Machine Learning

Carlos García Martínez

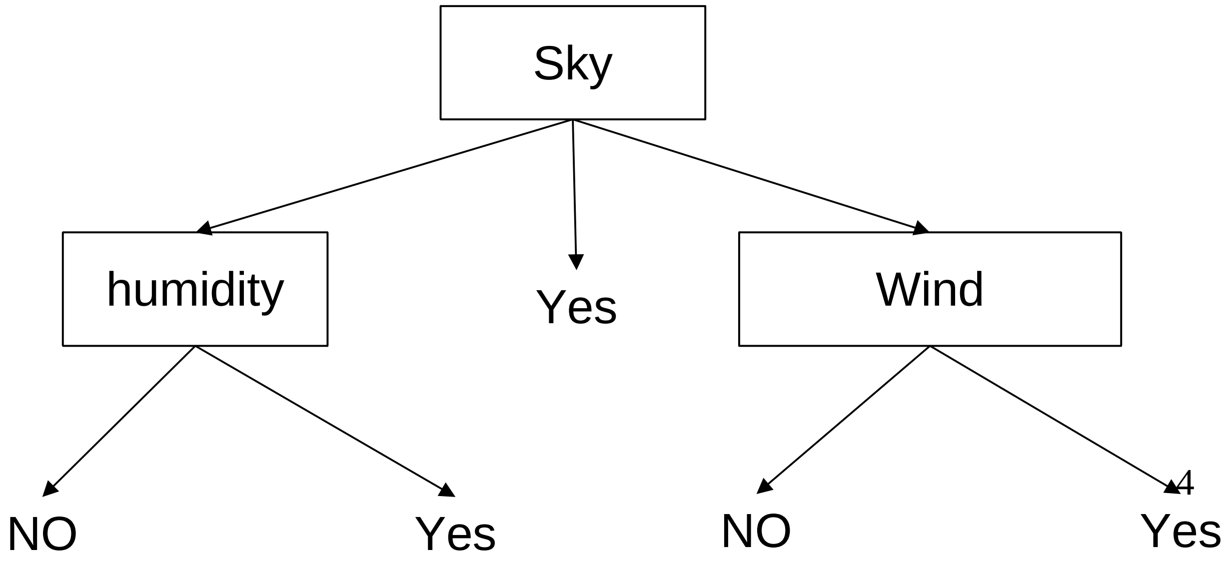
Machine Learning

- Decision trees
- The perceptron and neural networks
- Unsupervised learning

Decision Trees

Introduction

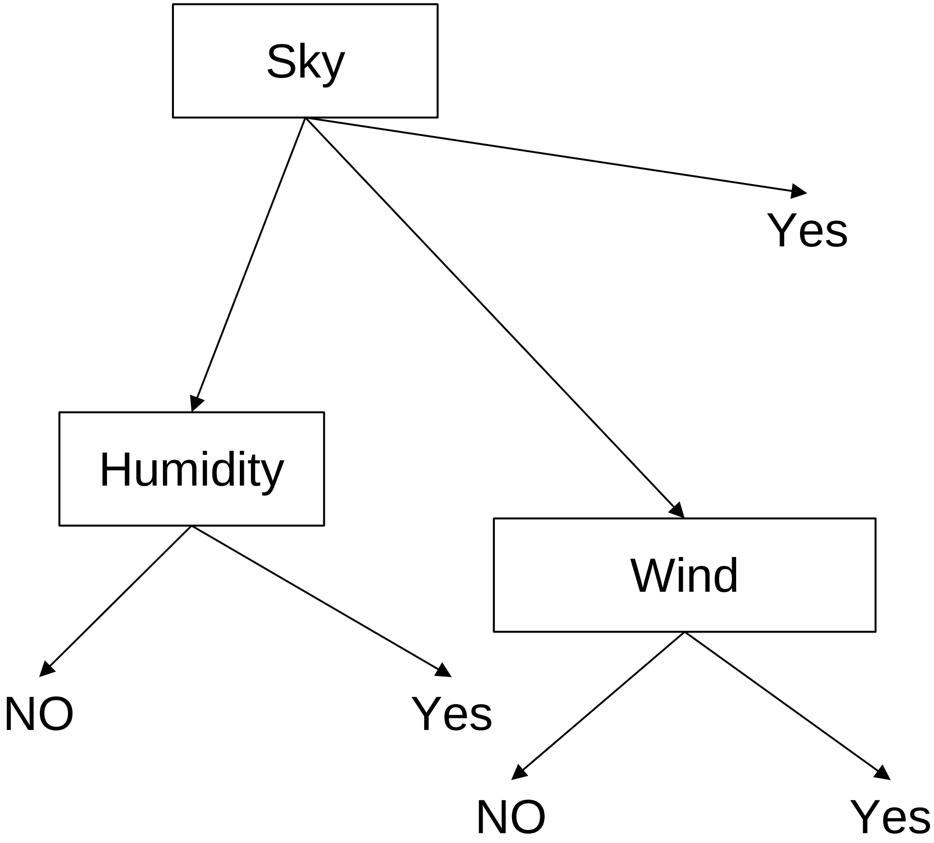
- Usage: Classification.
- Tree representation
  - Not terminal nodes: represent a question on an attribute.
  - Leaf nodes: answers or classes.
- Examples: Medical diagnostic, loan availability, ¿is today a good day to play tennis?



