Decision Trees

Isabelle Bloch

LIP6, Sorbonne Université - LTCI, Télécom Paris

isabelle.bloch@sorbonne-universite.fr, isabelle.bloch@telecom-paris.fr

Objective

- Learning set: description of each item by a list of attribute values, and assigned class.
- Find a model describing the links between attributes and classes.
- Find class assignments for new items based on their attributes. Should be interpretable by the users.

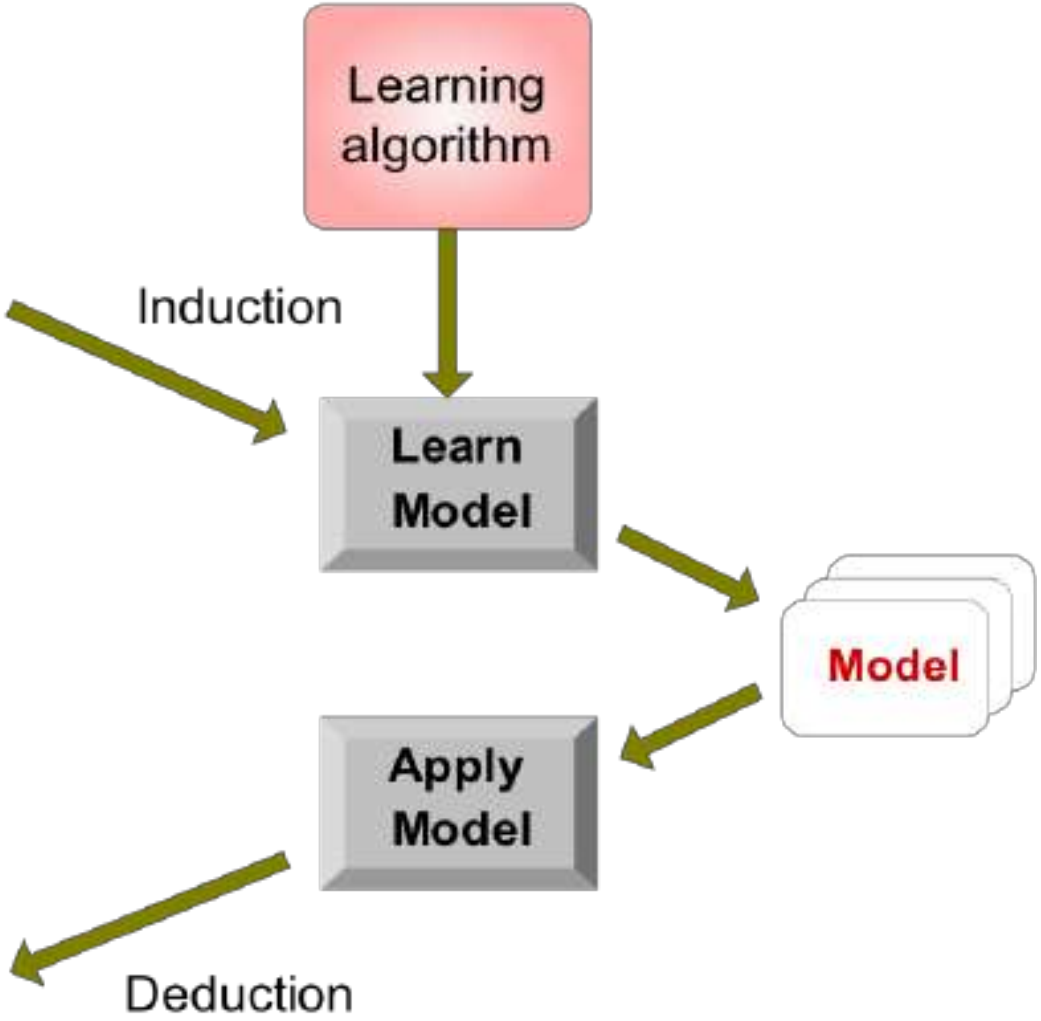
Example (Tan, Steinbach, Karpatne, Kumar, Minnesota University)

Tid	Attrib1	Attrib2	Attrib3	Class
1	Yes	Large	125K	No
2	No	Medium	100K	No
3	No	Small	70K	No
4	Yes	Medium	120K	No
5	No	Large	95K	Yes
6	No	Medium	60K	No
7	Yes	Large	220K	No
8	No	Small	85K	Yes
9	No	Medium	75K	No
10	No	Small	90K	Yes

Training Set

Tid	Attrib1	Attrib2	Attrib3	Class
11	No	Small	55K	?
12	Yes	Medium	80K	?
13	Yes	Large	110K	?
14	No	Small	95K	?
15	No	Large	67K	?

Test Set



Definition

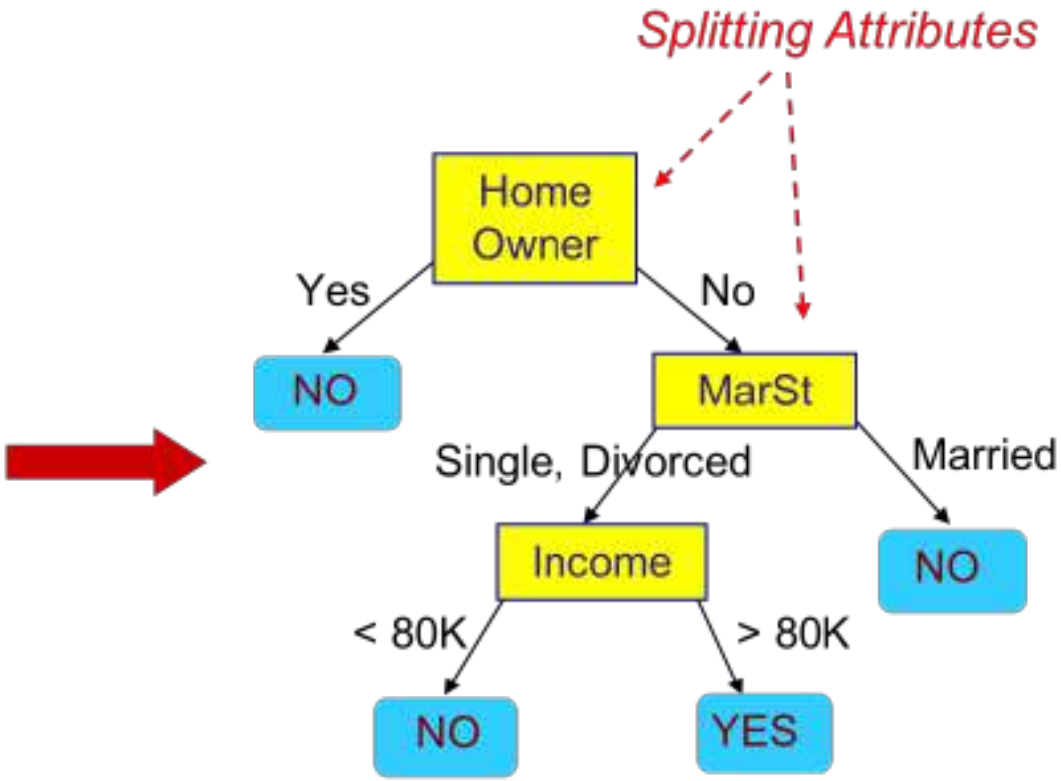
- Tree-like graph.
- Vertices: pick an attribute and ask a question.
- Edges: answers to the question.
- Leaves: actual output or class label.

Principle:

- Decision trees classify the examples by sorting them from the root to some leaf node.
- Each node in the tree acts as a test case for some attribute, and each edge descending from that node corresponds to one of the possible answers to the test case.
- Recursive process, for each sub-tree.

