## Data Mining II

## Classifications and rules

# Data Mining Process

* Data Gathering, e.g., data warehousing
* Data preparation and Cleansing
* Pattern Extraction and Discovery (mining)
* Visualisation of the Data
* Analysis and Evaluation of Results

**ID3 algorithm - *An example* (Fruit)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Conclusion | Skin | Colour | Size | Flesh |
| Safe | Hairy | Brown | Large | Hard |
| Safe | Hairy | Green | Large | Hard |
| Dangerous | Smooth | Red | Large | Soft |
| Safe | Hairy | Green | Large | Soft |
| Safe | Hairy | Red | Small | Hard |
| Safe | Smooth | Red | Small | Hard |
| Safe | Smooth | Brown | Small | Hard |
| Dangerous | Hairy | Green | Small | Soft |
| Dangerous | Smooth | Green | Small | Hard |
| Safe | Hairy | Red | Large | Hard |
| Safe | Smooth | Brown | Large | Soft |
| Dangerous | Smooth | Green | Small | Soft |
| Safe | Hairy | Red | Small | Soft |
| Dangerous | Smooth | Red | Large | Hard |
| Safe | Smooth | Red | Small | Hard |
| Dangerous | Hairy | Green | Small | Hard |

## ID3 Algorithm - Decision Tree

**[C1, C2, …, C16]**

**[10+, 6-]**

***Colou****r*

***Green***

***Brow***

***Red***

**[C2, C4, C8, C9, C12, C16] [2+, 4-]**

**Safe**

**[C3,C5,C6,C10,C13,C14,C15] [5+, 2-]**

***Size***

***Size***

***Large***

***Small***

**Safe Dangerous**

**Safe**

***Small***

***Large***

**[C3,C10,C14]**

**[1+, 2-]**

***Skin***

***Smooth***

***Hairy***

**Dangerous**

**Safe**

## ID3 - Cont'd

Decision trees do classification:

* Classify examples by sorting them down the tree from the root node to some leaf notes
* Learned function represented by tree
* Each node in tree is tested on some attribute of an instance
* Branches represent values of attributes
* Follow tree from root to leaves for output value

## ID3 – Cont’d

How do you determine which attribute best classifies data?

Entropy: Given a target attribute *C*, which can have values c1, c2, …, cn (in the example above, n = 2 and the values are `safe' or `dangerous') and a certain attribute *A*,

which can take values a1, …, am. Then the entropy of

attribute *A* with respect to the target attribute *C* is

*Entropy* 

*m*



*j* 1

*n*

*p*(*aj* )

*i* 1

*p*(*ci*

| *aj* ) log2

*p*(*ci*

| *aj* )

where *p*(*ci*|*aj*): the probability that *ci*, given *aj*

*p*(*Conclusion = safe* | *Skin = Hairy*) = 6/8 = 3/4.

Entropy can provide the following:

* Entropy is a measure of 'degree of doubt'
* The higher it is, the more doubt there is about

the possible conclusions

* The attribute which has the lowest entropy is the

most useful determiner

1. For each attribute, compute its entropy with

respect to the conclusion (the target attribute)

1. Select the attribute (say A) with lowest entropy
2. Divide the data into separate sets so that within a set, A has a fixed value (eg *Colour=green* in one set, *Colour=brown* in another, etc)
3. Build a tree with branches:

if *A=a1* then ... (subtree1) if *A=a2* then ... (subtree2)

...etc...

1. For each subtree, repeat this process from step 1
2. At each iteration, one attribute gets removed from consideration. The process stops when there are no attributes left to consider, or when all the data being considered in a subtree have the same value for the conclusion (e.g. they all say *Conclusion=safe*).