Learning

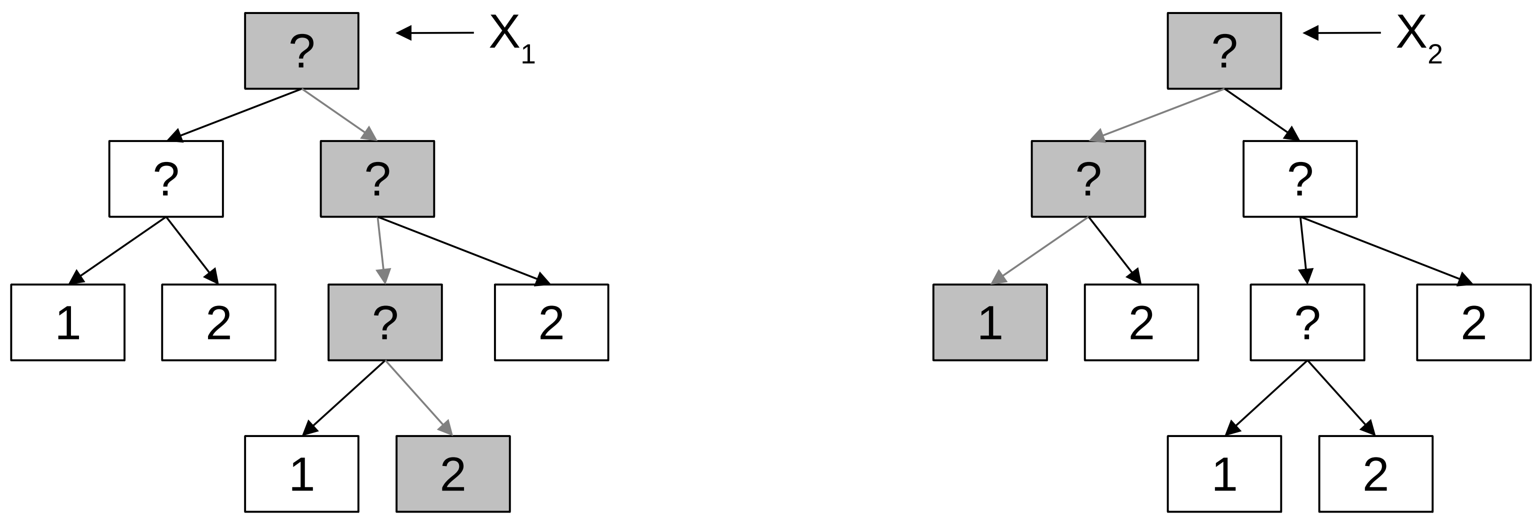
- To construct the tree from a data set:

day	*sky*	*temperature*	*humidity*	*wind*	*play-tennis?*
d1	sunny	hot	high	weak	no
d2	sunny	hot	high	strong	no
d3	cloudy	hot	high	weak	yes
d4	rain	warm	high	weak	yes
d5	rain	cold	normal	weak	yes
d6	rain	cold	normal	strong	no
d7	cloudy	cold	normal	strong	yes
d8	sunny	warm	high	weak	no
d9	sunny	cold	normal	weak	yes
d10	rain	warm	normal	weak	yes
d11	sunny	warm	normal	strong	yes
d12	cloudy	warm	high	strong	yes
d13	cloudy	hot	normal	weak	yes
d14	rain	warm	high	strong	no



Classification

- Given a pattern X, to obtain its class, regardless the learning set. For each no terminal node, to answer the question according to the attributes of X.



Automatic construction: ID3

- It is based on reducing the entropy
- It obtains good trees, although they are not optimal
- It is a greedy algorithm
- It requires the number of patterns to be highly superior to the number of classes.
- Applications: discrete attributes and finite set of classes

Entropy

- Def: "*Magnitude* that measures the information in a dataflow, i. e., how much new information is given." Wikipedia (Spanish) March 3, 2006.
- Def2: "It is a measure of the randomness of elements in a system".
- Example1: "Suppose a family with 4 members. It is three o'clock and all of them have lunch. All them decide to see the news on TVE". In this example, the entropy on the TV preferences is 0. The reason is that there is no diversity, every member prefers the same.
  - If three members prefer a channel and the another one, other channel, the entropy increases.
  - If two members prefer a channel and the another two, other channel, the entropy increases.
  - If every member prefer a different channel, the entropy is maximal.

ID3: Execution

- It starts from a set of patterns
- While there are sets with patterns of different classes
  - For each attribute, yet chosen, classify the patterns of the set.
  - Compute the profit of classifying by means of each attribute (it is based on the entropy).
  - Choose the attribute that produces the highest profit.
  - Repeat the procedure for every subset.

