Induction on Decision Trees

Séance « IDT »

de l'UE « apprentissage automatique »

Bruno Bouzy

bruno.bouzy@parisdescartes.fr

www.mi.parisdescartes.fr/~bouzy

Outline

- Induction task
- ID3
- Entropy (disorder) minimization
- Noise
- Unknown attribute values
- Selection criterion

The induction task

- Formalism:
  - objects with attributes
- Example:
  - objects = saturday mornings
  - attributes:
    - outlook {sunny, overcast, rain}
    - temperature {cool, mild, hot}
    - humidity {high, normal}
    - windy {true, false}

The induction task

- One particular saturday:
  - Outlook = overcast
  - Temperature = cool
  - Humidity = normal
  - Windy = false
- Classes mutually exclusive, here 2 classes:
  - Positive (P)
  - Negative (N)

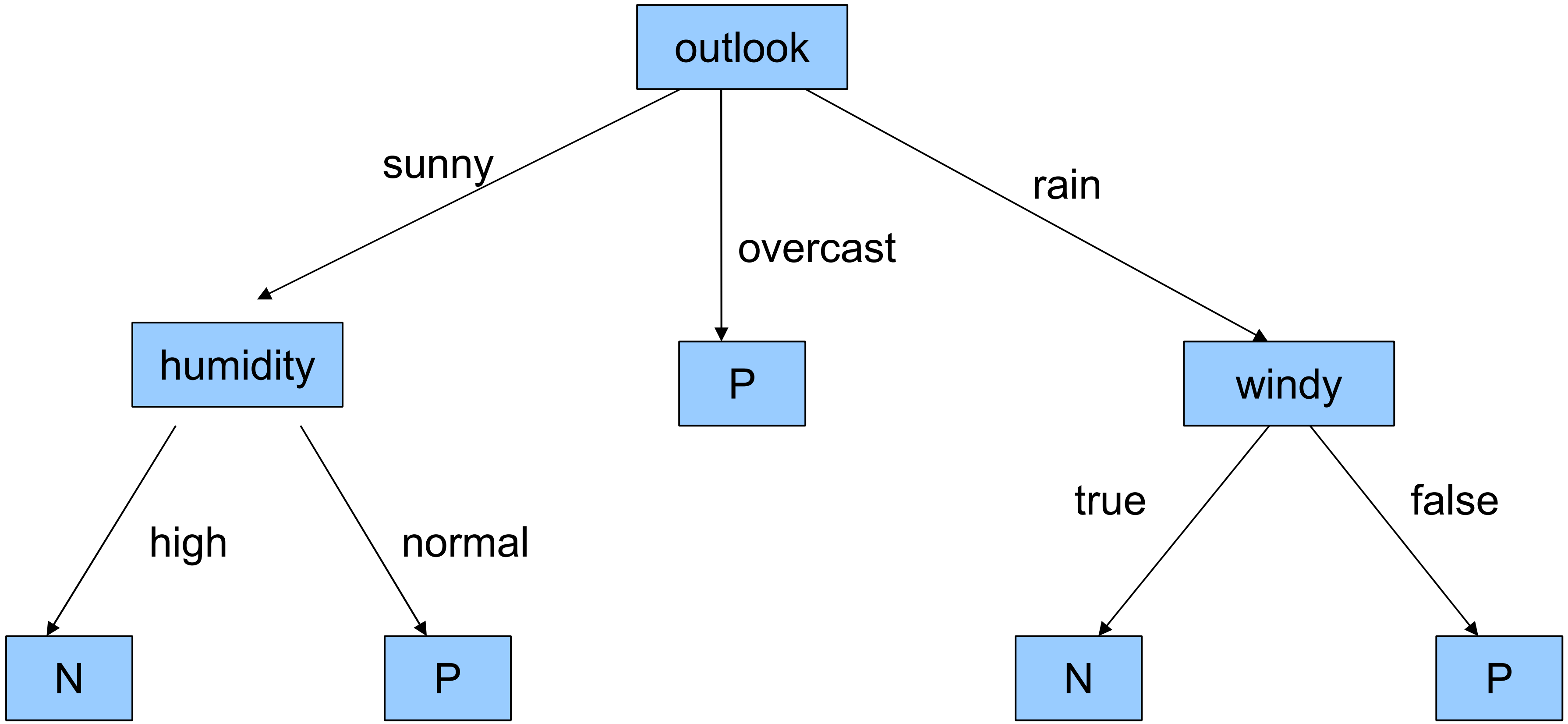
The induction task

- Training set:
  - objects whose class is known
- Goal:
  - Develop a classification rule