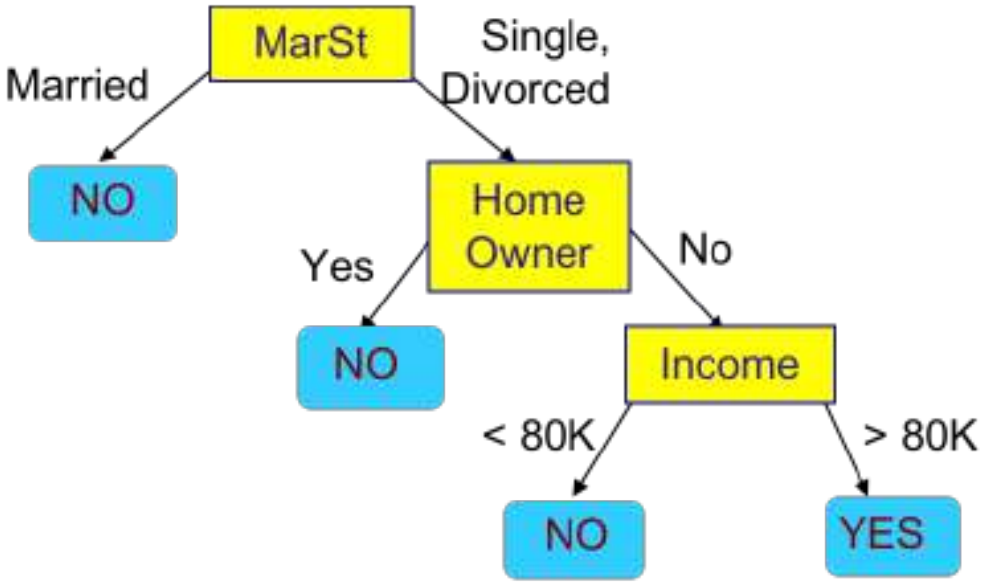
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Training Data



Model: Decision Tree

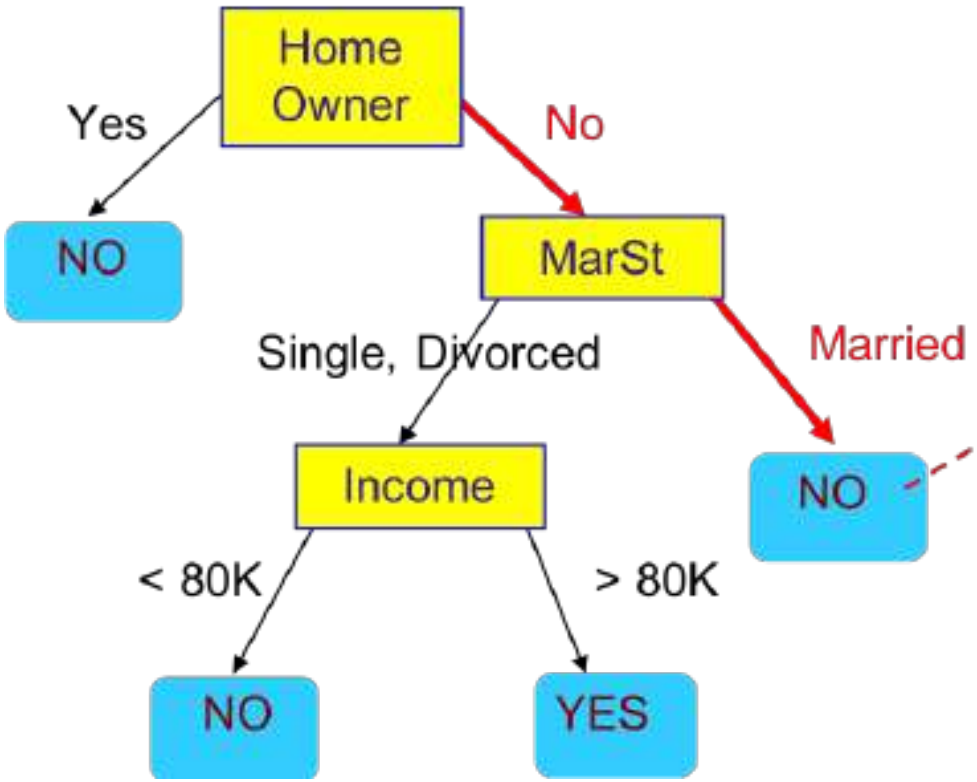
ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
	categorical	categorical	continuous	class
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



There could be more than one tree that fits the same data!

Test Data

Home Owner	Marital Status	Annual Income	Defaulted Borrower
No	Married	80K	?



Assign Defaulted to "No"

Tree construction (induction step)

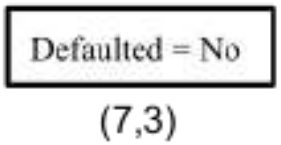
Many algorithms! with several problems to be solved:

- Splitting of the training set (binary vs multi-way).
- Test conditions: depend on the types of the attributes (nominal, ordinal, continuous, binary).
- Stopping criterion (depth of the tree): balance between precise classification and generalization capabilities.

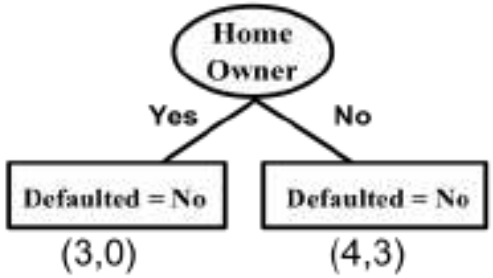
ID3 algorithm (Quinlan)

- Top-down, greedy approach.
- Select the best attribute A.
- Assign A as the decision attribute (test case) for the NODE.
- For each value of A, create a new descendant of the NODE.
- Sort the training examples to the appropriate descendant node leaf.
- If examples are perfectly classified, then STOP else iterate over the new leaf nodes.

How to choose the best attribute? $\Rightarrow$ notion of information gain (measure that expresses how well an attribute splits that data into groups based on classification).

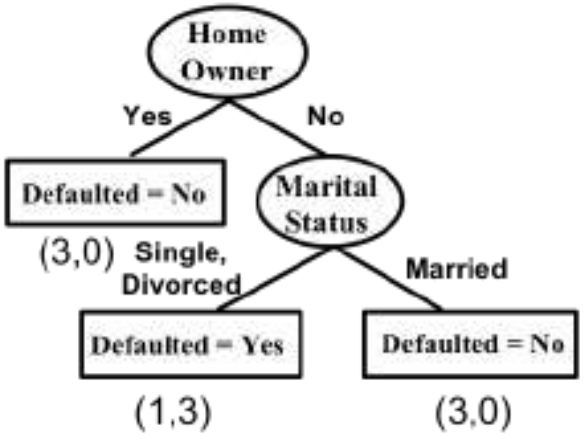


(a)

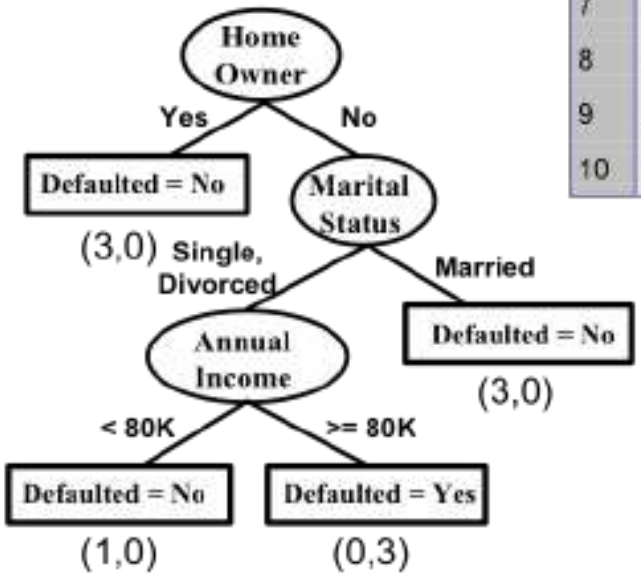


(b)

ID	Home Owner	Marital Status	Annual Income	Defaulted Borrower
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes



(c)

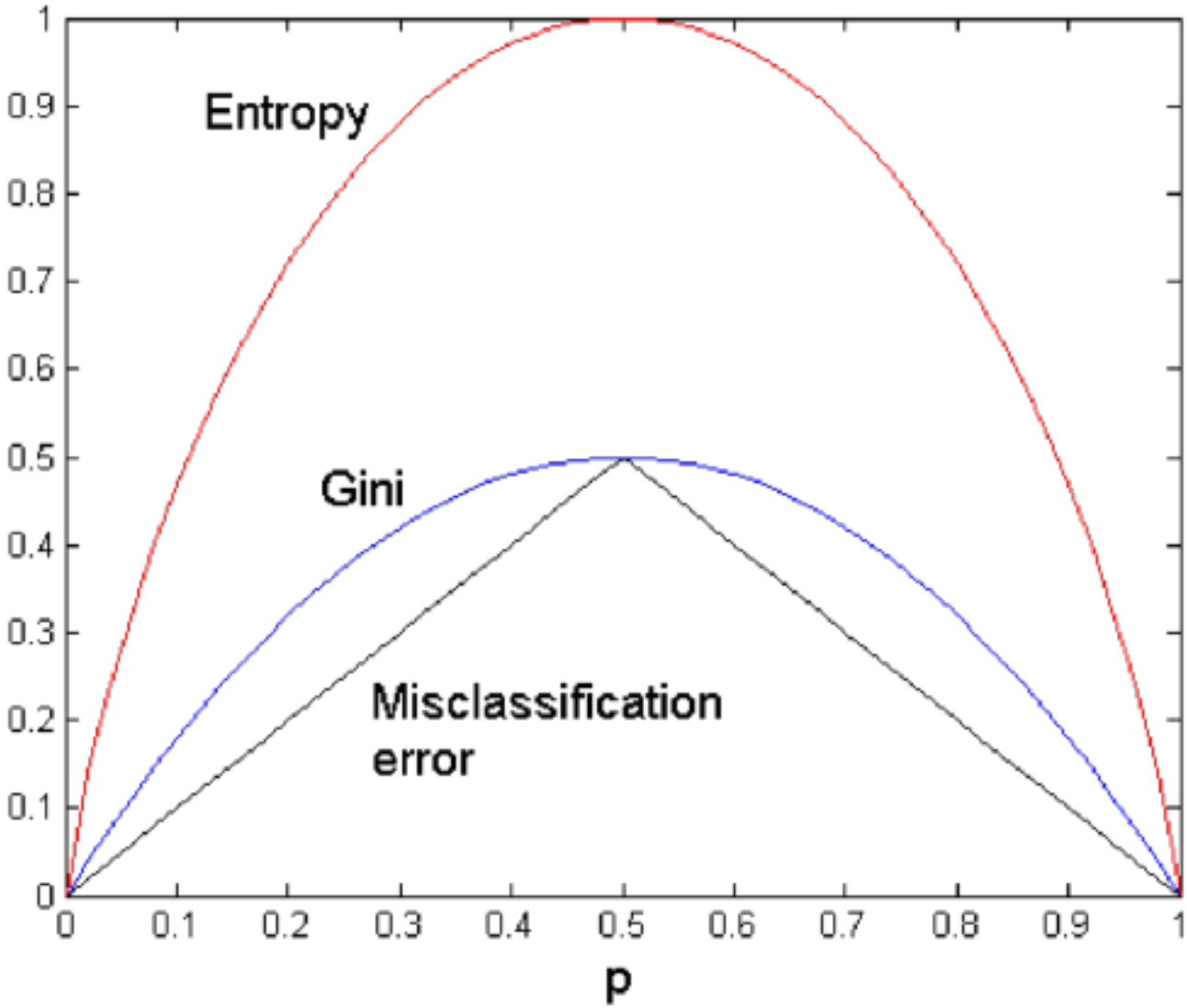


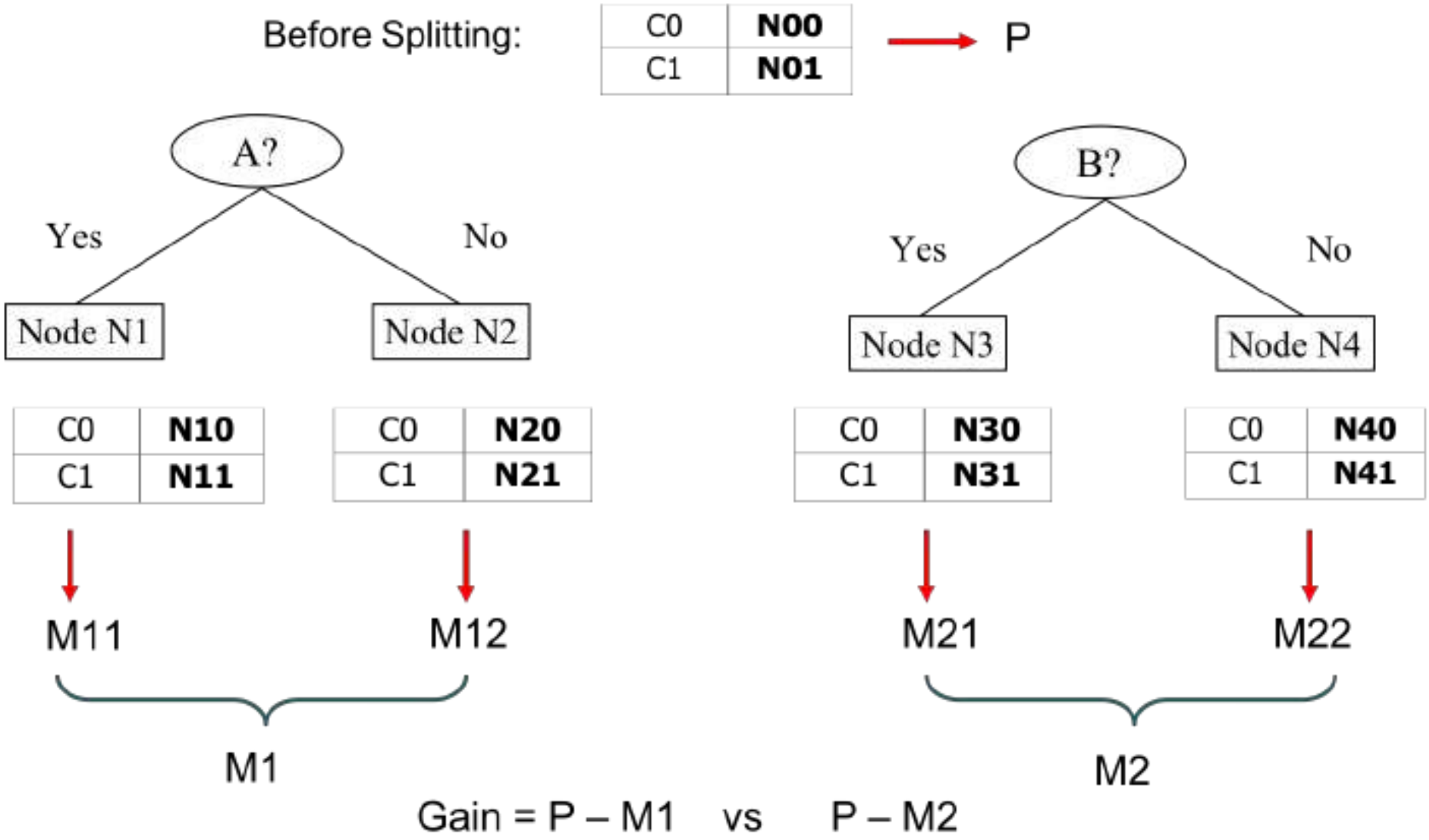
(d)

Split based on node "purity"

- Pure node: all associated items belong to the same class.
- Nodes with purer class distribution are preferred.
- Usual measures:
  - Gini index: for a node t, $GINI(t) = 1 - \sum_j p(j \mid t)^2$ where $p(j \mid t)$ is the relative frequency of class j at node t.
    $GINI(t) = 0$ if all items belong to a same class.
    Should be minimized.
  - Entropy: $-\sum_j p(j \mid t) \log p(j \mid t)$.
  - Classification error: $1 - \max_j p(j \mid t)$.
- Best split:
  - Compute purity before (P) and after (M) splitting.
  - Gain $= G = P - M$.
  - Choose the attribute test condition that produces the highest gain.
  - To avoid too small classes: normalize G by the partition entropy.