Formulae

- Probability for an attribute *K* to be set to a value *v* in the set *S*

$$P(K=v,S)=\frac{|S_{K=v}|}{|S|} \quad P(play=yes,Total)=\frac{9}{14} \quad P(play=no,Total)=\frac{5}{14}$$

- Entropy of the set *S*

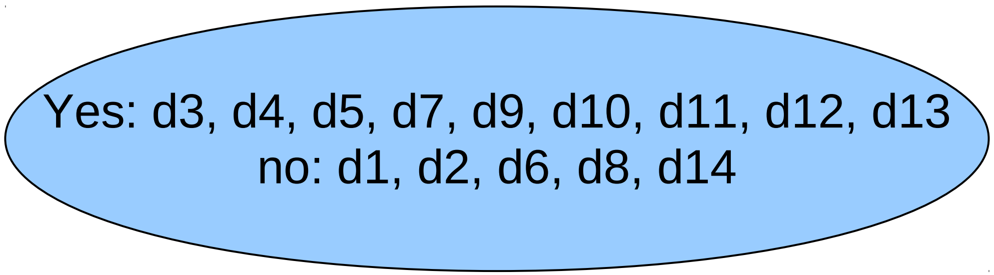
$$E\left(P(res=v_{1,}S),\ldots,P(res=v_n,S)\right)=\sum_{i=1}^{n}-P(res=v_i,S)\cdot\log_2\left(P(res=v_i,S)\right)$$

$$E\left(\frac{9}{14},\frac{5}{14}\right)=\frac{-9}{14}\cdot\log_2\left(\frac{9}{14}\right)+\frac{-5}{14}\cdot\log_2\left(\frac{5}{14}\right)=0{,}94$$

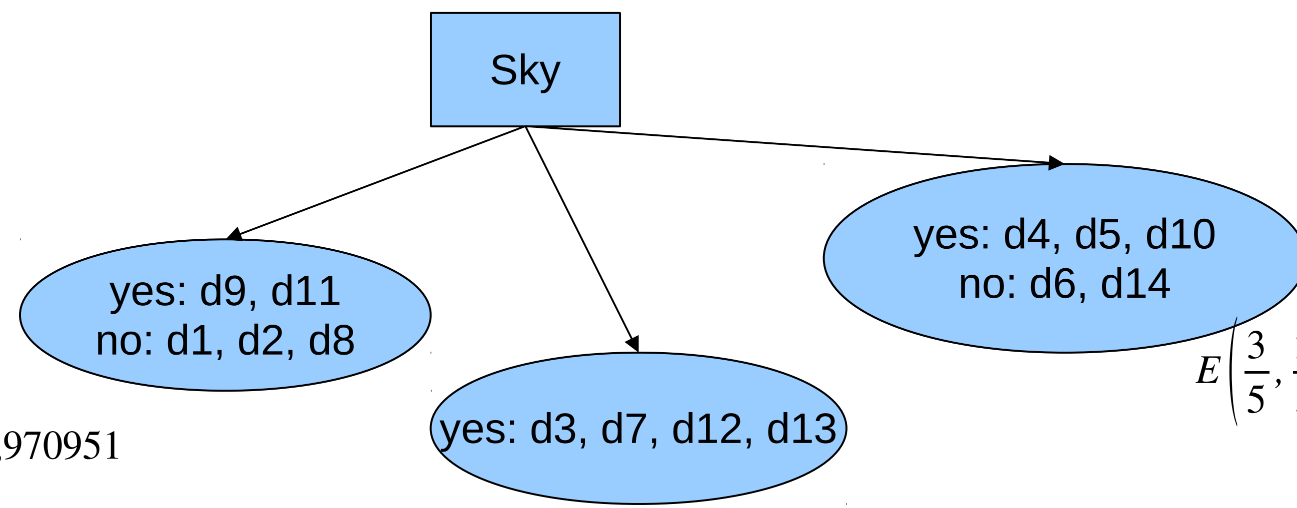
- *Profit* of classifying by means of the attribute K

$$Profit(S,K)=E(S)-\sum_{value\in K}P(K=value,S)\cdot E(S_v)$$

Example



E(S) = 0.94



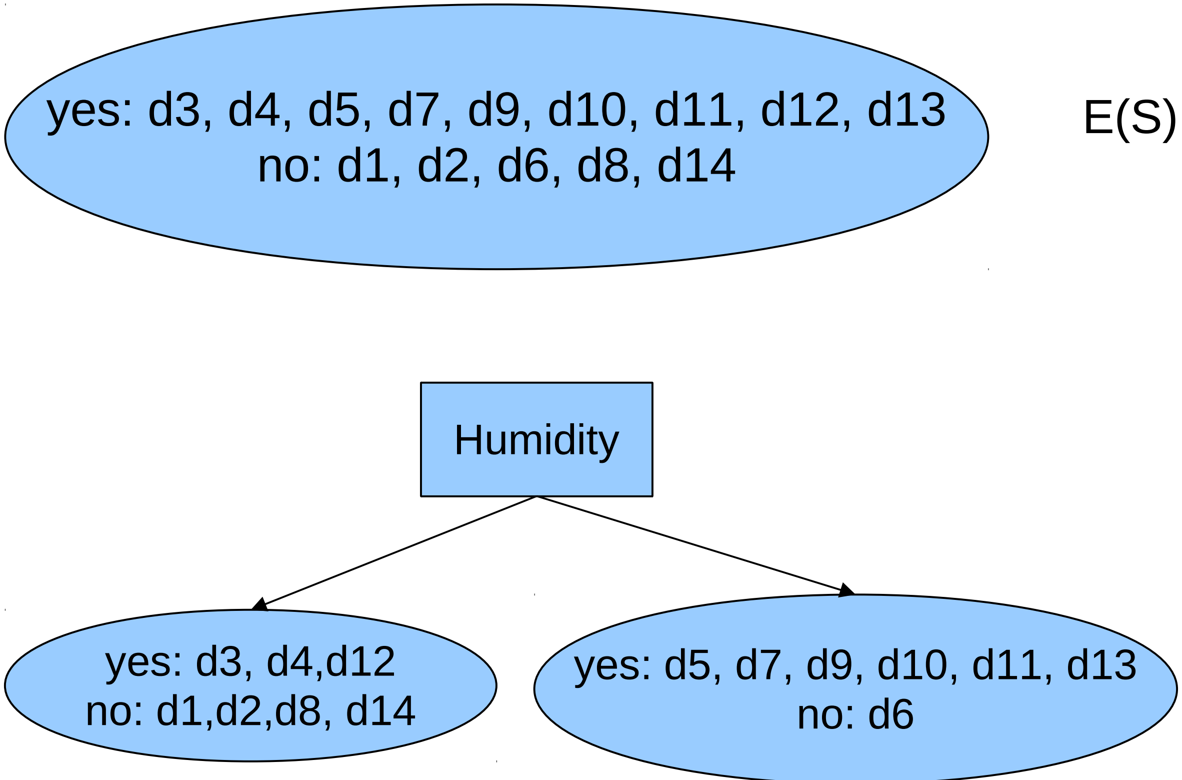
$$E\left(\frac{2}{5},\frac{3}{5}\right)=0{,}970951$$