A small training set

n	outlook	temperat.	humidity	windy	C
1	sunny	hot	high	false	N
2	sunny	hot	high	true	N
3	overcast	hot	high	false	P
4	rain	mild	high	false	P
5	rain	cool	normal	false	P
6	rain	cool	normal	true	N
7	overcast	cool	normal	true	P
8	sunny	mild	high	false	N
9	sunny	cool	normal	false	P
10	rain	mild	normal	false	P
11	sunny	mild	normal	true	P
12	overcast	mild	high	true	P
13	overcast	hot	normal	false	P
14	rain	mild	high	true	N

A simple decision tree



The induction task

- If the attributes are adequate, it is possible to build a correct decision tree.
- Many correct decision trees are possible.
- Correctly classify unseen objects ? (it depends...)
- Between 2 correct decision trees, choose the simplest one.

ID3

- Systematical approach:
  - Generate all decision trees and choose the simplest
  - Possible for small induction tasks only

- ID3 approach:
  - Many objects, many attributes.
  - A reasonably good decision tree is required.
  - Use the entropy minimization principle to select the « best » attribute

ID3

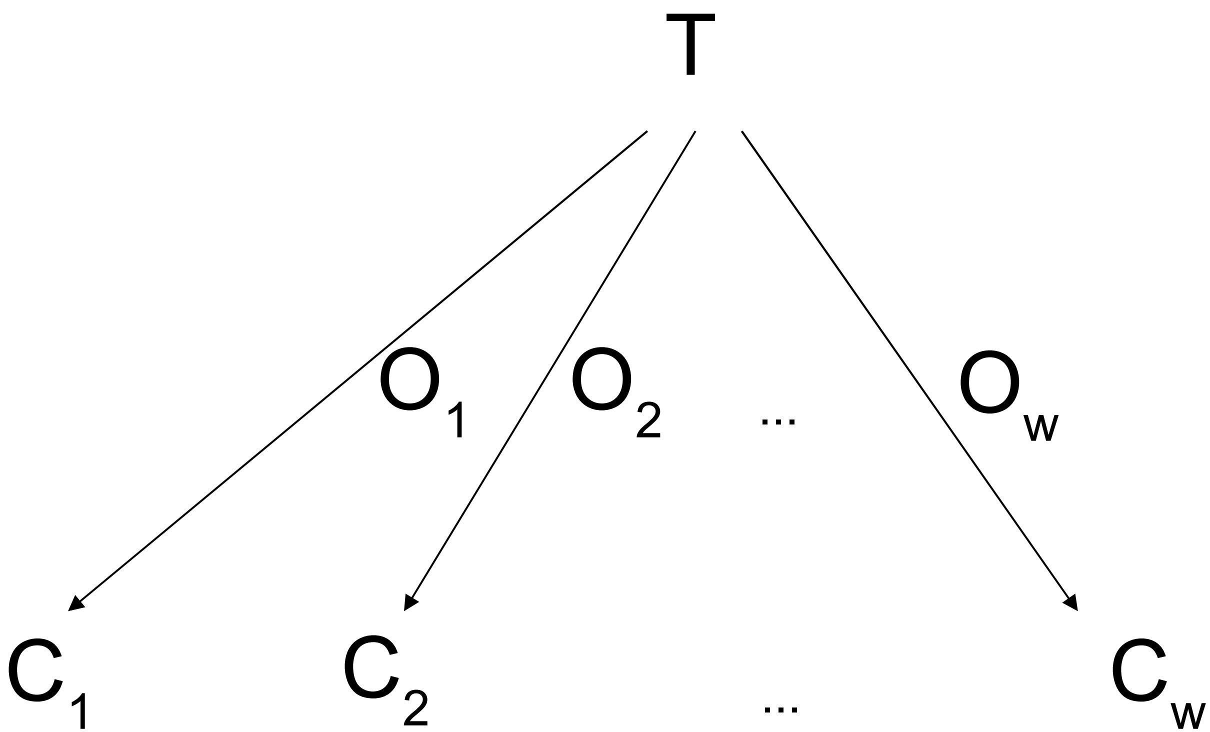
- Result:
  - Correct decision trees are found.
  - Training sets of 30,000 examples
  - Examples with 50 atttributes
  - No convergence garantee

ID3

- How to form a DT for a set C of objects ?
  - T = test of the value of a given attribute on an object
  - The possible values (outcomes) are:

    $O_1, O_2, ..., O_w$.
  - Partition = $\{C_1, C_2, ..., C_w\}$ of C.
  - C_i contains objects of C whose value (outcome) is O_i.

