Data Mining

Semi-supervised classification

Classification trees

24/11/2021

NOVA-IMS

Fernando Lucas Bação

bacao@isegi.unl.pt

http://www.isegi.unl.pt/fbacao

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

1

AGENDA

- Cluster analysis
  - Semi-supervised classification
    - Classification Trees

2

NOVA IMS Information Management School

Classification Trees

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

3

NOVA IMS Information Management School

Going from clustering into classification

Classification

- K-nearest neighbours — Use to classify new instances of the problem with the proper label;
- Classification Trees — Use to classify new instances and understand what distinguish different labels;

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

4

Classification Trees

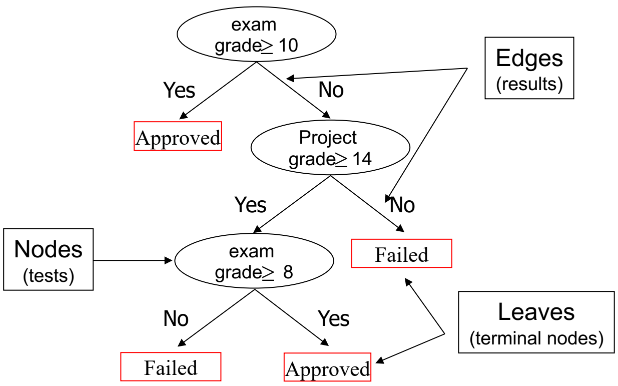
- **Classification Trees:**
  - Classification tree are typically considered to be classification and regression tools
  - One of its most important advantages relates with the simplicity of the interpretation of its results
  - Thus, the end result of a classification tree can easily be expressed in English or SQL.

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

5

Classification Trees

- A classification tree is a decisional algorithm
- It can be seen as a way of storing knowledge
- The objective is to discriminate between Class
- Obtain leaves as pure as possible
- If possible each leave should represent only individuals from a specific class



Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

6

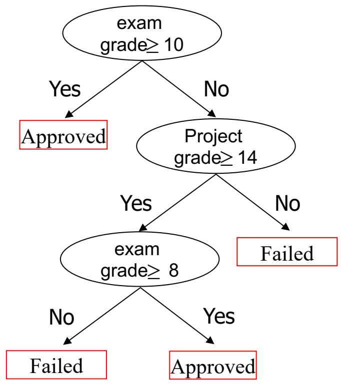
Classification Trees

$$aprovado \Leftrightarrow (exame \geq 10)$$

$$aprovado \Leftrightarrow (exame < 10) \wedge (projecto \geq 14) \wedge (exame \geq 8)$$

$$reprovado \Leftrightarrow (exame < 10) \wedge (projecto < 14)$$

$$reprovado \Leftrightarrow (exame < 10) \wedge (projecto > 14) \wedge (exame < 8)$$



Classification Trees

- **Classification Trees (strenghts):**
  - Interpretation
    - We can easily understand the reasons behind a specific classification decision
  - May use different types of data
    - Interval, ordinal, nominal, etc.
  - Insensitive to scale factors
    - Variables measured in different scales may be used without any type of normalization

NOVA IMS Information Management School

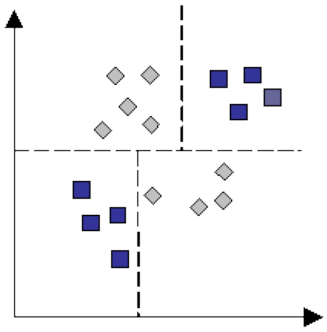
Classification Trees

- **Classification Trees (strenghts):**
  - Automatically defines the most relevant variables
    - These are the variables used at the top of the tree
  - Can be adapted to a regression
    - Each leave becomes a linear model

NOVA IMS Information Management School

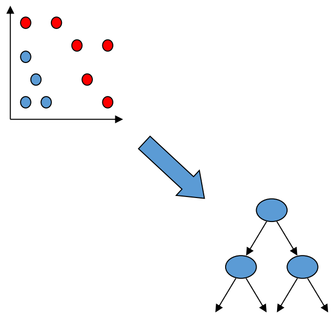
Classification Trees

- **Classification Trees (weaknesses):**
  - Boundaries are linear and perpendicular to the variables axys
  - Sensitive to small perturbations in the data



From Gahegan and West
http://divcom.otago.ac.nz/SIRC/GeoComp/GeoComp98/61/gc_61.htm

- **Classification Trees:**
  - Build (induce) a tree from data
  - Problems:
    - What to do?
      - Which variable to use?
      - What partition to use?
      - Which node to split?
      - How many edges per node?
      - When to stop?