$$E\left(\frac{4}{4},\frac{0}{4}\right)=0$$

$$E\left(\frac{3}{5},\frac{2}{5}\right)=0{,}970951$$

$$Profit\left(Sky,S\right)=E\left(S\right)-P\left(C=sunny,S\right)\cdot E\left(S_{C=sunny}\right)-P\left(C=cloudy,S\right)\cdot E\left(S_{C=cloudy}\right)-$$
$$-P\left(C=rain,S\right)\cdot E\left(S_{C=rain}\right)=0.94-\frac{5}{14}\cdot 0{,}970951-\frac{4}{14}\cdot 0-\frac{5}{14}\cdot 0{,}970951=0{,}24675$$

Ejemplo(2)



E(S) = 0.94

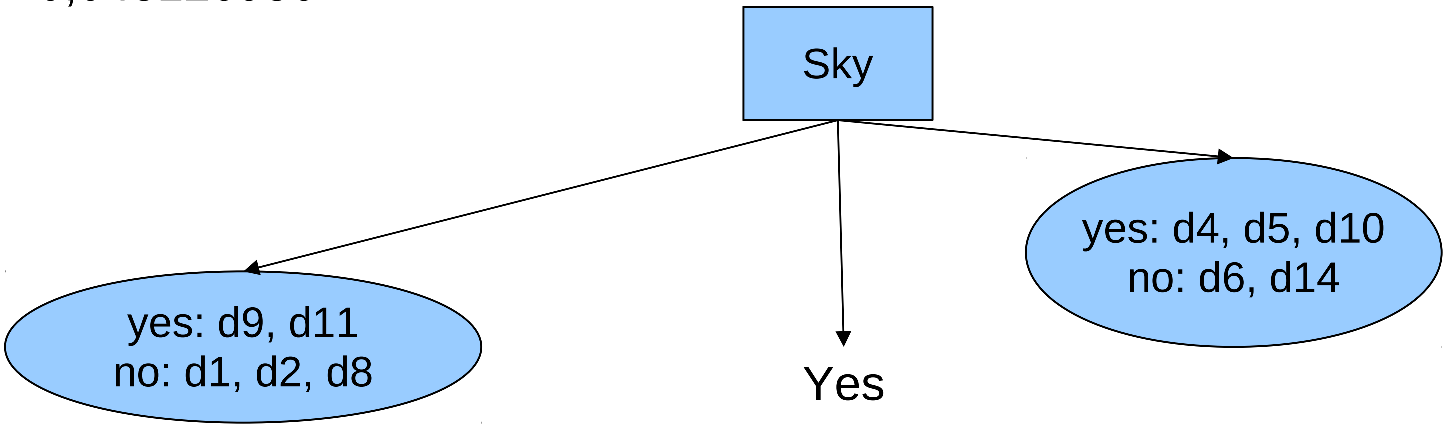
$$E\left(\frac{6}{7},\frac{1}{7}\right)=0{,}591673$$

$$E\left(\frac{3}{7},\frac{4}{7}\right)=0{,}985228$$

$$Profit\left(Humidity,S\right)=E\left(S\right)-P\left(H=high,S\right)\cdot E\left(S_{H=high}\right)-P\left(H=normal,S\right)\cdot E\left(S_{H=normal}\right)=$$
$$0.94-\frac{7}{14}\cdot 0{,}985228-\frac{7}{14}\cdot 0{,}591673=0{,}15183544$$

Choose attribute and repeat

Profit(Sky,S) = 0,24674976
Profit(Humidity,S) = 0,15183544
Profit(Temperature,S) = 0,029222548
Profit(Wind,S) = 0,048126936

