Inducing Decision Trees

Azadeh Samadian

In this project I implement and test the decision tree learning algorithm (Mitchell, Chapter 3).

- Download the two datasets available in the project data_set folder. Each data set is divided into three sets: the training set, the validation set and the test set. Data sets are in CSV format. The first line in the file gives the attribute names. Each line after that is a training (or test) example that contains a list of attribute values separated by a comma. The last attribute is the class-variable. Assume that all attributes take values from the domain {0,1}.

- Implement the decision tree learning algorithm. As discussed in class, the main step in decision tree learning is choosing the next attribute to split on. Implement the following two heuristics for selecting the next attribute.

-

1. Information gain heuristic (Mitchell Chapter 3).
2. Variance impurity heuristic described below.

Let K denote the number of examples in the training set. Let K0 denote the number of training examples that have class = 0 and K1 denote the number of training examples that have class = 1. The variance impurity of the training set S is defined as:

$$VI(S) = \frac{K_0}{K}\frac{K_1}{K}$$

Notice that the impurity is 0 when the data is pure. The gain for this impurity is defined as usual.

$$Gain(S, X) = VI(S) - \sum_{x \in Values(X)} \Pr(x) VI(S_x)$$

where X is an attribute, Sx denotes the set of training examples that have X = x and Pr(x) is the fraction of the training examples that have X = x (i.e., the number of training examples that have X = x divided by the number of training examples in S).

- Implement the post pruning algorithm given below as Algorithm 1 (See Mitchell, Chapter 3).

- Implement a function to print the decision tree to standard output. I will use the following format.

```
wesley = 0 :
| honor = 0 :
| | barclay = 0 : 1
| | barclay = 1 : 0
| honor = 1 :
| | tea = 0 : 0
| | tea = 1 : 1
wesley = 1 : 0
```

According to this tree, if wesley = 0 and honor = 0 and barclay = 0, then the class value of the corresponding instance should be 1. In other words, the value appearing before a colon is an attribute value, and the value appearing after a colon is a class value.

Algorithm 1: Post Pruning

Input: An integer L and an integer K
Output: A post-pruned Decision Tree
begin
  Build a decision tree using all the training data. Call it D;
  Let $D_{Best} = D$;
  for $i = 1$ *to* L **do**
    Copy the tree D into a new tree D';
    M = a random number between 1 and K;
    for $j = 1$ *to* M **do**
      Let N denote the number of non-leaf nodes in the decision tree D'. Order the nodes in D' from 1 to N;
      P = a random number between 1 and N;
      Replace the subtree rooted at P in D' by a leaf node. Assign the majority class of the subset of the data at P to the leaf node.;
      /* For instance, if the subset of the data at P contains 10 examples with $class = 0$ and 15 examples with $class = 1$, replace P by $class = 1$ */
    end
    Evaluate the accuracy of D' on the validation set;
    /* accuracy = percentage of correctly classified examples */
    if D' *is more accurate than* D_{Best} **then**
      $D_{Best} = D'$;
    end
  end
  return D_{Best};
end

- To run the code from the command line, it takes the following six arguments as the input:

.\program <L> <K> <training-set> <validation-set> <test-set> <to-print>

L: integer (used in the post-pruning algorithm)

K: integer (used in the post-pruning algorithm)

to-print:{yes,no}

- The accuracy on the test set for decision trees constructed is reported using the two heuristics mentioned above.

Below are 10 different combinations of L and K and the reported accuracies for two heuristic methods and accuracies for the post-pruned decision trees constructed using the two heuristics.

---

for L= 8 and K= 4
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.765
Tree Accuracy of Variance Impurity Pruned Tree: 0.7575

---

for L= 4 and K= 8
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.7605
Tree Accuracy of Variance Impurity Pruned Tree: 0.7615

---

for L= 30 and K= 10
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.7605
Tree Accuracy of Variance Impurity Pruned Tree: 0.7605

---

for L= 10 and K= 30
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.763
Tree Accuracy of Variance Impurity Pruned Tree: 0.7605

---

for L= 20 and K= 20
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.763
Tree Accuracy of Variance Impurity Pruned Tree: 0.7605

---

for L= 48 and K= 8
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.763
Tree Accuracy of Variance Impurity Pruned Tree: 0.7605

---

for L= 8 and K= 48
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.763
Tree Accuracy of Variance Impurity Pruned Tree: 0.7605

---

for L= 100 and K= 10
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.763
Tree Accuracy of Variance Impurity Pruned Tree: 0.7605

---

for L= 10 and K= 100
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.763
Tree Accuracy of Variance Impurity Pruned Tree: 0.7605

---

for L= 12 and K= 11
Tree Accuracy by Information Gain: 0.7585
Tree Accuracy by Variance Impurity: 0.752
Tree Accuracy of Information Gain Pruned Tree: 0.763
Tree Accuracy of Variance Impurity Pruned Tree: 0.7605