* It generates a detailed decision tree.
* With training data provided, it is always able to generate a tree.
* it is easily implemented
* The output is easily to be understood and

interpreted

* The process is simple process
* Its running time increases only linearly with the

complexity of the problem

* ID3 only consider nominal attributes
* Wholly spurious correlations are possible, since the algorithm takes no account of any meaning that the data it works on may have
* ID3 considers just one attribute at a time
* When inducing rules from large sets of examples in which there are a large number of possible outcomes, then the algorithm can be very sensitive to apparently trivial changes in the set of examples
* ID3 cannot generate uncertain rules or handle uncertain data

## ID3 - Limitations

Consider the data at right

|  |  |  |
| --- | --- | --- |
| *Outcome* | X | Y |
| Yes | 3 | 3 |
| No | 2 | 1 |
| Yes | 4 | 4 |
| No | 2 | 4 |
| No | 1 | 3 |
| Yes | 1 | 1 |
| Yes | 2 | 2 |
| No | 2 | 3 |

A rule:

IF *X = Y*

THEN *Outcome = yes*

ELSE *Outcome = no*

## Handling numeric attributes

Suppose temperature and outcome were as follows:

64 65 68 69 70 71 72 72 75 75 80 81 83 85

y n y y y n n y y y n y y n

64 |65 |68 69 70 |71 72 |72 75 75 |80 |81 83 |85

y |n | y y y |n n |y y y |n |y y |n

Problem: for a temperature of 72, both no and yes are valid results

## Handling numeric attributes – cont’d

### Within any group, allow mixed results (both no and yes) but the majority result in the group has at least a specified minimum number of instances

* Suppose we set that at two:

64 65 68 69 70 71 72 72 75 75 | 80 81 83 85

y n y y y n n y y y | n y y n

## Handling numeric attributes – Cont’d

* Suppose we set that at two:

64 65 68 69 70 71 72 72 75 75 | 80 81 83 85

y n y y y n n y y y | n y y n

* leading to the rules:

if temperature <= 77.5 (i.e. midway between 75 and 80) then outcome = yes (3 errors in 10 cases)

if temperature > 77.5

then outcome = no (2 errors in 4 cases)

for a total of 5 errors in 14 cases

* Weight the evaluation measure with the fraction of

cases with known values.

– Idea: The attribute provides information only if it is known.

* Try to find a surrogate test attribute with similar properties (CART, Breiman et al. 1984)
* Assign the case to all branches, weighted in each branch with the relative frequency of the corresponding attribute value (C4.5, Quinlan 1993)

Classification

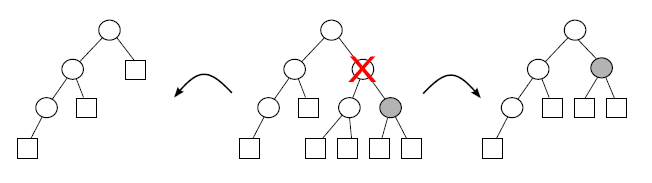
* Use the surrogate test attribute found.
* Follow all branches of the test attribute, weighted with their relative number of cases, aggregate the class distributions of all leaves reached, and assign the majority class of the aggregated class distribution

### Pruning serves the purpose

* to simplify the tree (improve interpretability),
* to avoid overfitting (improve generalization).

### Basic ideas:

* + Replace “bad” branches (subtrees) by leaves.
  + Replace a subtree by its largest branch if it is better.



### Common approaches:

* Reduced error pruning
* Pessimistic pruning
* Confidence level pruning
* It is an extension of ID3 algorithm
* It builds decision trees from a set of

training data in the same way as ID3.

