

# Aprendizaje basado en árboles

Rubén Acevedo  
racevedo@bioingenieria.edu.ar

Tópicos Selectos en Aprendizaje Maquinal  
Doctorado en Ingeniería, FICH-UNL

9 de noviembre de 2007



# Tópicos Selectos de Aprendizaje Maquinal

## Cronograma de clases

- 1 Introducción y repaso (26/10/07)
- 2 Análisis estadístico (02/11/07)
- 3 **Aprendizaje basado en árboles (09/11/07)**
- 4 Aprendizaje clásico (16/11/07)
- 5 Aprendizaje basado en núcleos (23/11/07)
- 6 Aprendizaje de datos secuenciales (30/11/07)
- 7 Técnicas de validación (07/12/07)
- 8 Aplicaciones (13/12/07)

## Class organization

### 1 Introduction

### 2 Decision Trees

### 3 Classification and Regression Trees (CART)

Number of split.

Query selection and node impurity.

When to stop splitting.

Pruning.

Assignment of leaf node labels.



























## Query selection and node impurity

- *Gini impurity*

$$i(N) = -\sum_{i \neq j} P(\omega_i|N)P(\omega_j|N) = \frac{1}{2} \left[ 1 - \sum_j P^2(\omega_j|N) \right]$$

Is the expected error rate at node  $N$  if the category label is selected randomly from the class distribution present at  $N$ .

