More on decision trees

Windowing in ID3

Windowing is applied in ID3 as a way of dealing with large sets of training instances. Without windowing, such an algorithm can be really slow, as it needs to do entropy calculations over huge amounts of data. With windowing, training is done on a relatively small sample of the data, and then checked against the full set of training data.
Here is the windowing algorithm:

1. Select a sample S of the training instances at random. The proportion chosen would need to be small enough that ID3 could run fairly fast, but large enough to be representative of the whole set of examples.

2. Run the ID3 algorithm on the set of training instances to obtain a decision tree.

3. Check the decision tree on the full data set, to obtain a set E of training instances that are misclassified by the current tree.

4. If E is empty, stop.

5. Let the new set of training instances S' be the union of S and E.

6. Go to step 2 and run the ID3 to induce a tree from S'.

Pruning in ID3

Pruning is the final step after growing tree. Purpose: avoid overfitting. The idea is to produce a smaller (simpler) tree with improved generalization properties.

- Split data into training and validation sets.
- Try removing each possible node and evaluate impact on validation set.
- Remove node that improves accuracy the most on validation set.
- Repeat until no additional improvement possible.

Handling continuous attributes in ID3

Discretize.

CART

CART stands for "Classification and Regression Trees". It is similar to ID3 with a few differences. The most important difference is in the way pruning is computed. The idea of CART is to build many "small" trees and combine them in order to form "larger" trees.