For a two-class problem:




Before Splitting:
C0 N00
C1 N01
P
A?
Yes
No
Node N1
Node N2
B?
Yes
No
Node N3
Node N4
C0 N10
C1 N11
C0 N20
C1 N21
C0 N30
C1 N31
C0 N40
C1 N41
M11
M12
M21
M22
M1
M2
Gain = P – M1 vs P – M2

Classification based on decision trees

Advantages:

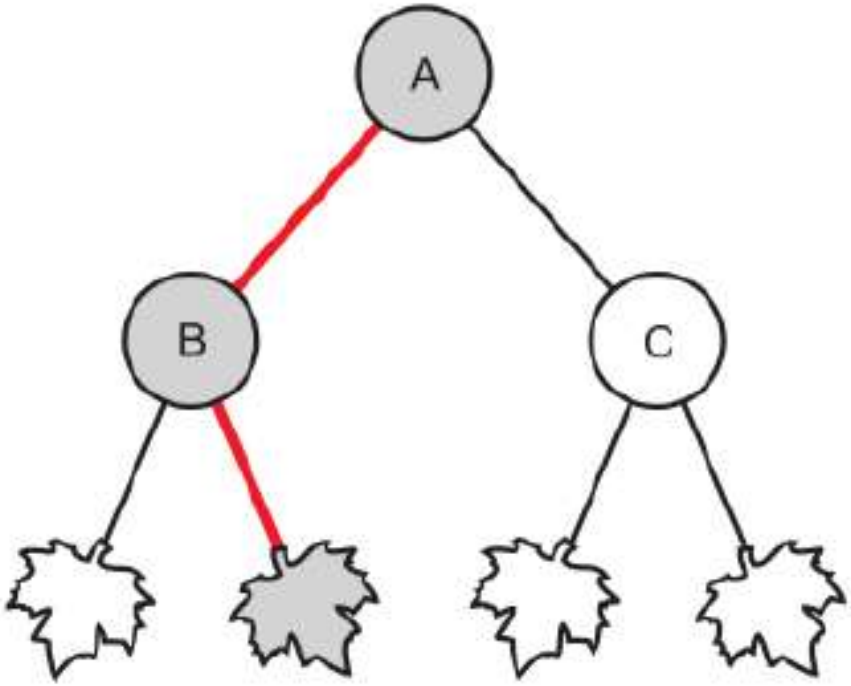
- Low construction cost.
- Very fast classification of unknown records.
- Easy interpretation (if trees are small enough).
- Robust to noise (especially when methods to avoid over-fitting are used).
- Can handle redundant or irrelevant attributes.

Disadvantages:

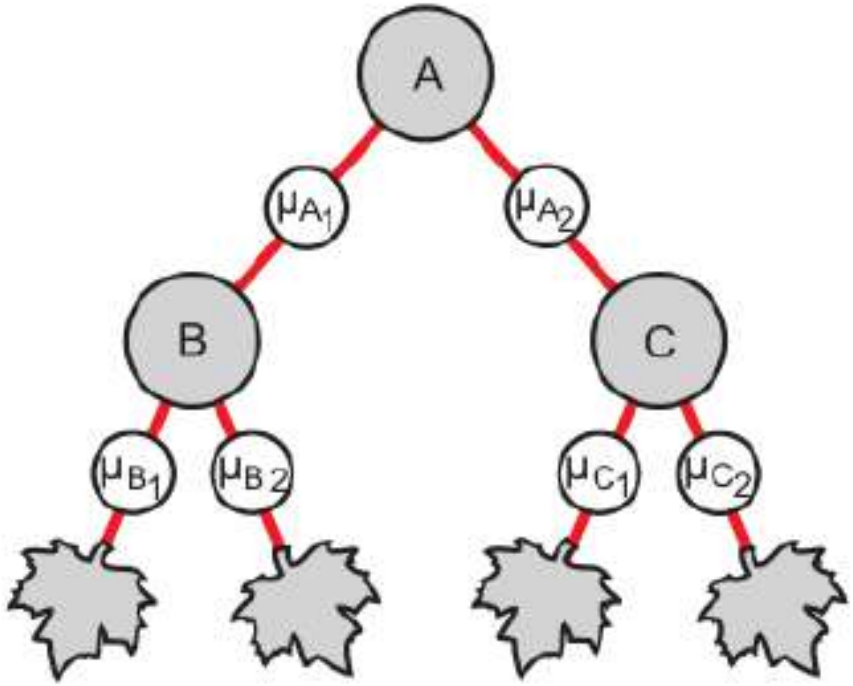
- Exponentially large space of possible decision trees.
- Greedy approaches are often unable to find the best tree.
- Potential interactions between attributes are not taken into account.
- Each decision boundary involves only a single attribute.

Extensions

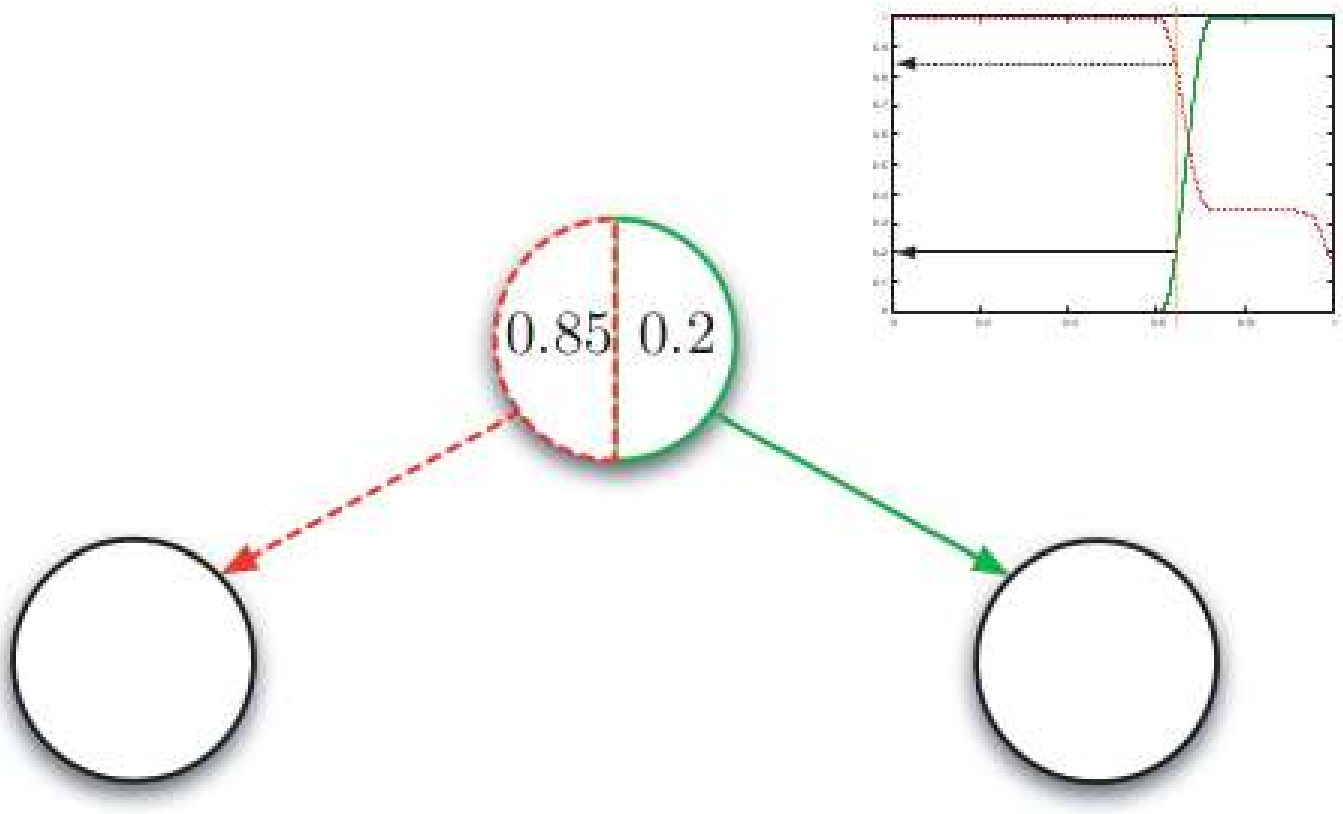
- Other algorithms: CART, C4.5...
  - CART: tree expansion (binary tree) based on *GINI* criterion, then tree pruning (minimizing the error on a validation set).
  - C4.5: uses the entropy criterion.
- Fuzzy decision trees.
- Links with Conditional Preference Nets.
- ...