Some of the improvements are:

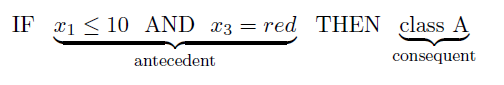
* Handling both continuous and discrete attributes
* Handling training data with missing

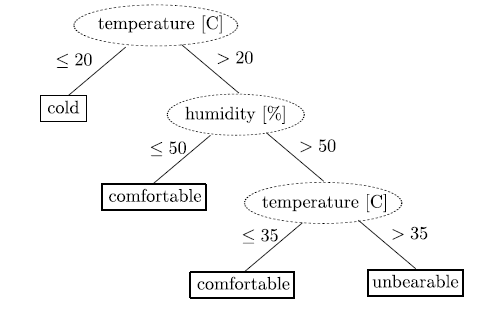
attribute

* Pruning trees after creation

Propositional Rules:

* Rules from Decision Trees
  + Propositional Rules are simple:
    - only atomic facts
    - combinations of facts using only logical operators
    - no variables (in contrast to first order rules)
  + Example





* + *Ra* : IF temperature ≤ 20 THEN class “cold”
  + *Rb* : IF temperature > 20 AND humidity ≤ 50

THEN class “comf.”

* + *Rc* : IF temperature ∈ (20; 35] AND humidity > 50

THEN class “comf.”

* + *Rd* : IF temperature > 35 AND humidity > 50

THEN class “unbearable”

* 𝑅1 : IF temperature ≤ 20 THEN class “cold”
* 𝑅2 : IF humidity ≤ 50 THEN class “comf.”
* 𝑅3 : IF temperature ≤ 35 THEN class “comf.”
* 𝑅4 : class = “unbearable”

↑ order of rules matters!

## Data Preparation

Convert Numeric to Nominal values:

* weather.arff
* weather.nominal.arff

# weather.arff

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.** | **Outlook** | **Temperature** | **Humidity** | **Windy** | **Play** |
| 1 | Sunny | 85 | 85 | False | No |
| 2 | Sunny | 80 | 90 | True | No |
| 3 | Overcast | 83 | 83 | False | Yes |
| 4 | Rain | 70 | 96 | False | Yes |
| 5 | Rain | 68 | 80 | False | Yes |
| 6 | Rain | 65 | 70 | True | No |
| 7 | Overcast | 64 | 65 | True | Yes |
| 8 | Sunny | 72 | 95 | False | No |
| 9 | Sunny | 69 | 70 | False | Yes |
| 10 | Rain | 75 | 80 | False | Yes |
| 11 | Sunny | 75 | 70 | True | Yes |
| 12 | Overcast | 72 | 90 | True | Yes |
| 13 | Overcast | 81 | 75 | False | Yes |
| 14 | Rain | 71 | 91 | True | No |

## Quantification

Group numerical values:

* Temperature
* Hot if value ≥ 80
* Mild if value ∈ [70, 80)
* Cool if value < 70
* Humidity
* High if value > 80
* Normal if value ≤ 80

# An Example

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **No.** | **Outlook** | **Temperature** | **Humidity** | **Windy** | **Play** |
| 1 | Sunny | Hot | High | False | No |
| 2 | Sunny | Hot | High | True | No |
| 3 | Overcast | Hot | High | False | Yes |
| 4 | Rain | Mild | High | False | Yes |
| 5 | Rain | Cool | Normal | False | Yes |
| 6 | Rain | Cool | Normal | True | No |
| 7 | Overcast | Cool | Normal | True | Yes |
| 8 | Sunny | Mild | High | False | No |
| 9 | Sunny | Cool | Normal | False | Yes |
| 10 | Rain | Mild | Normal | False | Yes |
| 11 | Sunny | Mild | Normal | True | Yes |
| 12 | Overcast | Mild | High | True | Yes |
| 13 | Overcast | Hot | Normal | False | Yes |
| 14 | Rain | Mild | High | True | No |

**References**

* I. Witten, E. Frank and M. Hall (2011)

*Data Mining: Practical Machine Learning Tools*

*and Techniques*,

Morgan Kaufmann Publishers.

* M. Berthold, C. Borgelt, F. Hoppner and F. Klawonn

Guide to Intelligent Data Analysis, Spriner-Verlag,

2010

– Chapter 8, Section 8.1, 8.4

* <http://www.kdnuggets.com/>