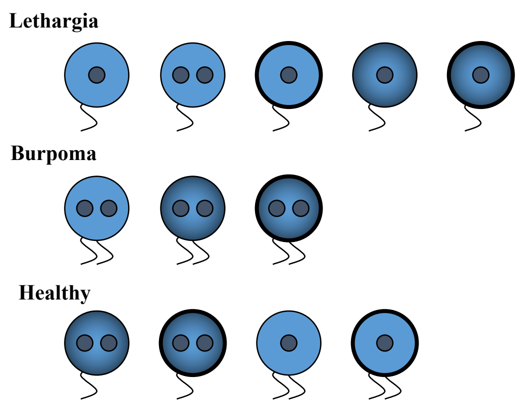


- **Classification Trees:**
  - ID3, C4.5 e C5 [Quinlan 86,93]
    - Iterative Dichotomizer 3
  - CART
    - Classification and regression trees [Breiman 84]
  - CHAID [Hartigan 75]
    - Used in SPSS and SAS…
  - Muitas (mesmo muitas) outras variantes...
    - In SAS you can choose different parameters to build your tree.

Worked-Example
General Idea

Langley, P: 1996, Elements of Machine Learning, Morgan and Kaufmann Publishers.

13



14

NOVA IMS Information Management School

Classification Trees

Table

# Nucleus	# Tails	Color	Membrane	*Class*
1	1	Light	Thin	*Lethargia*
2	1	Light	Thin	*Lethargia*
1	1	Light	Thick	*Lethargia*
1	1	Dark	Thin	*Lethargia*
1	1	Dark	Thick	*Lethargia*
2	2	Light	Thin	*Burpoma*
2	2	Dark	Thin	*Burpoma*
2	2	Dark	Thick	*Burpoma*
2	1	Dark	Thin	*Healthy*
2	1	Dark	Thick	*Healthy*
1	2	Light	Thin	*Healthy*
1	2	Light	Thick	*Healthy*

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

15

NOVA IMS Information Management School

Classification Trees

- **Classification Trees:**
  - Measure to discriminate the attribute

$$f(A) = \frac{1}{n}\sum_{i=1}^{|A|} C_i$$

  - n is the total number of examples and C_i the number of examples correctly classified based on the most frequent class.
  - This is a measure of “dominance” or “purity”

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

16

NOVA IMS Information Management School

Classification Trees

Table

# Nucleus	1	2
Lethargia	4	1
Burpoma	0	3
Healthy	2	2

Discrimination:

(4 + 3) / 12 = 0.58

# Nucleus	# Tails	Color	Membrane	Class
1	1	Light	Thin	Lethargia
2	1	Light	Thin	Lethargia
1	1	Light	Thick	Lethargia
1	1	Dark	Thin	Lethargia
1	1	Dark	Thick	Lethargia
2	2	Light	Thin	Burpoma
2	2	Dark	Thin	Burpoma
2	2	Dark	Thick	Burpoma
2	1	Dark	Thin	Healthy
2	1	Dark	Thick	Healthy
1	2	Light	Thin	Healthy
1	2	Light	Thick	Healthy

17

NOVA IMS Information Management School

Classification Trees

Table

# Tails	1	2
Lethargia	5	0
Burpoma	0	3
Healthy	2	2

Discrimination:

(5 + 3) / 12 = 0.67

# Nucleus	# Tails	Color	Membrane	Class
1	1	Light	Thin	Lethargia
2	1	Light	Thin	Lethargia
1	1	Light	Thick	Lethargia
1	1	Dark	Thin	Lethargia
1	1	Dark	Thick	Lethargia
2	2	Light	Thin	Burpoma
2	2	Dark	Thin	Burpoma
2	2	Dark	Thick	Burpoma
2	1	Dark	Thin	Healthy
2	1	Dark	Thick	Healthy
1	2	Light	Thin	Healthy
1	2	Light	Thick	Healthy

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

18

NOVA IMS Information Management School

Classification Trees

Table

Color	Light	Dark
Lethargia	3	2
Burpoma	1	2
Healthy	2	2

Discrimination:

(3 + 2) / 12 = 0.41

# Nucleus	# Tails	Color	Membrane	Class
1	1	**Light**	Thin	***Lethargia***
2	1	**Light**	Thin	***Lethargia***
1	1	**Light**	Thick	***Lethargia***
1	1	**Dark**	Thin	***Lethargia***
1	1	**Dark**	Thick	***Lethargia***
2	2	**Light**	Thin	***Burpoma***
2	2	**Dark**	Thin	***Burpoma***
2	2	**Dark**	Thick	***Burpoma***
2	1	**Dark**	Thin	***Healthy***
2	1	**Dark**	Thick	***Healthy***
1	2	**Light**	Thin	***Healthy***
1	2	**Light**	Thick	***Healthy***

NOVA IMS Information Management School

Classification Trees

Table

Membrane	Thin	Thick
Lethargia	3	2
Burpoma	2	1
Healthy	3	1

Discrimination:

(3 + 2) / 12 = 0.41

# Nucleus	# Tails	Color	Membrane	Class
1	1	Light	**Thin**	***Lethargia***
2	1	Light	**Thin**	***Lethargia***
1	1	Light	**Thick**	***Lethargia***
1	1	Dark	**Thin**	***Lethargia***
1	1	Dark	**Thick**	***Lethargia***
2	2	Light	**Thin**	***Burpoma***
2	2	Dark	**Thin**	***Burpoma***
2	2	Dark	**Thick**	***Burpoma***
2	1	Dark	**Thin**	***Healthy***
2	1	Dark	**Thick**	***Healthy***
1	2	Light	**Thin**	***Healthy***
1	2	Light	**Thick**	***Healthy***

NOVA IMS Information Management School

Classification Trees

Choice: # Tails

# Nucleus	1	2
Lethargia	4	1
Burpoma	0	3
Healthy	2	2

0.58

Color	Light	Dark
Lethargia	3	2
Burpoma	1	2
Healthy	2	2

0.41

# Tails	1	2
Lethargia	5	0
Burpoma	0	3
Healthy	2	2

0.67

Membrane	Thin	Thick
Lethargia	3	2
Burpoma	2	1
Healthy	3	1

0.41

NOVA IMS Information Management School

Classification Trees

Initial Partition

Tails

one

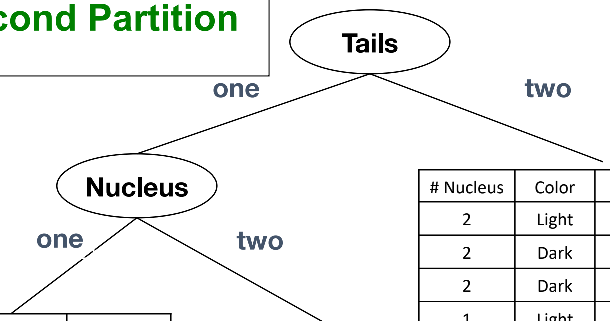
# Nucleus	Color	Membrane	Class
1	Light	Thin	Lethargia
2	Light	Thin	Lethargia
1	Light	Thick	Lethargia
1	Dark	Thin	Lethargia
1	Dark	Thick	Lethargia
2	Dark	Thin	Healthy
2	Dark	Thick	Healthy

two

# Nucleus	Color	Membrane	Class
2	Light	Thin	Burpoma
2	Dark	Thin	Burpoma
2	Dark	Thick	Burpoma
1	Light	Thin	Healthy
1	Light	Thick	Healthy