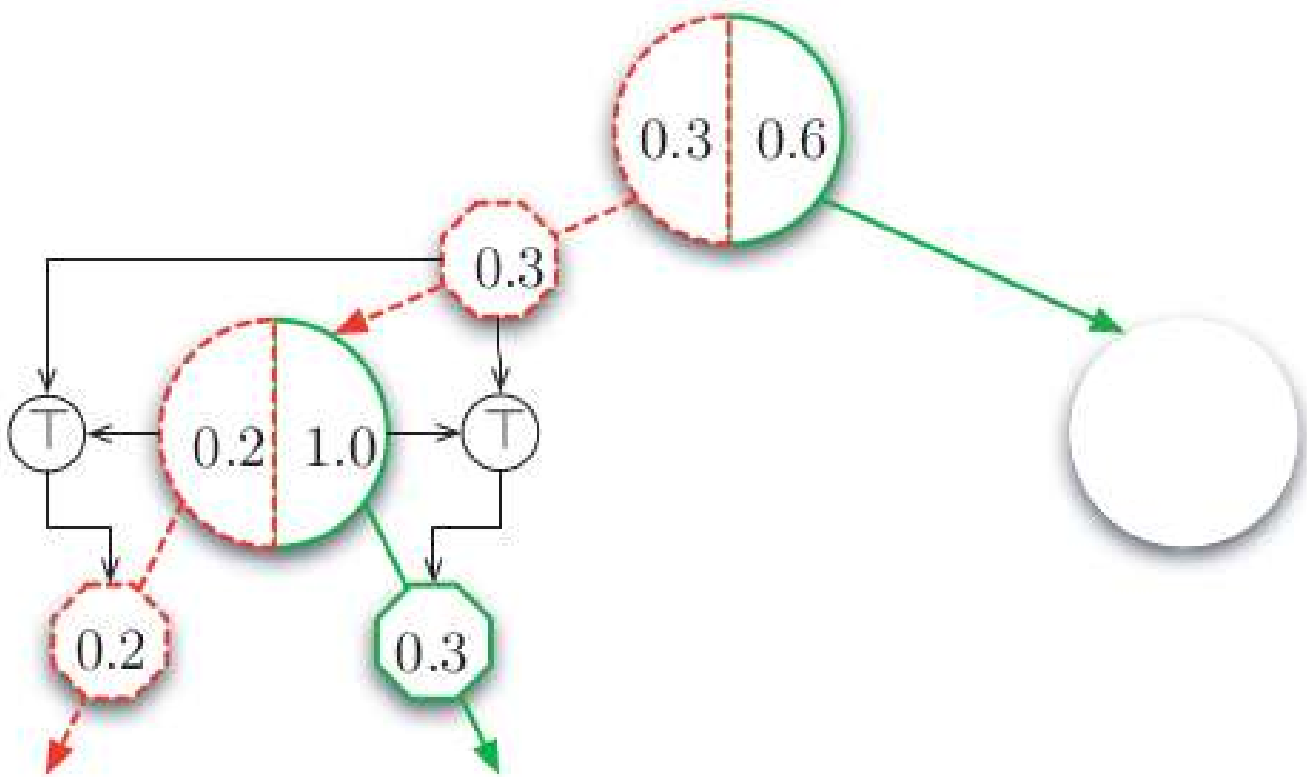
(a) Crisp Decision Tree



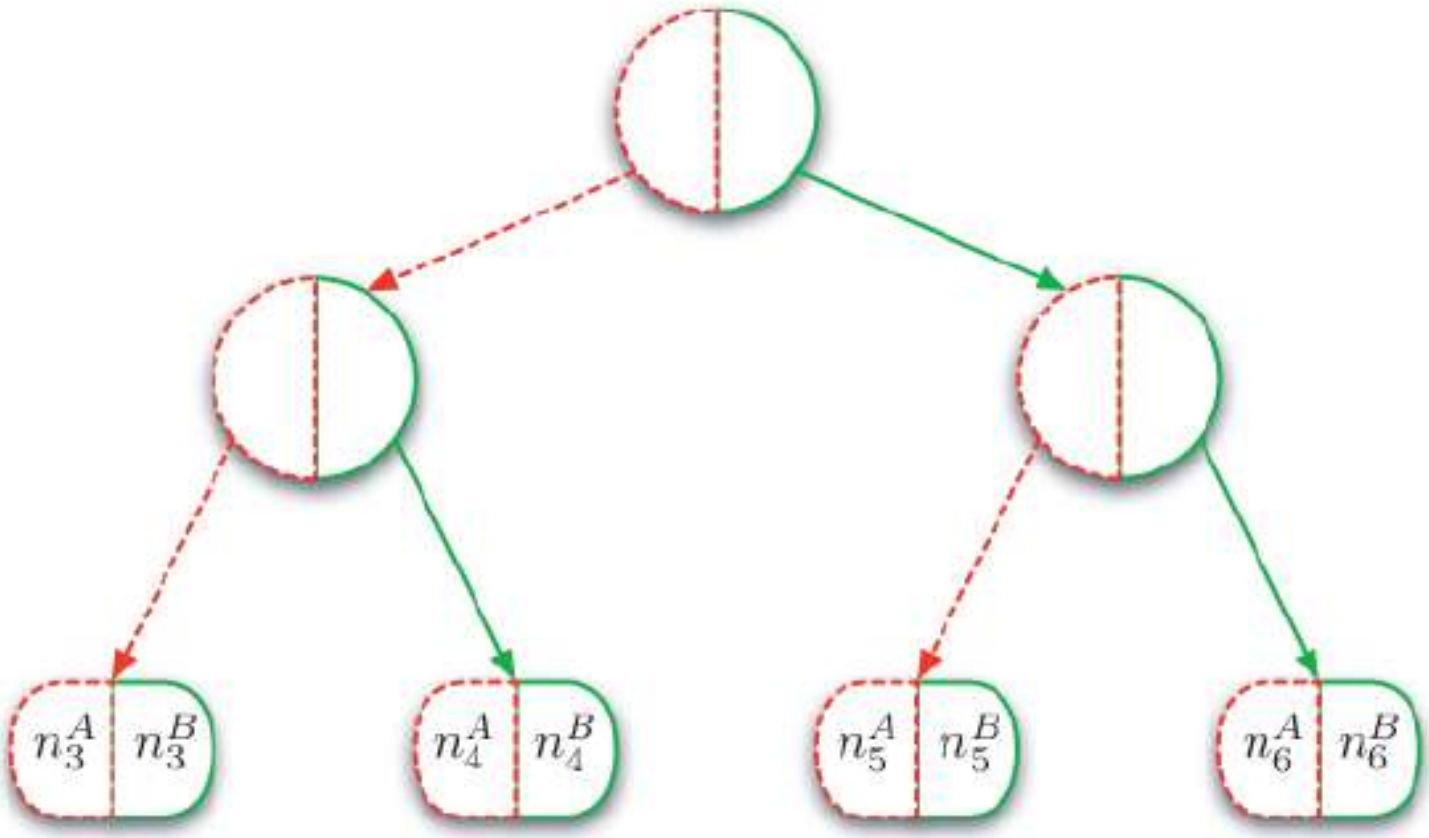
(b) Fuzzy Decision Tree



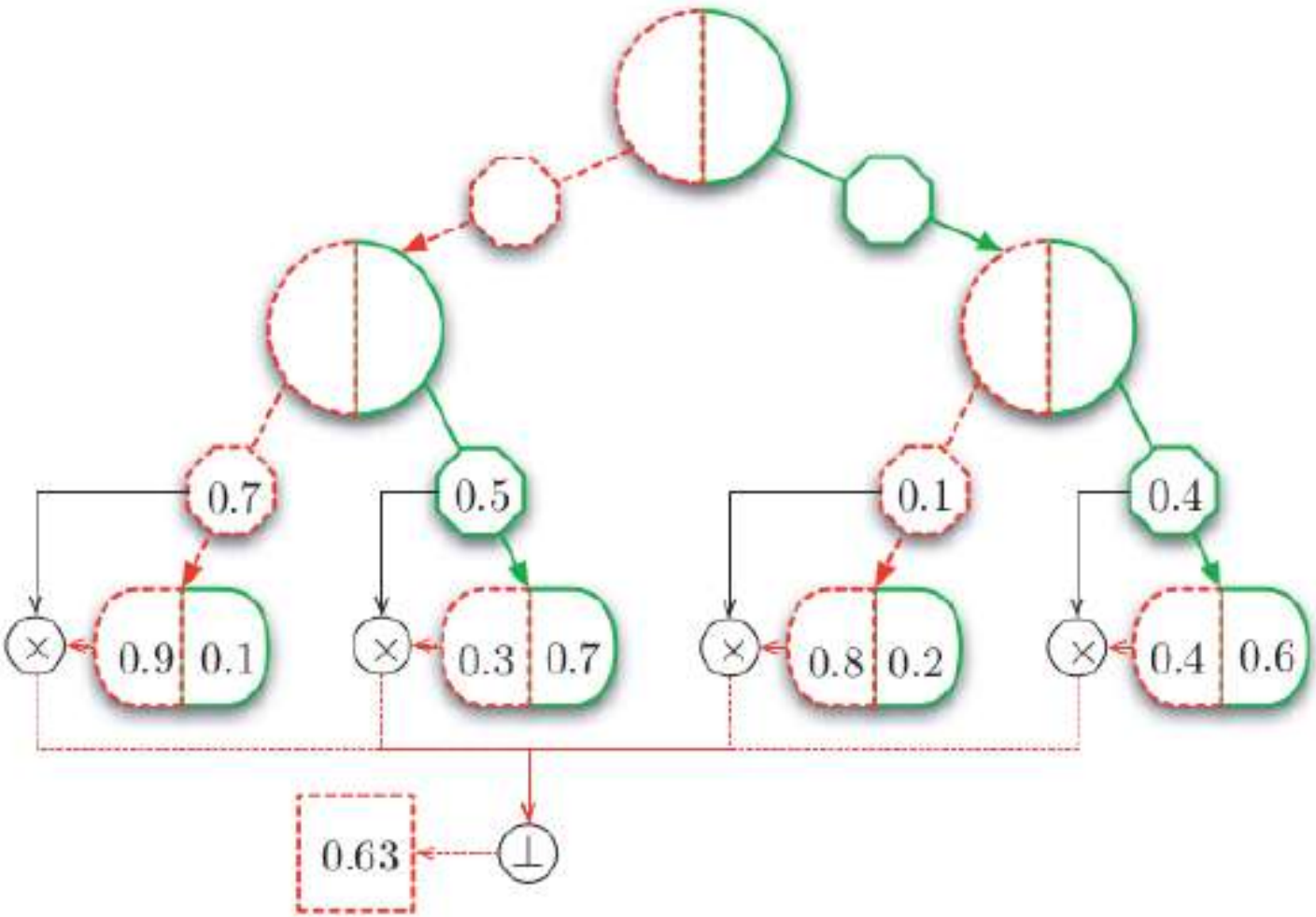
A sample arriving at the root node is tested with the corresponding node test. This results in a membership degree to each of the two classes (class A and class