**Machine Learning**

**Assignment-1 Decision Tree**

Manish Biyani | mxb172930

* Report the accuracy of two data sets available on class webpage.
* Each data set is divided into three sets: the training set, the validation set and the test set. Data sets are in CSV format.
* Accuracy is calculated for 2 heuristics:

1. Information Gain Heuristic
2. Variance Impurity Heuristic

* Accuracy measured are as follows:

1. Data-set 1:

Without Pruning (Information Gain)

1. Training: 100%
2. Validation: 75.9%
3. Test: 75.25%

Without Pruning (Variance Impurity)

a. Training: 100%

1. Validation: 77.05%
2. Test: 76.7%
3. Data-set 2:

Without Pruning (Information Gain)

a. Training: 99.83%

1. Validation: 76.16%
2. Test: 72.67%

Without Pruning (Variance Impurity)

a. Training: 99.83%

1. Validation: 76.5%
2. Test: 73%

* Choose 10 suitable values for L and K. For each of them, report the accuracies for the post-pruned decision trees constructed using the two heuristics.
* **Data-set 1**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Index | L | K | Accuracy-Info Gain-Test  Validation | Accuracy- Info Gain-Imp-  Test | Accuracy-Variance Imp-Validation | Accuracy-Variance Imp-Test |
| 1. | 5 | 10 | 76.3% | 75.7% | 77.15% | 77.25% |
| 2. | 5 | 15 | 76.25% | 74.45% | 77.75% | 76.7% |
| 3. | 7 | 10 | 76.05% | 75.35% | 77.05% | 76.7% |
| 4. | 10 | 10 | 77.3% | 76.5% | 77.9% | 77.5% |
| 5. | 10 | 15 | 76.95% | 75.9% | 77.10% | 77.3% |
| 6. | 10 | 20 | 78.25% | 77.65% | 77.05% | 76.7% |
| 7. | 10 | 30 | 76.0% | 75.1% | 78.25% | 78.15% |
| 8. | 12 | 35 | 75.95% | 76.1% | 77.05% | 76.7% |
| 9. | 15 | 40 | 75.95% | 75.4% | 77.05% | 76.7% |
| 10. | 15 | 45 | 75.9% | 75.25% | 77.05% | 76.7% |

**Data-set 2**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Index | L | K | Accuracy-Info Gain-Test  Validation | Accuracy- Info Gain-Imp-  Test | Accuracy-Variance Imp-Validation | Accuracy-Variance Imp-Test |
| 1. | 5 | 10 | 77.16% | 73.5% | 76.6% | 73.3% |
| 2. | 5 | 15 | 76.16% | 72.67% | 77.67% | 71.16% |
| 3. | 7 | 10 | 76.16% | 72.67% | 77.16% | 71.33% |
| 4. | 10 | 10 | 78% | 73.67% | 77.33% | 72% |
| 5. | 10 | 15 | 76.67% | 73.33% | 77.67% | 72.83% |
| 6. | 10 | 20 | 77.16% | 72.5% | 77.5% | 73.16% |
| 7. | 10 | 30 | 78.16% | 74.83% | 77.33% | 75.67% |
| 8. | 12 | 35 | 76.33% | 72.33% | 76.5% | 73% |
| 9. | 15 | 40 | 76.16% | 72.67% | 76.67% | 73% |
| 10. | 15 | 45 | 76.5% | 70.5% | 79.5% | 76.33% |

Various parts of the Decision Tree:

* The projects consists of 6 files:
  1. Main.java
  2. Constant.java
  3. Node.java
  4. DecisionTree.java
  5. Prune1.java
  6. Test.java
  7. Import.java
* Main consists of various connections with other classes. It defines the loops to prune the decision tree as well as function to count leaf nodes and display results.
* Constant is used to define the inputs which remain constant throughtout the program such as test\_set, validation\_set and others.
* Node is used to define various characteristics of nodes and writing getter and setter method for nodes.
* DecisionTree is the heart of the program and consists of all the important methods like building the decision tree and calculating various values such as entropy, information gain, noOfzeroes, variance impurity, and other important things.
* Prune1 is used to implement the pruning in which the random number is used to delete random number of non-leaf nodes.
* Test is used to measure the accuracy of various sets of data and is used for pruning and further predictions.
* Import is used to read the CSV files and defines functions to extract features and data.

Working of decision tree:

* Once we input L,K, Test set, Validation set, test set and print(Y/N), the training set is processed and used to build the decision tree. The decision tree is build for information gain heuristic as well as variance impurity heuristic.
* The validation data is used to calculate the accuracy of our decision tree and the tree is stored as the current best tree.
* In the pruning, L trees are built and max K nodes are removed and accuracy is calculated for validation. The tree with maximum accuracy is used to predict the test set and its accuracy is calculated for the same.
* The accuracy information for all the data sets are printed.
* The tree can be printed by selecting ‘Y’ in the beginning of the tree.

Future aspects:

* I researched with a number of inputs that we can increase the accuracy of a tree if we train the tree again after testing it with validation set.
* All we need to do is take the data where our tree predicted the wrong classification for validation data and then train our tree with that data again. It can increase the accuracy upto 1% in several cases.
* It is yet to be confirmed and I have commented it in my program and can be used with little modifications.

References:

* ML by Tom Mitchell
* Various internet sites
* Pruning algorithm as in our Homework Assignment