Classification Trees

Second Partition



Color	Membrane	Class
Light	Thin	Lethargia
Light	Thick	Lethargia
Dark	Thin	Lethargia
Dark	Thick	Lethargia

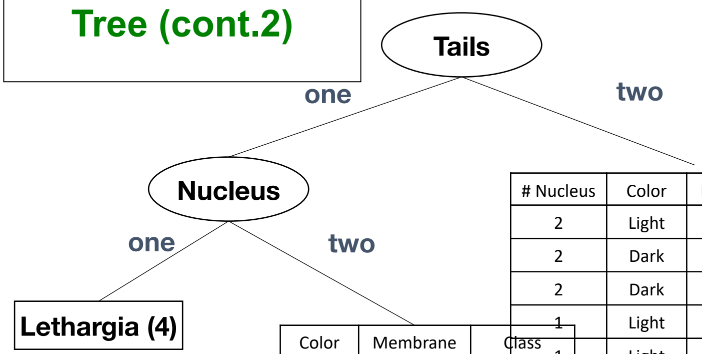
Color	Membrane	Class
Light	Thin	Lethargia
Dark	Thin	Healthy
Dark	Thick	Healthy

# Nucleus	Color	Membrane	Class
2	Light	Thin	Burpoma
2	Dark	Thin	Burpoma
2	Dark	Thick	Burpoma
1	Light	Thin	Healthy
1	Light	Thick	Healthy

23

Classification Trees

Tree (cont.2)



Color	Membrane	Class
Light	Thin	Lethargia
Dark	Thin	Healthy
Dark	Thick	Healthy

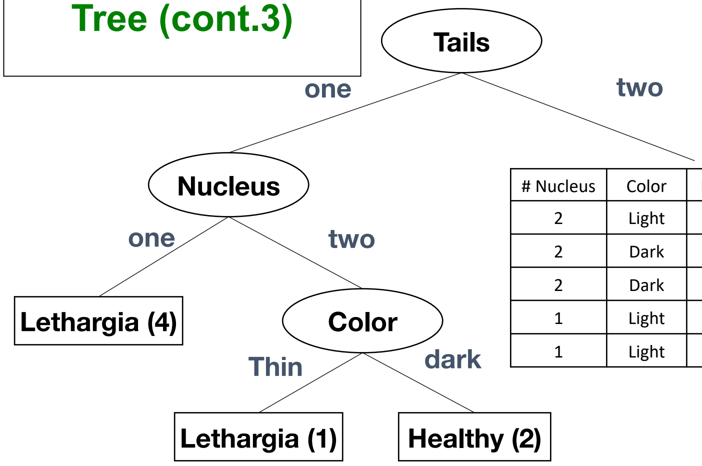
# Nucleus	Color	Membrane	Class
2	Light	Thin	Burpoma
2	Dark	Thin	Burpoma
2	Dark	Thick	Burpoma
1	Light	Thin	Healthy
1	Light	Thick	Healthy

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

24

Classification Trees

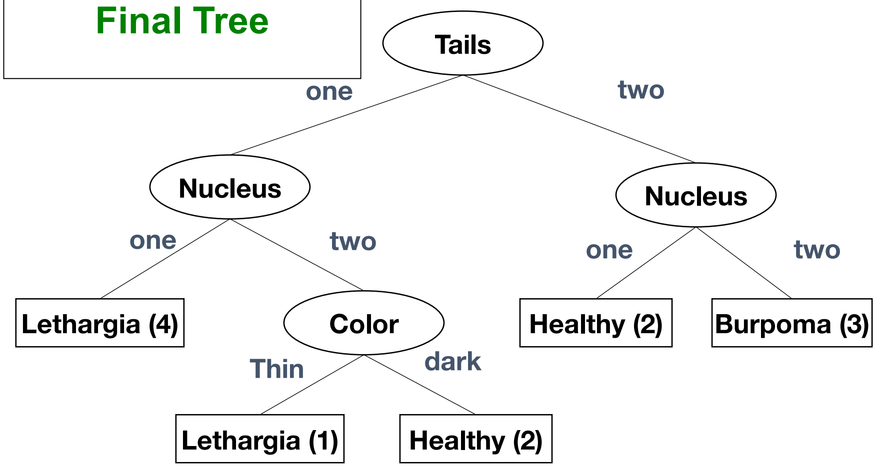
Tree (cont.3)