B, left half and right half of the root node respectively).



Computation of cumulative degrees using the minimum t-norm (fuzzy conjunction). Represented by the octagons.

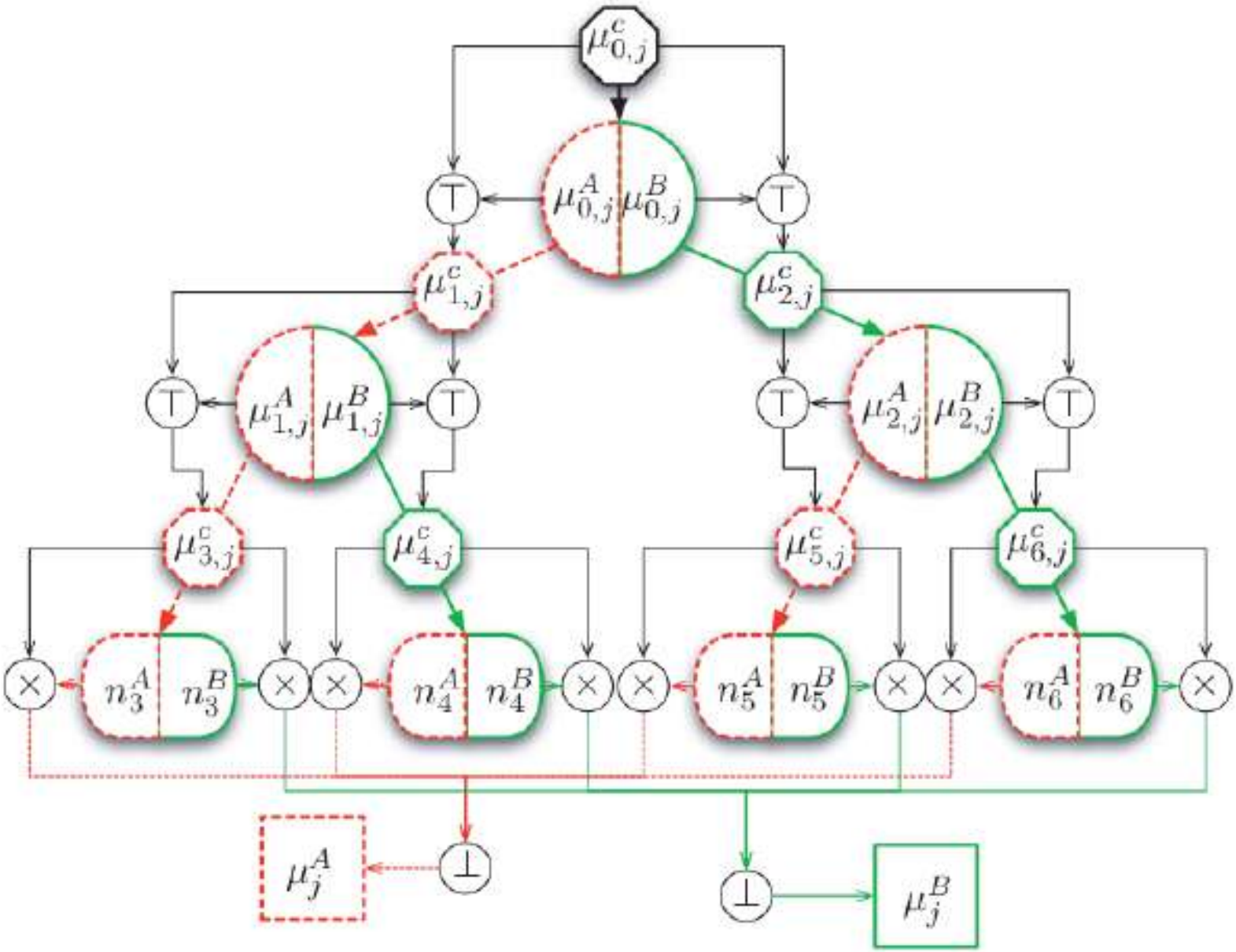


Class densities at leaves: from the sum of the cumulative degrees in a given leaf of all samples belonging to a class.

