Machine learning: classification

Roberto Innocente
inno@sissa.it

Terminology

- Another way to call rows and columns of spreadsheets :
  - Columns = Attributes : Categorical, Numerical
  - Rows = Instances, Records

Outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rain	mild	high	strong	no

Table 1: Play tennis table

Prediction

- From a ***training set*** of examples :
  - we want to learn to predict a class of the instances : ***classification***
  - we want to learn to predict a numerical attribute : ***regression***

Classification problem

- We are given a set of pairs(***training set***) :
  - x(i), y(i) : where x is a vector (an array) of multiple attributes (columns), and y has a finite set of values.
- Learn a function f: X->Y that fits in a good way the examples given
- The problem is that there are |Y|^|X| such functions, the training set is very small compared to the domain X, and there is uncertainty on the data

Naive Bayes

- We can apply Bayes theorem and for each function we can compute
  - p(f | d) = p(d | f)*p(h)/p(d) : where d is the data of training set
- Then according to the Maximum A Posteriori principle select the one wich maximizes p(f|d)
- Uncertainty in the data and unknown cross correlations between attributes make things very hard

Play tennis table

Outlook	temperature	humidity	wind	play
sunny	hot	high	weak	no
sunny	hot	high	strong	no
overcast	hot	high	weak	yes
rain	mild	high	weak	yes
rain	cool	normal	weak	yes
rain	cool	normal	strong	no
overcast	cool	normal	strong	yes
sunny	mild	high	weak	no
sunny	cool	normal	weak	yes
rain	mild	normal	weak	yes
sunny	mild	normal	strong	yes
overcast	mild	high	strong	yes
overcast	hot	normal	weak	yes
rain	mild	high	strong	no

Table 1: Play tennis table

- 5 attributes : 4+ 1 target attribute(play)
- Outlook: 3 outcomes, temperature: 3 outcomes, humidity: 2, wind: 2
- |Domain|:3x3x2x2=36
- |functions| = subsets of domain = 2^36 ~10^12
- |training set| = 14

Occam's razor

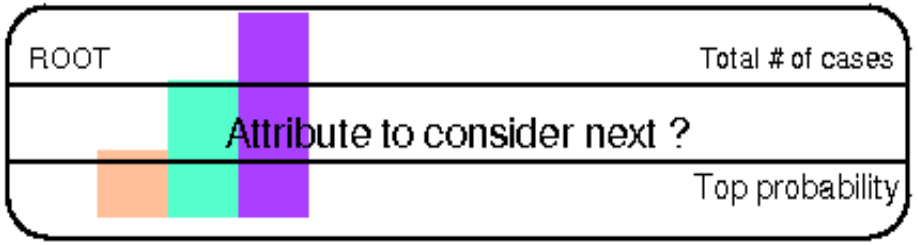
- **lex parsimoniae** : "entia non sunt multiplicanda praeter necessitatem" or "entities should not be multiplied beyond necessity"
- The simplest hypotheses that fit are probably the right ones

Induction learning

- We build up knowledge growing a knowledge base, in this way we try to obey to Occam's law :
  - **Rule induction :** we start with simple propositions, and we add in Normal Conjunctive Form till we drop all negative examples
  - **Tree induction :** we analyze level after level different attributes that reduce ***impurity*** of classification
- They reduce one to other : every node of a tree can be seen as the disjunction of all previous branching values, and every rule can be seen as a leaf node of a tree

Node types

ROOT node:

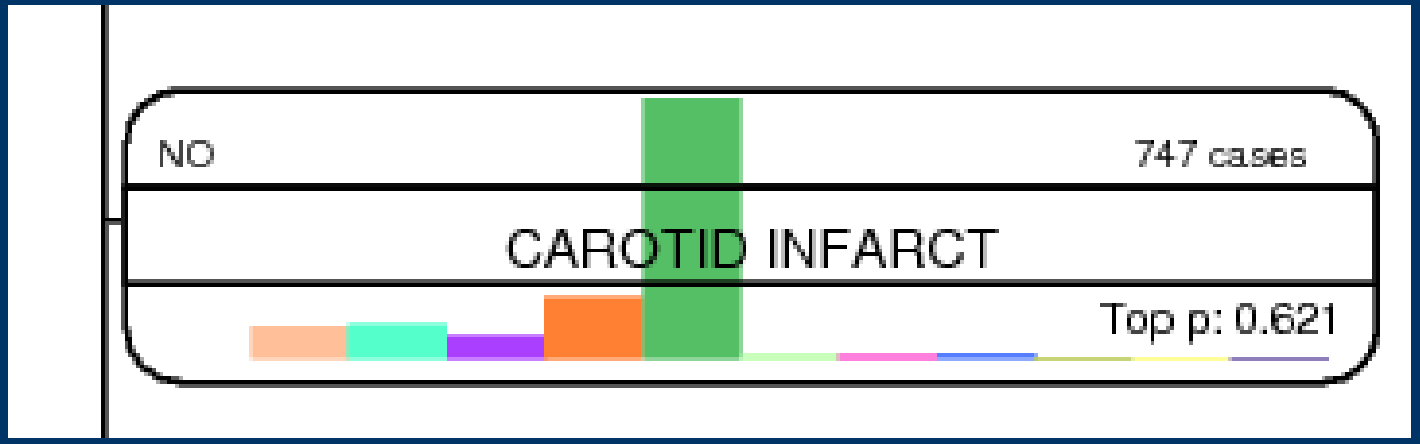


Internal node:

Leaf node:

Bar chart

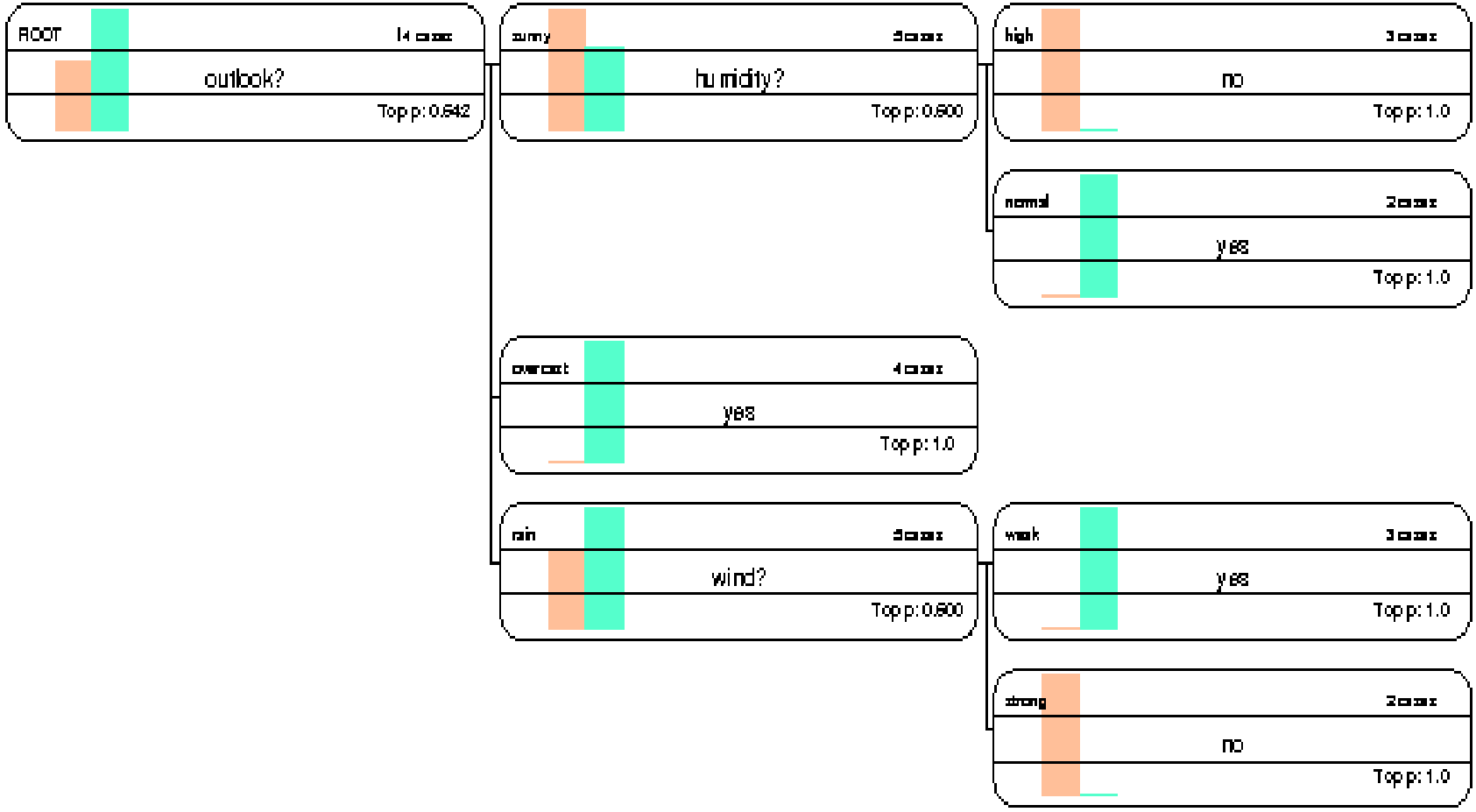
- 3 bands : top probability covers all of them
- 2/3 of top probability lower 2 bands
- 1/3 of top probability lower band
- In this example: the green bar is 62 %, and hence the orange bar is ~18 % (1/3*62%)



Impure nodes / Misprediction

- A node having instances with multiple classifications is also called ***impure***
- A rational guess of the classification at that point would be the one with the top probability
- The probability of doing a mistake predicting the most probable outcome is called ***misprediction rate*** and is (1-Top probability)
- For the previous example the misprediction rate is : 1-0.621= 0.379 ~ 37%

Play tennis tree



Rule/node equivalence

- Nodes as rules and viceversa :
  - outlook=sunny and humidity=high=>play=no
  - outlook=sunny and humidity=low=>play=yes
  - outlook=overcast => play=yes
  - outlook=rain and wind=weak => play=yes
  - outlook=rain and wind=strong => play=no
- We can also use probabilistic rules and impure nodes:
  - outlook=rain => play=yes (with prob 0.6)

Simplified diagnostic tree

Stroke data

- More than 100 attributes (columns)
- 11 possible outcomes
- Counting as if all attributes were binary:
  - 11^(2^100) ~ 11^(10^33)
- By contrast we have a training set of only around 1000 instances (rows)