# Nucleus	Color	Membrane	Class
2	Light	Thin	Burpoma
2	Dark	Thin	Burpoma
2	Dark	Thick	Burpoma
1	Light	Thin	Healthy
1	Light	Thick	Healthy

Classification Trees

Final Tree



Classification Trees

The description of the three Classs

$$(tails = 1) \wedge (nucleous = 1)$$
$$\vee$$
$$(tails = 1) \wedge (nucleous = 2) \wedge (color = light) \rightarrow Lethargic$$

$$(tails = 2) \wedge (nucleous = 1)$$
$$\vee$$
$$(tails = 1) \wedge (nucleous = 2) \wedge (color = dark) \rightarrow Healthy$$

$$(tails = 2) \wedge (nucleous = 2) \rightarrow Burpoma$$

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

27

Classification Trees

- **Classification Trees:**
  - In each level it divides the set into alternative partitions.
    - Using a measure of quality selects the best partition.
  - The process is repeated for each element of the partition.
  - Stops when a given criteria is reached

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

28

NOVA IMS Information Management School

Classification Trees

- **Classification Trees:**
  - It assumes the existence of a target variable "Class" meaning the examples were previously classified.
  - Each node specifies a unique attribute which is used as test.
  - N – node N
  - ASET – Atribute Set
  - ISET – Instance Set

29

NOVA IMS Information Management School

Classification Trees

```
If Se the ISET is empty then the terminal node N is an unknown class
    if not
        If all the examples of ISET are of the same class
        then the terminal node N has the name of the class
            if not
                For each attribute A of the set of attribute ASET
                    Evaluate A acColording to its capability to discriminate a class
                    Select the attribute B which has the best discriminate value
                    For each value V of the best attribute B
                        Create a new node C from node N
                        Place the par attribute value (B, V) in C
                        Let JSET be the set of examples of ISET with value V in B
                        Let KSET be the set of attributes of ASET with B removed
                        DDT(C, KSET, JSET)
```

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

30

NOVA IMS Information Management School

Classification Trees

Worked-Example

Tree Accuracy

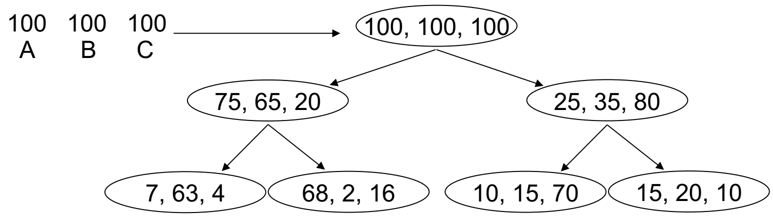
31

NOVA IMS Information Management School

Classification Trees

Quality of the results

Error rate

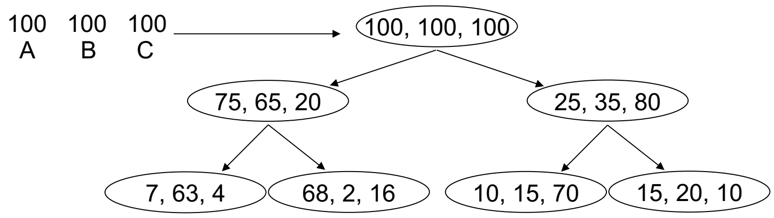


32

NOVA IMS Information Management School

Classification Trees

Quality of the results
Error rate



Error Calculation

$$TEA = \frac{(74 * 14{,}9\% + 86 * 20{,}9\% + 95 * 26{,}3\% + 45 * 55{,}5\%)}{300} = 26{,}3\%$$

Instituto Superior de Estatística e Gestão de Informação
Universidade Nova de Lisboa

33