INSTALLATION

The following packages are required:

PYTHON 2.7

Additional packages required:

graphviz

pydot

Installation in Windows:

1. Download from <http://www.graphviz.org/Download_windows.php>
2. Install the package
3. Open command prompt with administrator privileges
4. Goto C:\Python27\Scripts
5. run command: “ pip install pydot”

Installation in Ubuntu:

1. open terminal
2. Give command sudo apt-get install pydot ( This also installs graphviz )

EXECUTION

Run the python file in the following format:

.\program <L> <K> <file location of training-set> <file locatin of validation-set> <file location of test-set> <to-print>

ex:

python tree.py 1000 10 data\_sets1/training\_set.csv data\_sets1/validation\_set.csv data\_sets1/test\_set.csv yes

OUTPUT

if print is 'yes'

then two pictures of decision trees will be created in the location of tree.py

file1 → tree.jpg (Original tree)

file2 → prune\_tree.jpg (Pruned tree)

if print is 'no'

the two pictures will not be created

on the output screen the following will be printed:

No.of nodes in Rbest : 41

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.762

pruned nodes , path or ID : ['0-+---++-', '0++-+++', '0+++----+-', '0+++++--+', '0+--++', '0+++', '0++--+-+-', '0++--+--', '0-+-+-++', '0++--+-+', '0++---+', '0-+---+', '0++-', '0+-+--+-', '0-+++--+-', '0----+-+', '0-+-+--+-', '0++', '0----++', '0+-+---++--', '0+-+--+', '0+-+---']

Rbest is pruned tree

Root is original tree

Accuracy of original tree with test data

Accuracy of pruned tree with test data

The path or ID indicates the location of intermediate node removed from original tree

0 → start from root

+ → move right

- → move left

On moving to the specified path in original tree, the node can be identified.

DEBUGGING

You can use python debugger , pdb to debug.

Remove the '#' comment symbol to see the print values of intermediate nodes .

OBSERVATIONS

Here we used Information gain , heuristic for splitting the data.

**for Dataset1**

1. L = 1000 K = 10

No.of nodes in Rbest : 70

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.781

pruned nodes , path or ID : ['0-+---++-', '0----+-+', '0+++----+', '0++-+++-+', '0++-+++++', '0-+-+-++', '0++---+-+', '0++++--++-', '0-+---+-+', '0+++++-', '0++-+--+', '0-+---+', '0+++++', '0-+++--+-+', '0++--+---++', '0++-+-++-', '0++++--+-+', '0+++---', '0----++', '0++-+++-', '0+++-+-', '0++++--', '0+--++-+', '0++--+-+', '0++--+--', '0++++', '0++-++++', '0+-+---+', '0-+++--+-']

1. L = 100 K = 50

No.of nodes in Rbest : 85

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.757

pruned nodes , path or ID : ['0++--+---++', '0-+--', '0+++--+-', '0-+-+-+', '0+++------', '0++-+', '0+++++', '0----+++', '0++++--+-+', '0++--+---+', '0- +++--', '0---+']

1. L = 500 K = 55

No.of nodes in Rbest : 103

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.767

pruned nodes , path or ID : ['0+++----+-', '0++-+++-', '0-+---++-', '0+--++- +-', '0+-+---+', '0++-+', '0----++', '0++---+-+', '0+++++-', '0++++--+-']

1. L = 200 K = 101

No.of nodes in Rbest : 103

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.7485

pruned nodes , path or ID : ['0+-+--+-', '0-+---++', '0-+-+--+', '0+++++-+', '0-+--+', '0-+---+-+-', '0-+++--+-', '0+++-+--', '0++++', '0++-+++-', '0++-- +---+']

1. L = 800 K = 60

No.of nodes in Rbest : 98

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.76

pruned nodes , path or ID : ['0-+-+-+++', '0-+---++-', '0+--+--+-', '0-+-+- +', '0++++--+', '0-+++--+-+', '0+--++-+', '0++-+++-', '0+--+--', '0-+++--', '0+-+---+', '0+++++--', '0+-----+', '0-+---+']

