ID3 Algorithm

Abbas Rizvi
CS157 B
Spring 2010
What is the ID3 algorithm?

• ID3 stands for Iterative Dichotomiser 3
• Algorithm used to generate a decision tree.
• ID3 is a precursor to the C4.5 Algorithm.
History

• The ID3 algorithm was invented by Ross Quinlan.
• Quinlan was a computer science researcher in data mining, and decision theory.
• Received doctorate in computer science at the University of Washington in 1968.
Decision Tree

- Classifies data using the attributes
- Tree consists of decision nodes and decision leafs.
- Nodes can have two or more branches which represents the value for the attribute tested.
- Leaf nodes produces a homogeneous result.
The algorithm

- The ID3 follows the Occam’s razor principle.
- Attempts to create the smallest possible decision tree.
The Process

- Take all unused attributes and calculates their entropies.
- Chooses attribute that has the lowest entropy is minimum or when information gain is maximum.
- Makes a node containing that attribute.
The Algorithm

- Create a root node for the tree
- If all examples are positive, Return the single-node tree Root, with label = +.
- If all examples are negative, Return the single-node tree Root, with label = -.
- If number of predicting attributes is empty, then Return the single node tree Root, with label = most common value of the target attribute in the examples.
The Algorithm (cont.)

• Else
  – A = The Attribute that best classifies examples.
  – Decision Tree attribute for Root = A.
  – For each possible value, $v_i$, of A,
    • Add a new tree branch below Root, corresponding to the test $A = v_i$.
    • Let Examples($v_i$), be the subset of examples that have the value $v_i$ for A
    • If Examples($v_i$) is empty
      – Then below this new branch add a leaf node with label = most common target value in the examples
    • Else below this new branch add the subtree ID3 (Examples($v_i$), Target_Attribute, Attributes – {A})
  
• End
• Return Root
Entropy

- Formula to calculate
- A complete homogeneous sample has an entropy of 0
- An equally divided sample as an entropy of 1
- Entropy $= - p+ \log_2 (p+) - p- \log_2 (p-)$ for a sample of negative and positive elements.

$$Entropy(S) = \sum_{i=1}^{C} p_i \log_2 p_i$$
Exercise

- Calculate the entropy
- Given:
  - Set S contains 14 examples
  - 9 Positive values
  - 5 Negative values
Exercise

- Entropy(S) = \(-\frac{9}{14} \log_2 \left(\frac{9}{14}\right) - \frac{5}{14} \log_2 \left(\frac{5}{14}\right)\)
- = 0.940
Information Gain

- Information gain is based on the decrease in entropy after a dataset is split on an attribute.
- Looking for which attribute creates the most homogeneous branches
Information Gain Example

- 14 examples, 9 positive 5 negative
- The attribute is Wind.
- Values of wind are Weak and Strong
Exercise (cont.)

- 8 occurrences of weak winds
- 6 occurrences of strong winds
- For the weak winds, 6 are positive and 2 are negative
- For the strong winds, 3 are positive and 3 are negative
Exercise (cont.)

• Gain(S, Wind) =

• Entropy(S) - (8/14) * Entropy (Weak) - (6/14) * Entropy (Strong)

• Entropy(Weak) = - (6/8) * log₂(6/8) - (2/8) * log₂(2/8) = 0.811

• Entropy(Strong) = - (3/6) * log₂(3/6) - (3/6) * log₂(3/6) = 1.00
Exercise (cont.)

- So…
- \(0.940 - (8/14)*0.811 - (6/14)*1.00\)
- = 0.048
Advantage of ID3

• Understandable prediction rules are created from the training data.
• Builds the fastest tree.
• Builds a short tree.
• Only need to test enough attributes until all data is classified.
• Finding leaf nodes enables test data to be pruned, reducing number of tests.
Disadvantage of ID3

- Data may be over-fitted or over-classified, if a small sample is tested.
- Only one attribute at a time is tested for making a decision.
- Classifying continuous data may be computationally expensive, as many trees must be generated to see where to break the continuum.
Questions