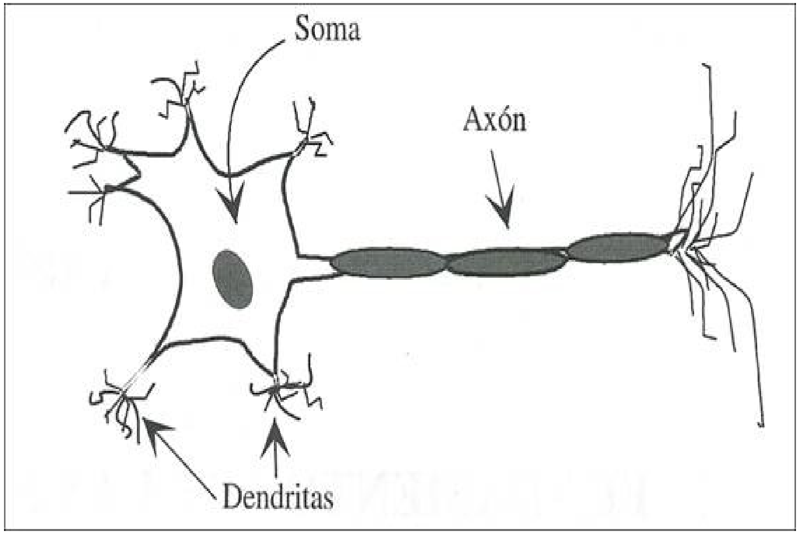


Profit(Temperature,$S_{c=sunny}$) = 0,5709506
Profit(Humidity,$S_{c=sunny}$) = 0,9709506
Profit(Wind,$S_{c=sunny}$) = 0,01997304

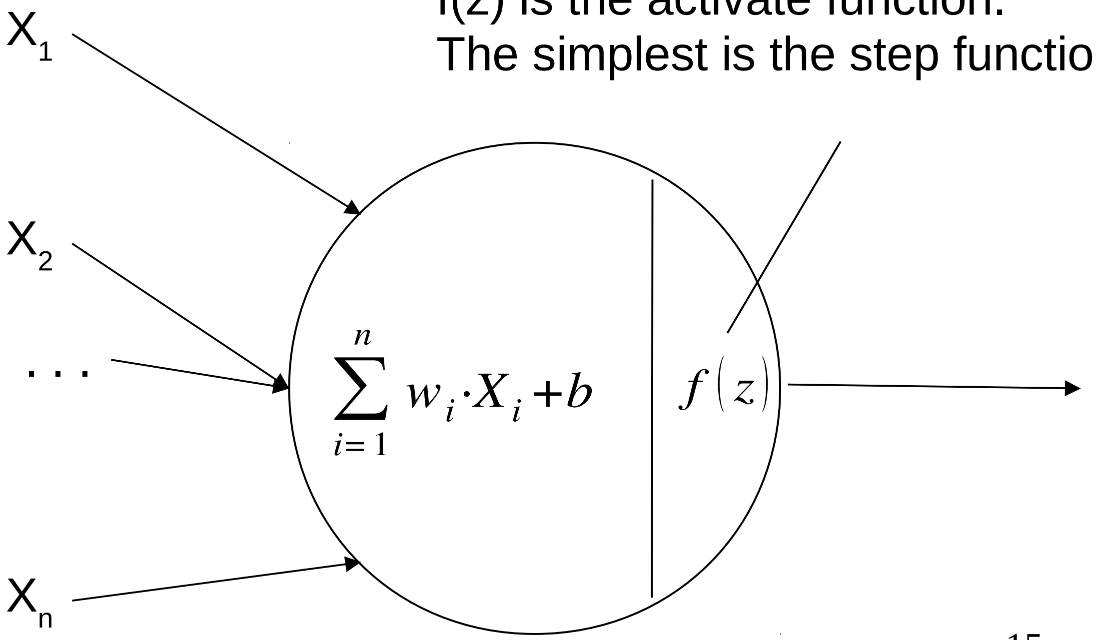
The Perceptron and the neural networks

Introduction

- They are inspired in the neural connections of our brain. In particular, in the multipolar neurons.