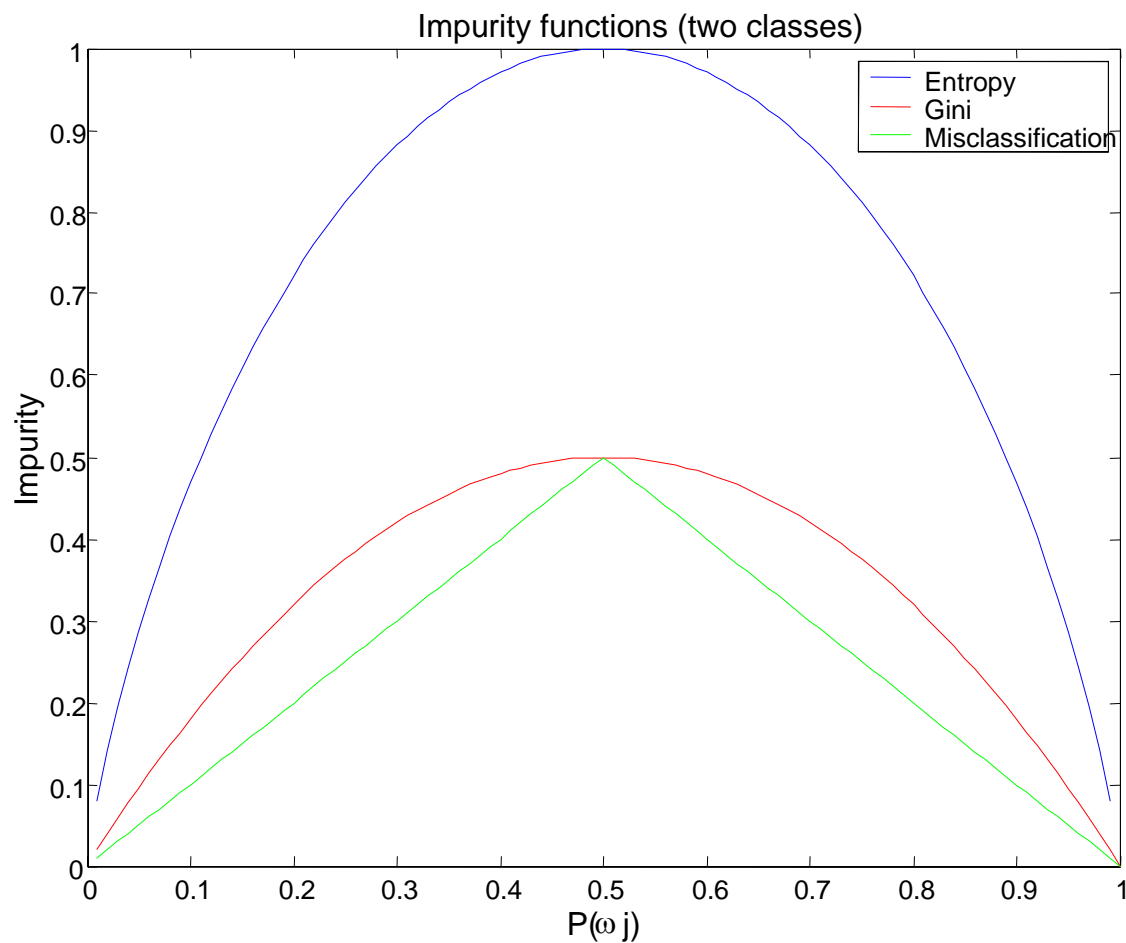
## Query selection and node impurity

- *Misclassification impurity*

$$i(N) = 1 - \max_j P(\omega_j | N)$$

Is a measure of the minimum probability that a training pattern would be misclassified at node  $N$ .

## Query selection and node impurity



## Query selection and node impurity

Given a partial tree down to node  $N$ , what value  $S$  should be choose for the property test T?

- *Impurity decrement (information gain)*

$$\Delta i(N) = i(N) - P_L i(N_L) - (1 - P_L) i(N_R)$$

where  $N_L$  and  $N_R$  are the left and right descendent nodes,  $i(N_L)$  and  $i(N_R)$  are their impurities, and  $P_L$  is the fraction of pattern at node  $N$  that will go to  $N_L$  when the property query T is used.

## When to stop splitting

Extreme cases:

- Each leaf correspond to a single training pattern.

Full tree is a *look up table* and not generalize well.

- Splitting is sttoped too early.

High training error and poor performace.

## When to stop splitting

- Without limits.
- Cross-validation.
- Set a small threshold ( $\max_S \leq \beta$ )
- Number of patterns in a node.

## Example: classification of animals

Patrones: (*pelos, nadan, color, tamaño*) → clase

Atributos o características:

*pelos: si / no.*

*nadan: si / no.*

*color: blanco, gris, marrón.*

*tamaño: pequeño, mediano, grande.*

Clases: (canguro, delfín, ballena) = (A, B, C)

## Example: classification of animals

Training patterns:

*(si, no, gris, mediano)* → canguro.

*(si, no, marrón, mediano)* → canguro.

*(no, si, gris, grande)* → delfín.

*(no, si, blanco, mediano)* → delfín.

*(no, si, marrón, grande)* → ballena.

## Example: classification of animals

Splitting root node

a) Calculation of *entropy impurity*

$$I(A, B, C) = -P(A) \cdot \log_2 P(A) - P(B) \cdot \log_2 P(B) - P(C) \cdot \log_2 P(C)$$

$$P(A) = \frac{a}{a+b+c}, \quad P(B) = \frac{b}{a+b+c}, \quad P(C) = \frac{c}{a+b+c} \quad a = 2, \quad b = 2, \quad c = 1$$

$$I(A, B, C) = -\frac{2}{5} \cdot \log_2 \frac{2}{5} - \frac{2}{5} \cdot \log_2 \frac{2}{5} - P\left(\frac{1}{5}\right) \cdot \log_2 P\left(\frac{1}{5}\right) = 1.5219 \text{ bits}$$

## Example: classification of animals

*pelos = si*

$$a = 2, b = 0, c = 0$$

$$I(a_1, b_1, c_1) = -\frac{2}{2} \cdot \log_2 \frac{2}{2} - 0 \cdot \log_2 P(0) - 0 \cdot \log_2 P(0) = 0 \text{ bits}$$