A structuring tree of C



Choice of the test

- 2 assumptions:

(1) the test set is in the proportion of the training set:

p: number of positive (+) examples

n: number of negative (-) examples

P_+: probability to be positive = p/(p+n)

P_-: probability to be negative = n/(p+n)

(2) Information gain based on the entropy E(p, n):

$$E(p, n) = - P_+ \log(P_+) - P_- \log(P_-)$$

(entropy ≈ disorder)

Choice of the test

- A attribute with values in $\{A_1, A_2, ..., A_w\}$
- $C = \{C_1, C_2, ..., C_w\}$
  - objects in C_i have $A = A_i$.
- C_i has p_i objects in P and n_i objects in N.
- $E(p_i, n_i)$ = entropy of of C_i.

Entropy function

A measure of disorder

For x in $]0, 1[$: $E(x) = -x\log(x) - (1-x)\log(1-x)$

- $E(0) = E(1) = 0$
  - No disorder
- E is a bell function
  - maximum for $x=1/2$ (maximal disorder)
  - Vertical in 0 and 1.
  - $E(1/2) = \log(2) \approx 0.7$
- (... approximate values: $\log(3) \approx 1.1$ $\log(4) \approx 1.4$ $\log(5) \approx 1.6$ $\log(7) \approx 2$)

Entropy function

- p positive objects and n negative objects...

- What is the entropy $E(p|n)$ of the proportion $(p|n)$?

- $E(p|n) = - p/(p+n)\log(p/(p+n)) - n/(p+n)\log(n/(p+n))$

  $= \log(p+n) - p/(p+n)\log(p) - n/(p+n)\log(n)$

Choice of the test

« Entropy a priori » (Eap) of attribute A:

A measure of what could be the average entropy if we ask the value of attribute A

A weighted sum of the entropies associated to each value of A

The weight of value Ai is in proportion of the number of objects with value Ai

$$Eap(A) = \Sigma_i \, E(p_i, n_i)(p_i+n_i)/(p+n)$$

Choose attribute $A^* = \text{argmin}_b \, Eap(b)$

(i.e. looking for the attribute that minimizes disorder...)

Choice of the test

- Example, the entropy « a priori » of each attribute
  - Eap(outlook) = **0.45**
  - Eap(temperature) = 0.65
  - Eap(humidity) = 0.55
  - Eap(windy) = 0.65

- ID3 chooses « outlook » as the DT root attribute.

ID3

- Complexity:
  - O (|C|.|A|.D)
  - |C| : size of the training set
  - |A| : number of attributes
  - D : depth of the decision tree

Noise

- Error in attribute values
  - Object 1 . outlook = overcast
  - 1 and 3 identical, but belong to different classes.

- Misclassification:
  - Object 3 corrupted to belong to N
  - The DT becomes complex (12 nodes)

Noise

- Two requirements:
  - (R1) Being able to work with inadequate attributes
  - (R2) Being able to decide that testing further attributes will not improve the predictive accuracy of the DT.

Noise

- What to do when an attribute is inadequate or irrelevant ?
- Create a leaf with which kind of value ?
  - Most numerous class: P or N
  - Probability of belonging to P

Unknown attribute values

- 2 questions:
  - How to build the DT ?
  - How to deal them during classification ?

Unknown attribute values

- How to build the DT ?

  - Bayesian approach -
  - DT approach -
  - « most common value » approach -
  - « unknown » as a value - -
  - the « proportion » approach ++

Unknown attribute values

Assume the value of A is unkown for few objects (= '?')

p_u number of objects in P with A unknown

n_u number of objects in N with A unknown

- Objects with unknown values are distributed across the values of in proportion the relative frequency of these values in C
- $p_i := p_i + p_u r_i$ where $r_i = (p_i+n_i)/((p+n)-(p_u - n_u))$
- (number of objects with value Ai: p_i+n_i)
- (Number of objects with A value known: $(p+n)-(p_u - n_u)$)

Summary

- Induction task = find out DT for classification
- 2 classes, ~1000 attributes, ~50 values
- Choice of root test based on information theory
- Minimization of entropy
- Noise
- Unknown attribute values
- Approximate method

Reference

- J.R. Quinlan, « Induction on decision trees », Machine Learning (1986)