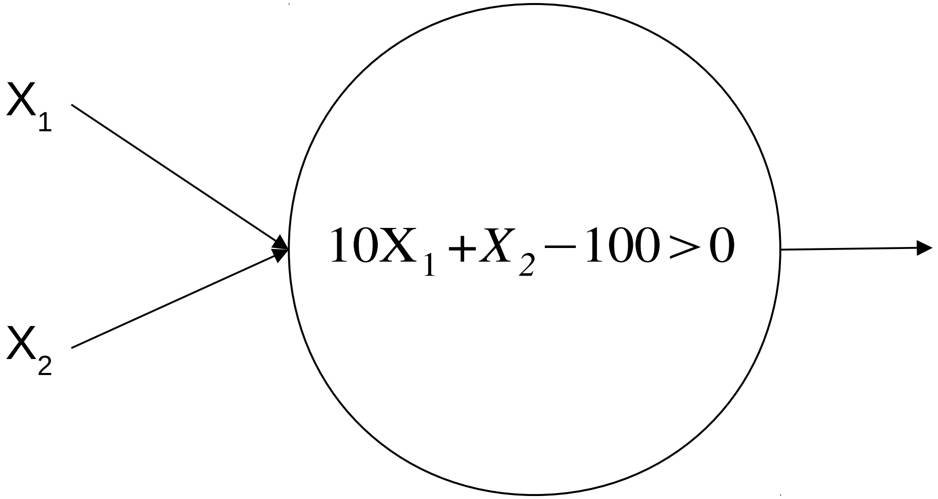


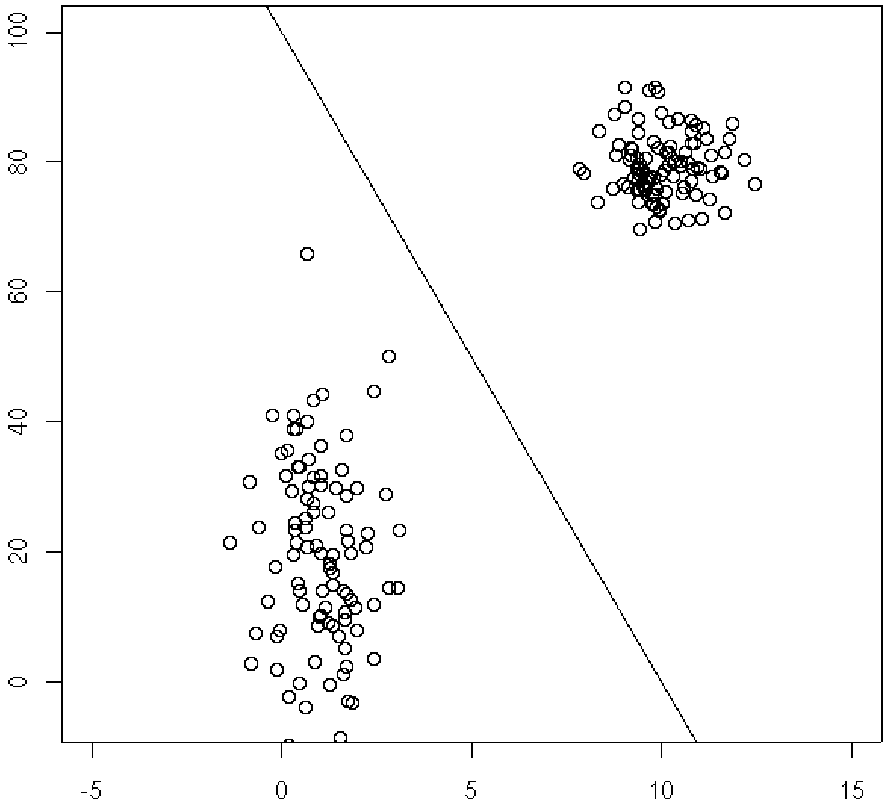
f(z) is the activate function.
The simplest is the step function.



Utility

- It is only able to solve hyperplane functions (they perform a simple cut)





Training: Delta rule

- The data is presented to the perceptron and the weights are updated if the output is not right:
  - α is the learning rate, and it should be small

$$w_i = w_i + \alpha \cdot \left(Y_{expected} - Y_{obtained} \right) \cdot X_i$$

$$b = b + \alpha \left(Y_{expected} - Y_{obtained} \right)$$

Exercise

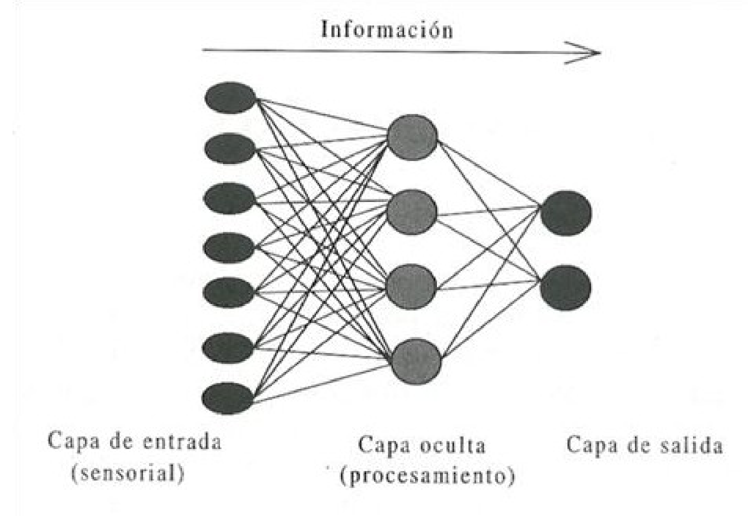
patrón	*x1*	*x2*	*y*
1	2	1	0
2	1	1	0
3	3	2	1
4	3	3	1

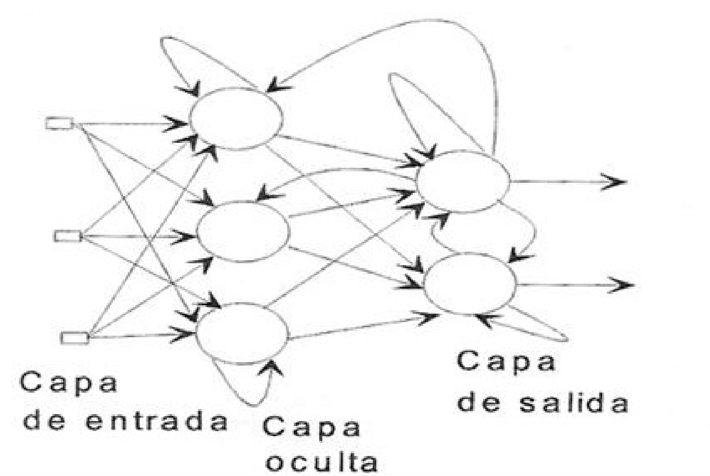
$$w_1 = 0, \quad w_2 = 0, \quad b = 0, \quad f(z) = \begin{cases} 1, & z > 0 \\ 0, & z \leq 0 \end{cases}$$

$$\alpha = 0{,}01$$

Comments

- Neural networks presents several layers to tackle complex problems:
  - to obtain the past tense of English verbs
  - to predict a temporal series.
  - artificial vision
- Feedforward networks are trained by means of an algorithm derived from the delta rule*: backpropagation.*
- *Recurrent networks are trained by means of evolutionary algorithms.*