*pelos = no*

$$a = 0, b = 2, c = 1$$

$$I(a_2, b_2, c_2) = -0 \cdot \log_2 0 - \frac{2}{3} \cdot \log_2 P\left(\frac{2}{3}\right) - \frac{1}{3} \cdot \log_2 P\left(\frac{1}{3}\right) = 0.9183 \text{ bits}$$

## Example: classification of animals

b) Calculation of *impurity decrement* for attribute “*pelos*”

$$\Delta i(\textit{pelos}) = I(A, B, C) - \frac{2}{5}I(a_1, b_1, c_1) - \frac{3}{5}I(a_2, b_2, c_2) = 1.5219 - 0.5509 = 0.9710$$

Likewise for other attributes:

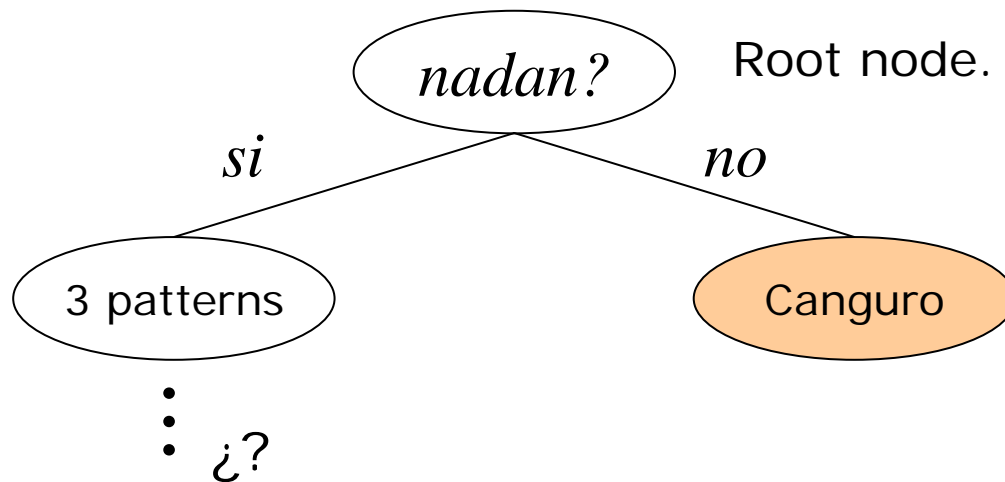
$$\Delta i(\textit{nadan}) = 0.9710 \quad \leftarrow \text{Selected attribute for splitting}$$

