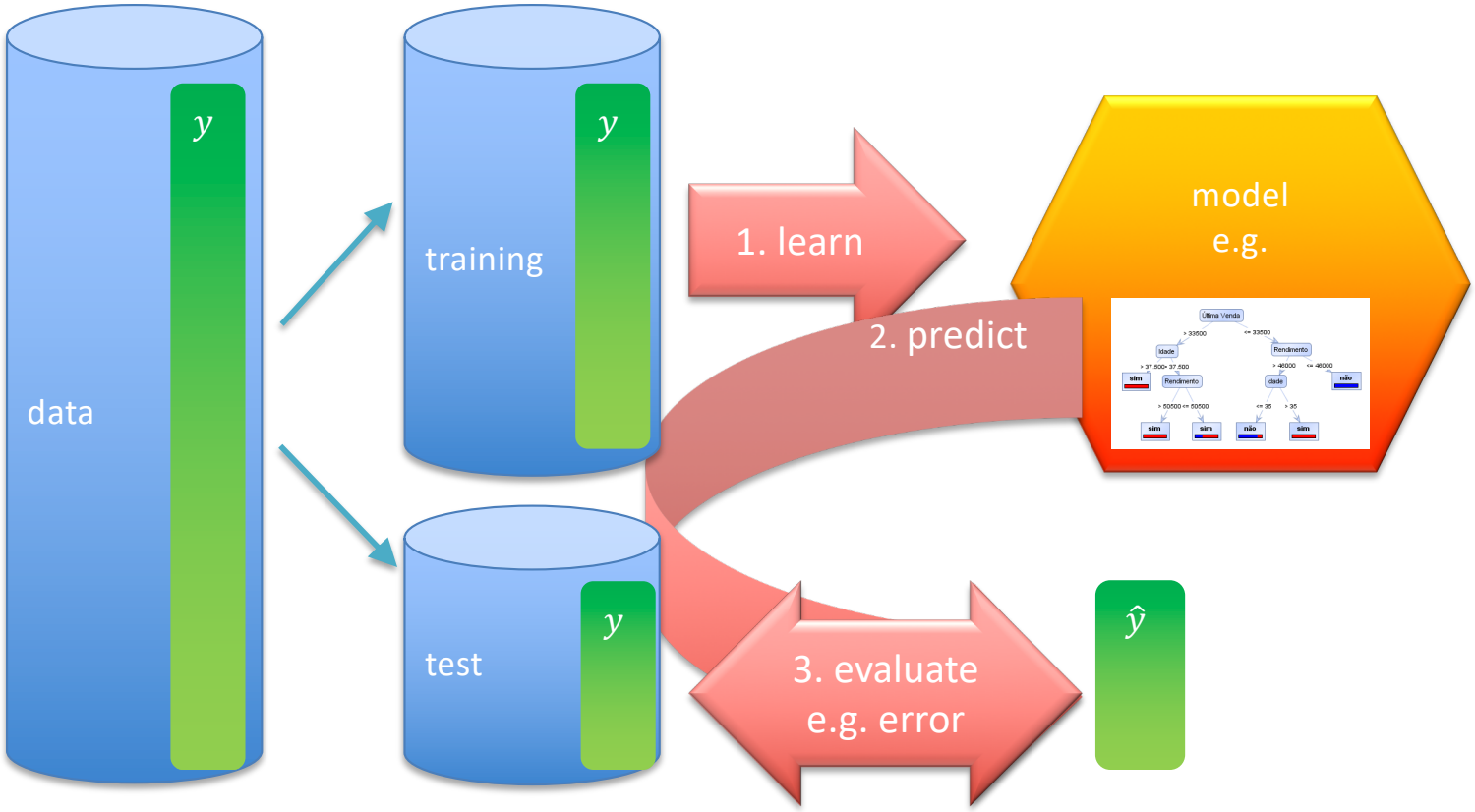
- goal: apply the model to **new** cases
- but, so far, same data to
  - **train** model
  - ... and evaluate it



- evaluation with training data
  - it's easy(ier) to make predictions in cases you already observed
  - assumes that future cases will be equal to those of training
    - similar to giving an exam with the same problems that were solved in the last class
  - unreliable estimator of model behavior in new examples

using past data to estimate predictive performance

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

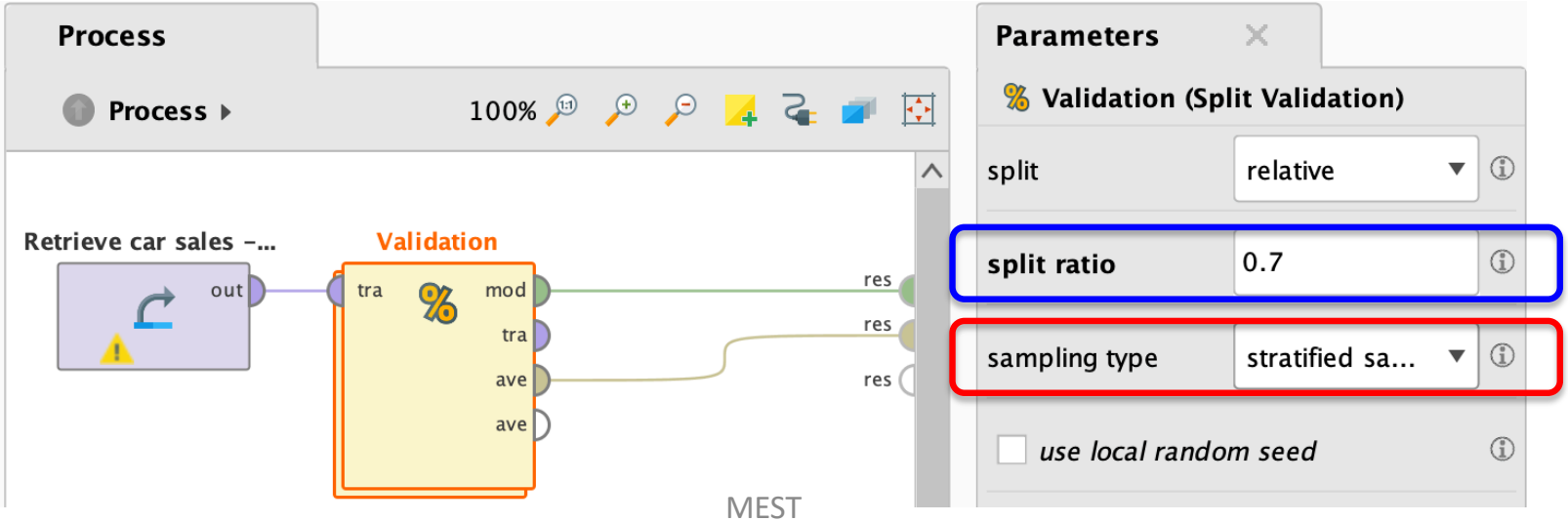


exercise II: marketing campaign (1/2)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- operator split validation
  - operator: split validation
- sub-process
  - operator that groups operators

- distribute data randomly between the training and test sets
  - ensuring the same class proportion
- proportion
  - 70% of the cases for training
  - 30% of the cases for testing



exercise II: marketing campaign (2/2)

U.PORTO
FEUP FACULDADE DE ENGENHARIA
UNIVERSIDADE DO PORTO

- split validation
  - different operations for training data and for test data
  - model obtained on the train side is passed on to the test side