Unsupervised learning

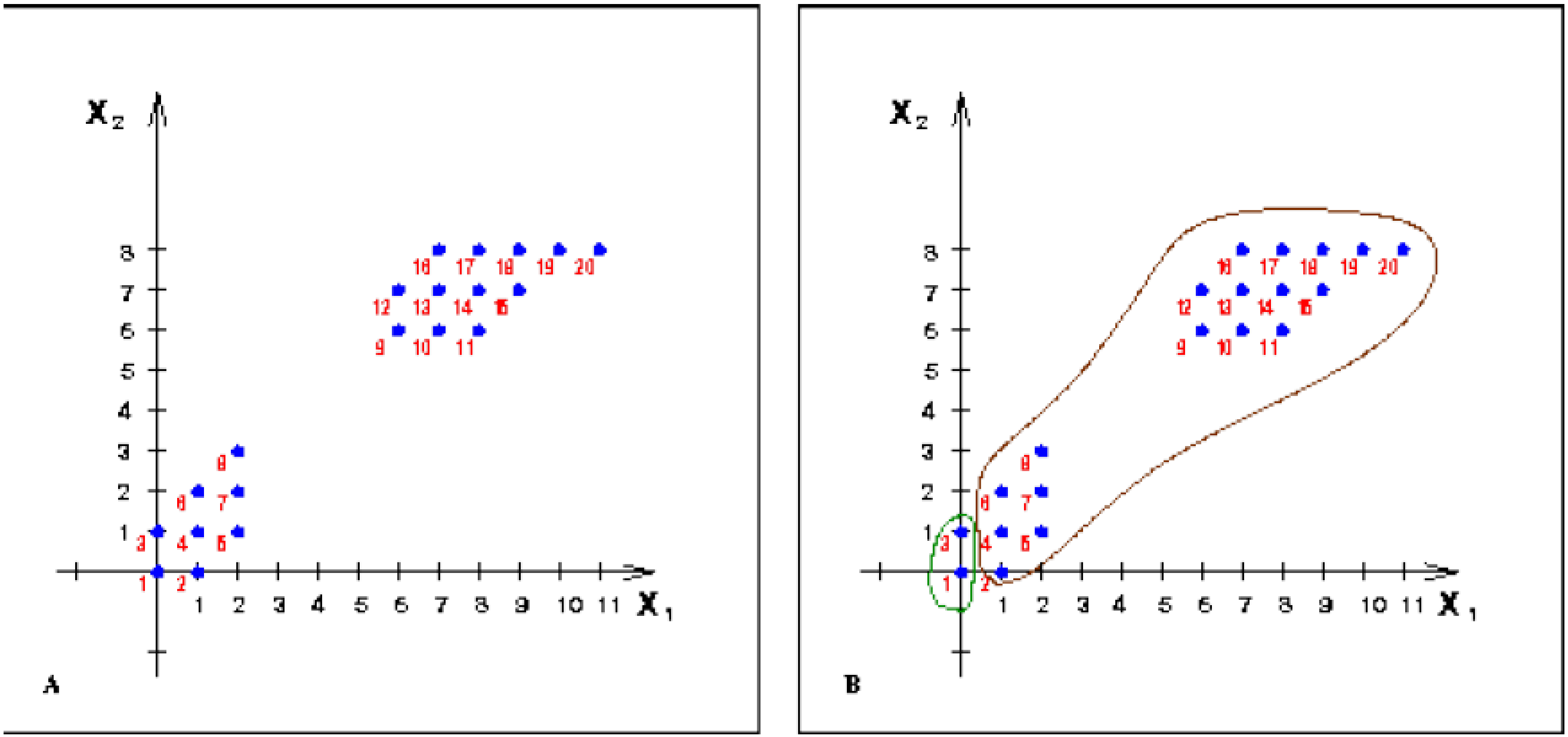
Introduction

- Unsupervised learning: to classify a data set without labels.
- *Clustering* techniques identify groups of data.
- Each group should be homogeneous and different to the other groups.
  - Distance: the distance between elements of the same group should be small, and large between elements of different groups.
- The clustering algorithms require parameters that affect the results.
- Usually, each group is characterized by means of its mean and variance.