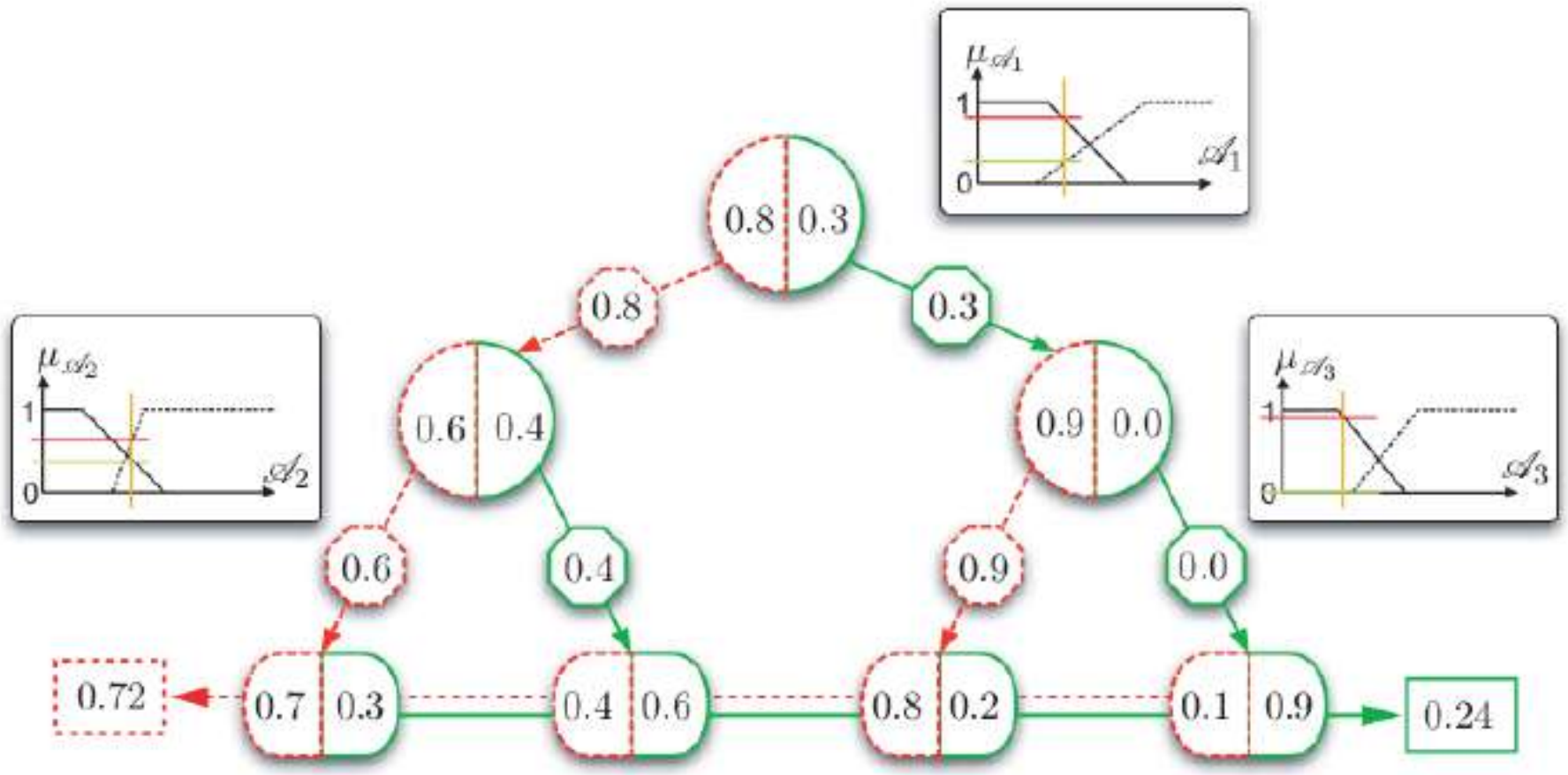


Obtaining a final decision: membership degree to class A is computed by weighting the cumulative degrees in each leaf with the corresponding class distribution for class A. The resulting values are then combined over the full set of leaves using the max t-conorm (fuzzy disjunction).

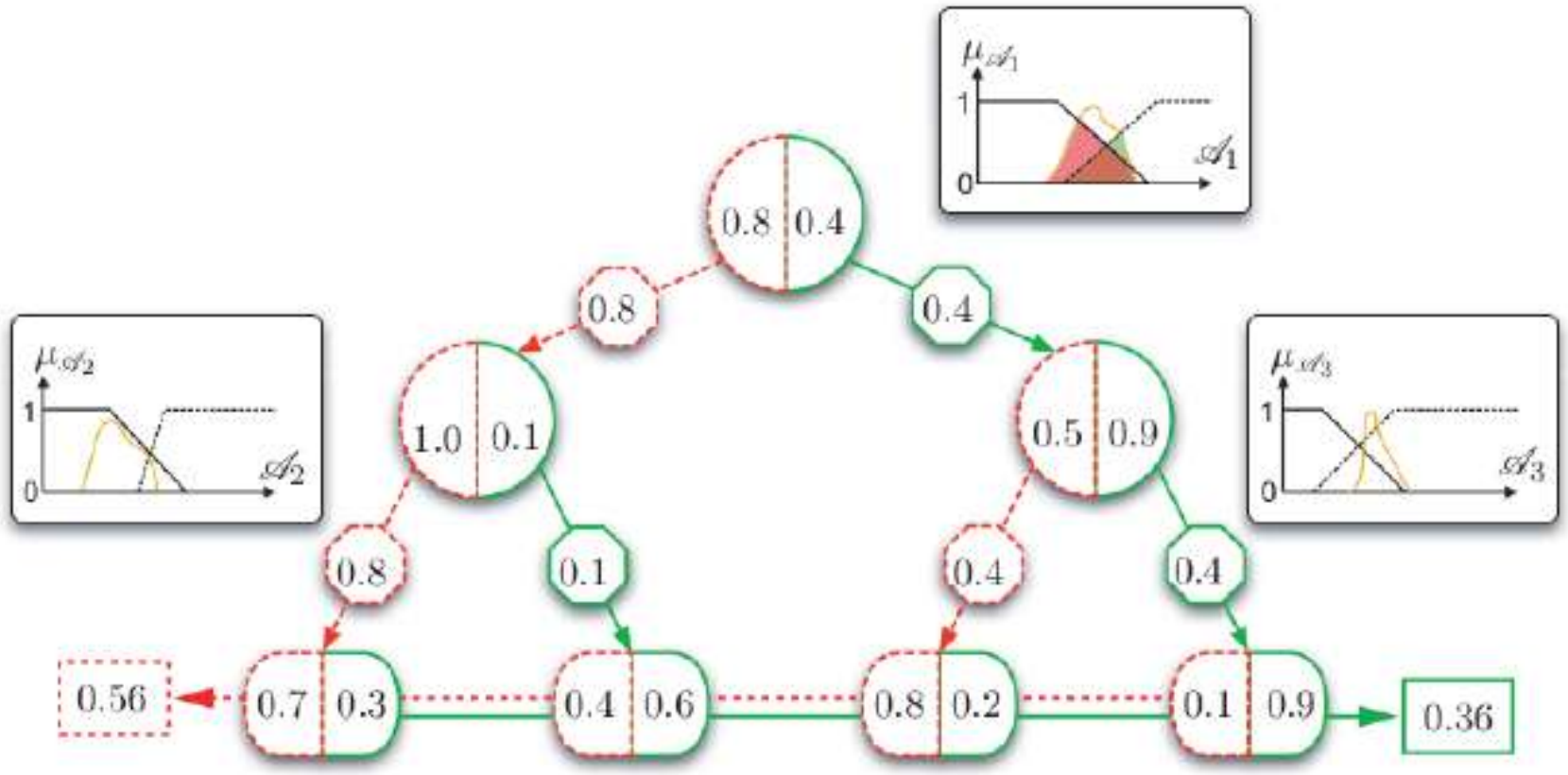
Complete decision tree:



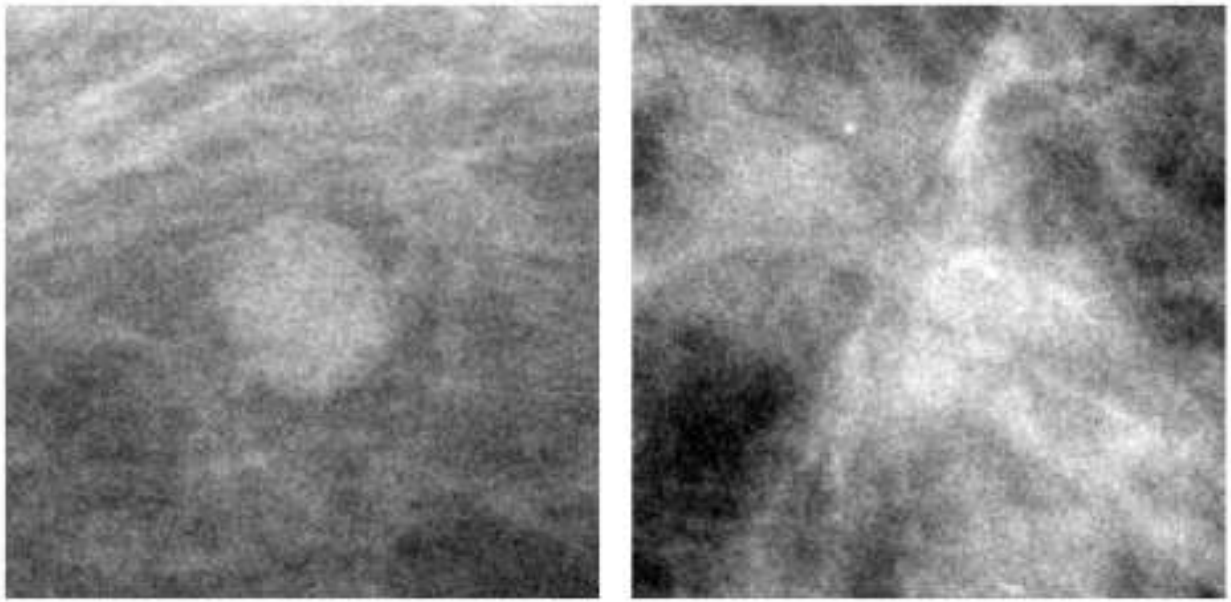
Example for a crisp sample:

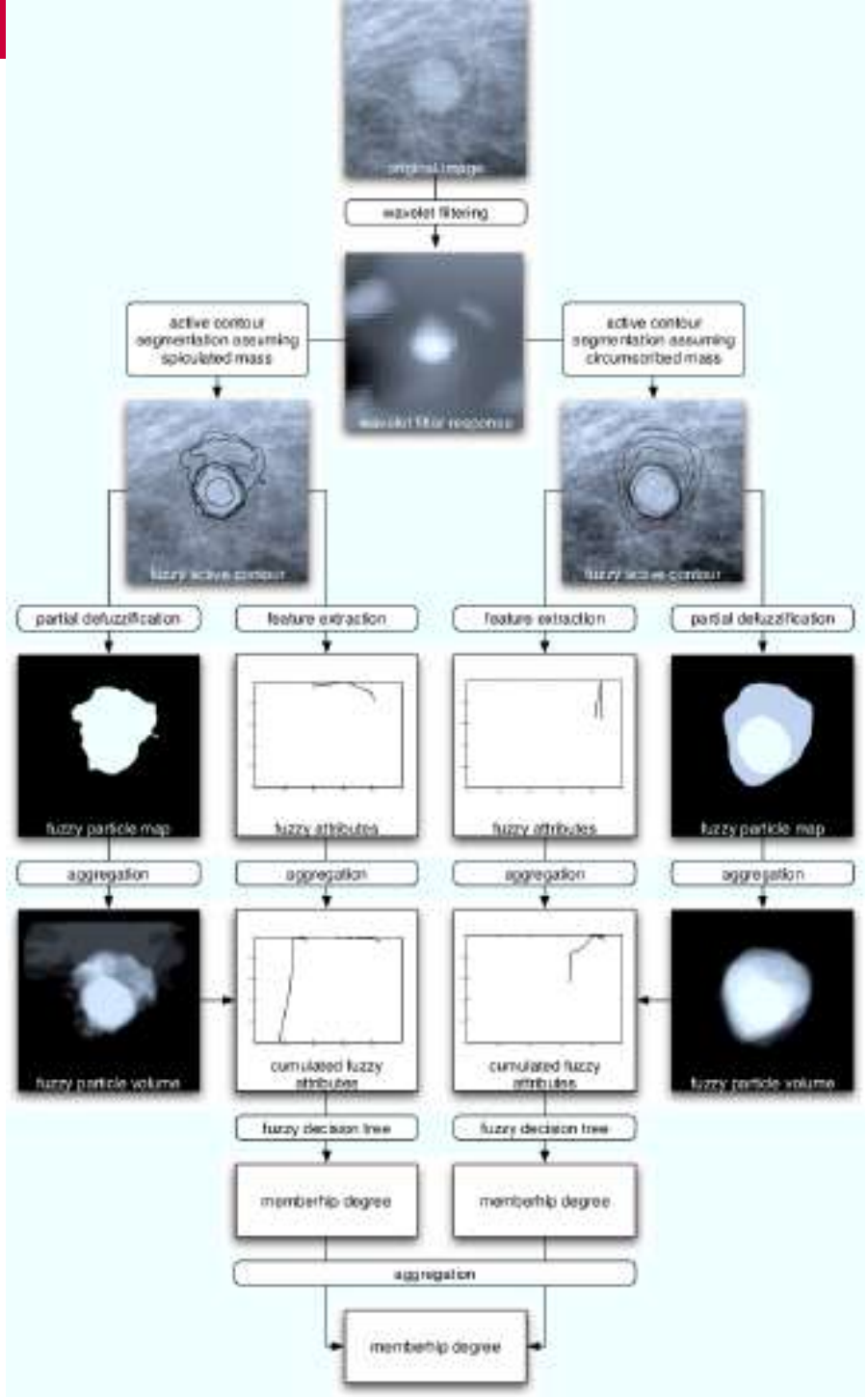


Example for a fuzzy sample:



Example in mammography (PhD Gero Peters)




original image
wavelet filtering
active contour segmentation assuming spiculated mass
active contour segmentation assuming circumscribed mass
wavelet filter response
fuzzy active contour
fuzzy active contour
partial defuzzification
feature extraction
feature extraction
partial defuzzification
fuzzy particle map
fuzzy attributes
fuzzy attributes
fuzzy particle map
aggregation
aggregation
aggregation
aggregation
fuzzy particle volume
cumulated fuzzy attributes
cumulated fuzzy attributes
fuzzy particle volume
fuzzy decision tree
fuzzy decision tree
membership degree
membership degree
aggregation
membership degree