$$\Delta i(\textit{color}) = 0.7219$$

$$\Delta i(\textit{tamaño}) = 0.5170$$

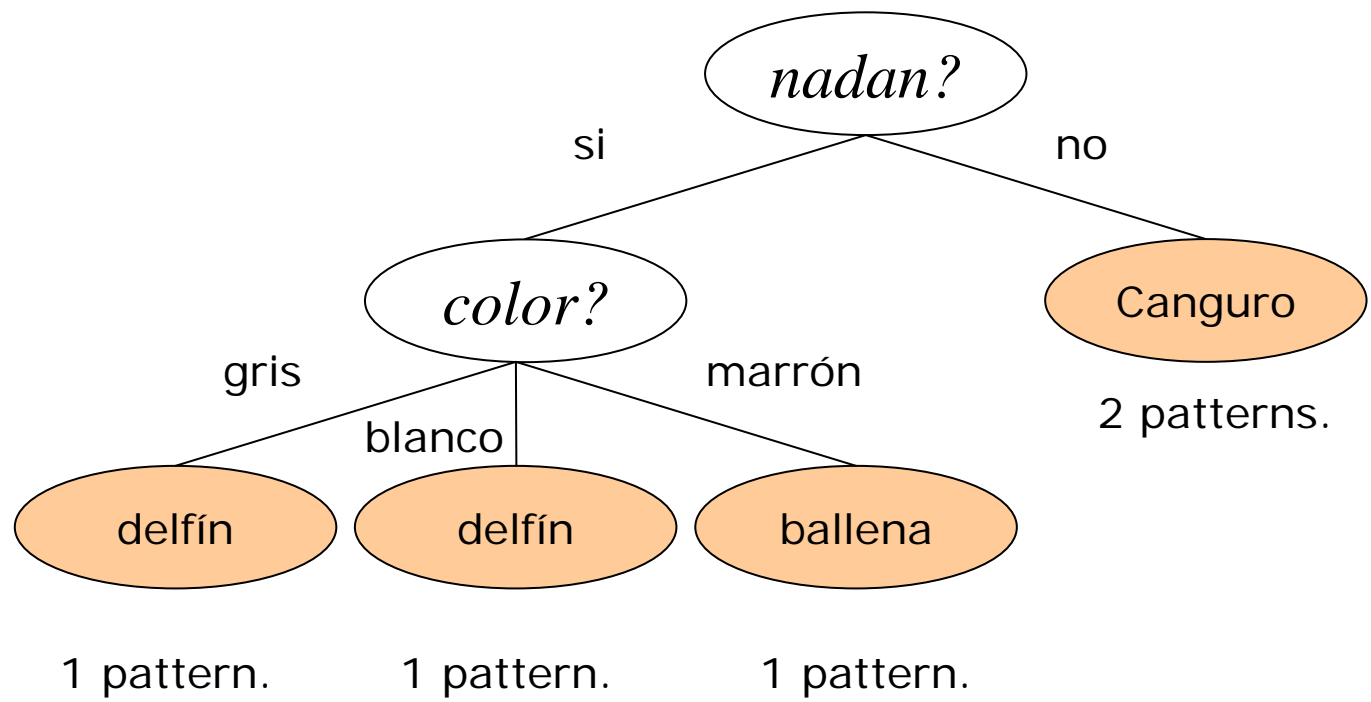
$$\Delta i(\textit{pelos}) = 0.9710$$

# Example: classification of animals

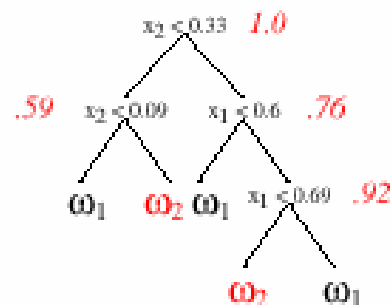
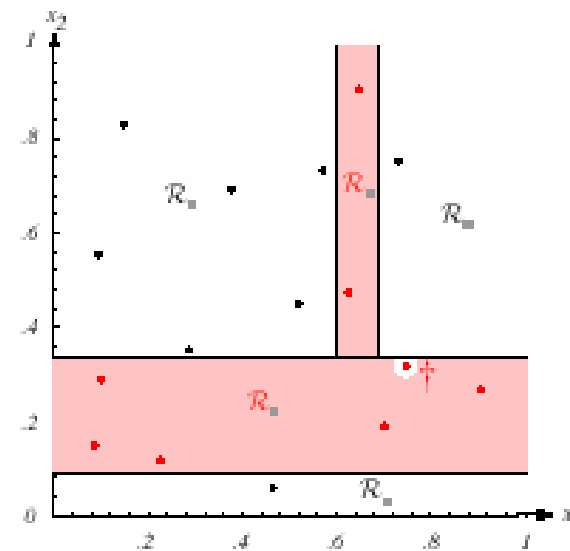
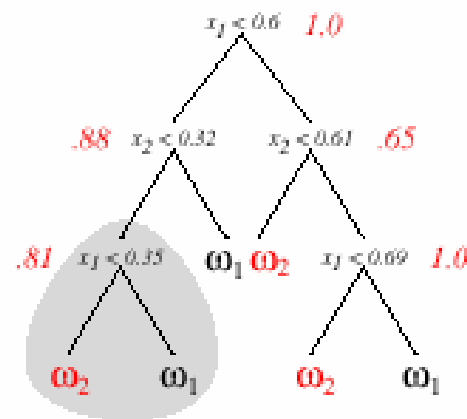
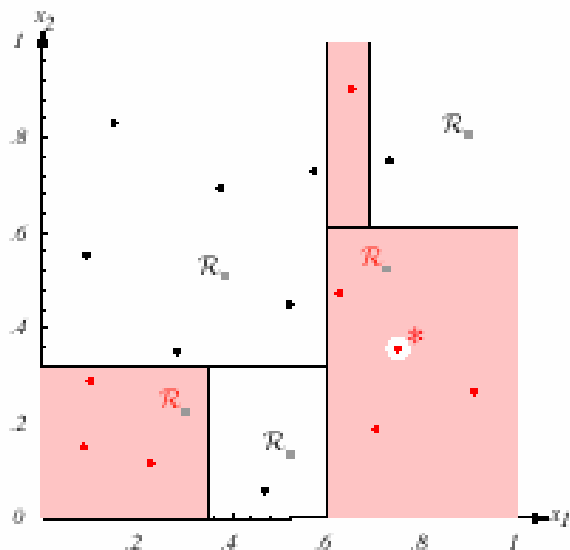