KMeans

1. **Initialize** randomly the centres of *K* groups (parameter)
2. ***Assignation and update of the centres***.
   - Each pattern is assigned to its nearest centre.
   - Compute the new centres according to the mean of the assigned patterns
3. **Repeat step 2 until convergence.**
   - **When updating the centres, some patterns change to another nearest centre.**

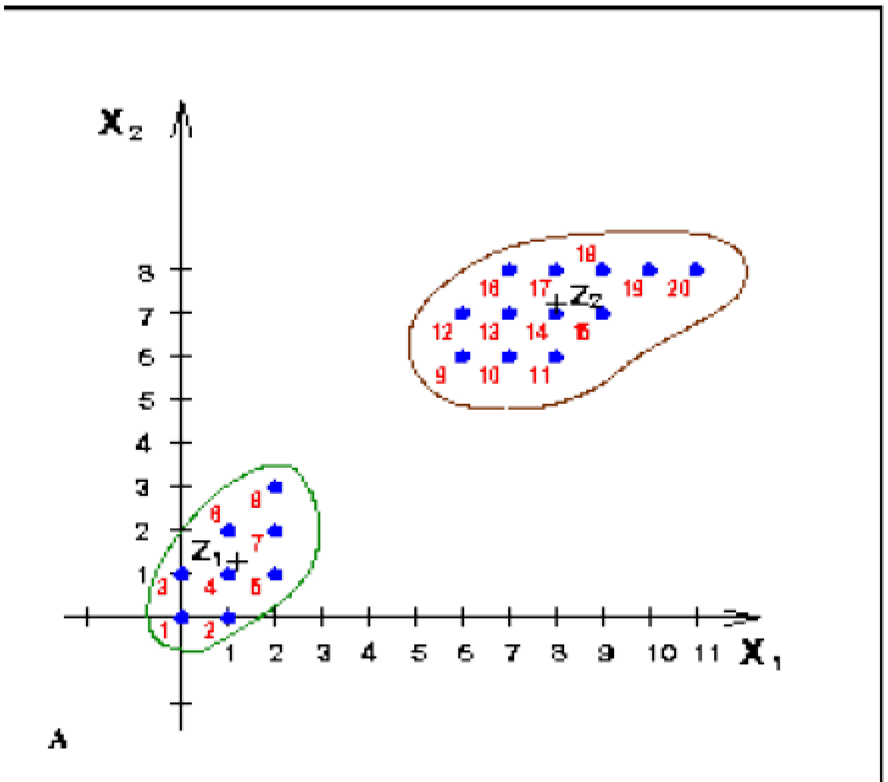
Example

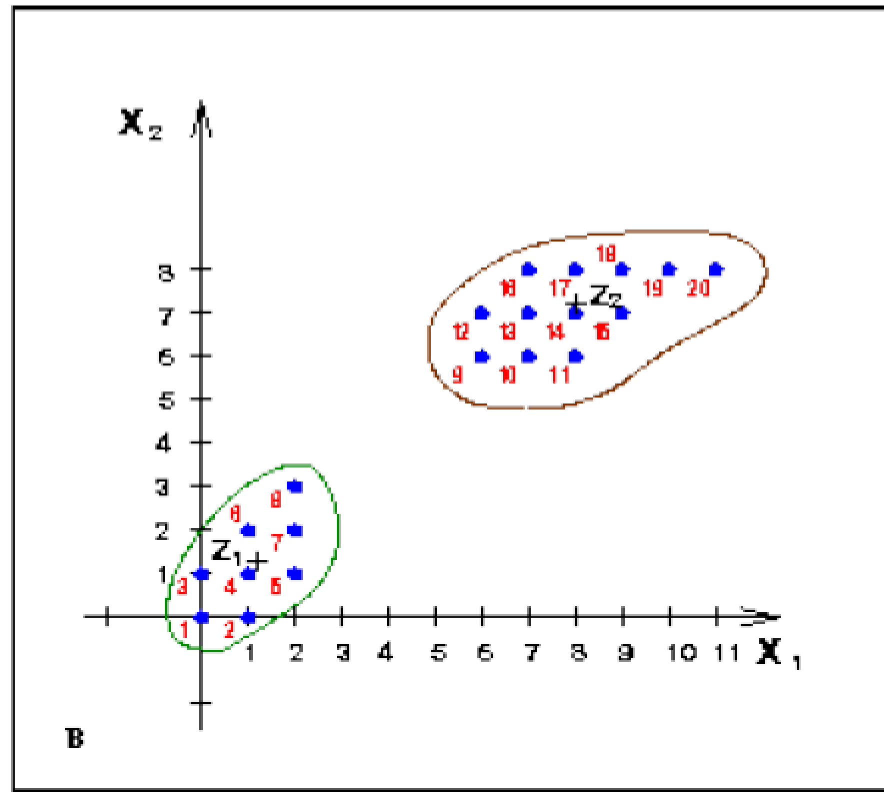


$S_1(0) = \{X_1\}$	$Z_1(0) = (0, 0)$
$S_2(0) = \{X_2\}$	$Z_2(0) = (1, 0)$

$S_1(1) = \{X_1, X_3\}$	$Z_1(1) = (0, 0.5)$
$S_2(1) = \{X_2, X_4, X_5, \ldots, X_{20}\}$	$Z_2(1) = (5.8, 5.3)$

Example (2)





$S_1(2) = \{X_1, X_2, \ldots, X_8\}$	$Z_1(2) = (1.1, 1.3)$
$S_2(2) = \{X_9, X_{10}, \ldots, X_{20}\}$	$Z_2(2) = (8.0, 7.2)$

$S_1(3) = \{X_1, X_2, \ldots, X_8\}$	$Z_1(3) = (1.1, 1.3)$
$S_2(3) = \{X_9, X_{10}, \ldots, X_{20}\}$	$Z_2(3) = (8.0, 7.2)$