ULB

UNIVERSITÉ LIBRE DE BRUXELLES

Implementation of Decision Tree Classifiers
ID3 versus C4.5

Depuydt Antoine
Dansy Efila
Mudura Mircea

May, 2017

Introduction

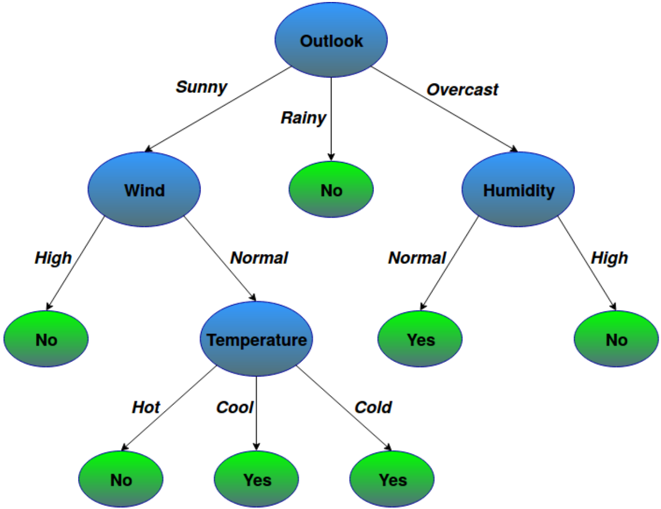
- Data mining: compress, understand and predict
  - Clustering
  - Classification
  - Regression
  - ...
- Techniques to find links
  - Linear Regression
  - Decision Trees
  - Neural Networks
  - ...

Classification

- Classical example: play tennis today?
  - **Features**:
    - Outlook: sunny, overcast, rainy
    - Temperature: hot, cool, cold
    - Wind: high, weak
    - Humidity: high, normal
  - **Class labels**:
    - Yes
    - No

Decision Tree

- Visual model, easily understandable
- Model: tree with decision and leaf nodes



Premise

- Given a training data-set
- Recursively split on a node:
- If node is pure return leaf (class value)
- Else compute entropy & info gain:
  - Shannon's entropy: $E(S) = \sum_i - p_i log_2(p_i)$
  - Subtree gain: $Gain(T, X) = E(T) - E(T, X)$

ID3 versus C4.5

- Goal: implement ID3 and C4.5 algorithms
- Objectives: compare ID3 and C4.5 output
    - Compare ID3 and C4.5
    - Create an application that classifies any data using both algorithms

ID3

- Initial implementation of decision trees
- Top down approach
- Split current node based on information gain:

Improvements?

- Entropy & information gain not sufficient metrics
- Missing data has to be handled
- Numerical values could provide order or dimension to a problem set
- Tree can be simplified

C4.5

Missing data

K-fold cross validation

Demonstration