# Example: classification of animals

Final decision tree



Example: real-valued data.



# Pruning

Large decision trees can have two problems:

a) Overfitting.

b) Understanding and interpreting is a complicated process.

- *Complexity of a tree*

The complexity of a tree is measured by the number of its terminal nodes.

Trade-off between *accuracy* and *tree complexity*.

# Pruning

Assumption: in a node all classes have the same probability .

S: set of patterns.

N: quantity of patterns.

K: quantity of classes.

n: quantity of patterns of the class with more examples in the node

$$E(S) = \frac{N - n + k - 1}{N + k} \quad \text{Expected probability of error classification}$$

# Pruning

- If the node is a leaf

$$Error( S ) = E( S )$$

- If the node is internal

$$BackedUpError( A ) = \sum_i p_i Error( A_i )$$

$p_i$  : relative frequency of the patterns that are in the node  $A_i$ .

$$Error( A ) = \min\{ E( A ), BackedUpError( A ) \}$$

If  $E(A) < BackedUpError(A)$  then *pruning*.





## Assignment of leaf node labels

- Leaf node impurity = 0

Patterns of a single class → label is assigned to the leaf.

- Leaf node impurity > 0

Label should be labeled by the class that has more patterns represented.

## General considerations

- There are not noisy data.
- Examples are adequate.
- If there are new examples must be the tree to grow again.

## Exercise

Using the data in Table generate a decision tree and classify the following pattern:

(age  $\leq$  30, income=medium, student=yes, credit-rating=fair)

<i>RID</i>	<i>age</i>	<i>income</i>	<i>student</i>	<i>credit_rating</i>	<i>Class: buys_computer</i>
1	$\leq 30$	high	no	fair	no
2	$\leq 30$	high	no	excellent	no
3	31 ... 40	high	no	fair	yes
4	$> 40$	medium	no	fair	yes
5	$> 40$	low	yes	fair	yes
6	$> 40$	low	yes	excellent	no
7	31 ... 40	low	yes	excellent	yes
8	$\leq 30$	medium	no	fair	no
9	$\leq 30$	low	yes	fair	yes
10	$> 40$	medium	yes	fair	yes
11	$\leq 30$	medium	yes	excellent	yes
12	31 ... 40	medium	no	excellent	yes
13	31 ... 40	high	yes	fair	yes
14	$> 40$	medium	no	excellent	no

## Bibliography

- ✓ R.O. Duda, P.E. Hart, D.G. Stork, “*Pattern Classification*”, 2<sup>o</sup> Ed., Wiley-Interscience, 2001.

The End of Part One!