1. L = 1500 K = 100

No.of nodes in Rbest : 73

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.776

pruned nodes , path or ID : ['0++--+-+', '0+-+---', '0-+++--+-+', '0+++-+- +', '0----+++', '0-+---+', '0++-+-', '0----++', '0++--+---', '0+++---', '0- +-+-+++-', '0++--+--', '0-+-+-+', '0---++', '0+++++-+', '0++-+++-++', '0+++ +--+', '0+++++', '0+--+-', '0+++-+--']

1. L = 400 K = 300

No.of nodes in Rbest : 123

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.7645

pruned nodes , path or ID : ['0-+++--+-+', '0-+-+-+', '0---+++', '0+++-+--', '0+++++']

1. L = 2000 K = 12

No.of nodes in Rbest : 56

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.778

pruned nodes , path or ID : ['0++-+-+', '0+-+---++', '0-+---+', '0++-++', '0++++', '0-+-+-+++', '0-+-+--+', '0-+-+-++', '0-+-+--', '0+-+---', '0+++-+- +-', '0+++-----', '0+--++-+', '0++--+---', '0++-+', '0----++', '0-+-+-+', '0+--+-', '0++--+-+', '0+++-+--', '0+++----+-', '0++--+', '0----+-+', '0+ +---+-', '0+++-+-']

1. L = 2500 K = 175

No.of nodes in Rbest : 58

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.77

pruned nodes , path or ID : ['0++--+----', '0++--+-+', '0----+++', '0+++-+-+ ', '0++-+--+', '0++-+++', '0++---+-', '0-+---++---', '0++++--', '0+--+--+', '0+++++', '0+--++-', '0++-++', '0-+---+', '0+-+---++', '0+-+---', '0-+-+-+++ ', '0++++', '0----+-+', '0++-+-++-', '0+++-+--', '0+--+--', '0++-+-++', '0++ +---', '0-+++--+', '0-+-+-++', '0++--+--', '0+++-+-', '0+--++', '0++--+']

1. L = 700 K = 180

No.of nodes in Rbest : 109

No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.755

pruned nodes , path or ID : ['0+--+--+-', '0-+-+-++', '0++--+', '0++-++++', '0++++--', '0+--+-', '0+++-+-+--', '0+-+---++--']

**for Dataset2**

1. L = 1000 K = 10

No.of nodes in Rbest : 64

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.783333333333

pruned nodes , path or ID : ['0--+-++--+-', '0+-+---+', '0++--++-', '0----+ ', '0--+--+++++', '0+----++', '0+---', '0+-++++', '0--+--', '0--+++----+--', '0+--+-+', '0++--++', '0++--+', '0-+-+', '0++-+--++', '0+-+--+++', '0+-+-+++ -', '0-+--++-+', '0--+++---', '0++-+--+', '0+-+-+-+-+-', '0++-+-++', '0-+--- +', '0+-+-++-']

1. L = 100 K = 50

No.of nodes in Rbest : 86

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.74

pruned nodes , path or ID : ['0--+---', '0--+--++', '0+--++-', '0++--++', '0+-+--+', '0+--+-+++', '0++-+-', '0-+---+', '0+-+-+', '0+---++']

1. L = 500 K = 55

No.of nodes in Rbest : 55

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.753333333333

pruned nodes , path or ID : ['0----++-+-', '0-+-++', '0+-+-+-+', '0--+--', '0+-+--+++', '0+--+-+---', '0+-+--', '0----+', '0-+---++', '0+-+-+-', '0-+-+ ', '0++-+-+++', '0++-+-', '0+----++-', '0++----+', '0+--++--', '0++----', '0--+-++--', '0+----+++++', '0++--+', '0+---++', '0--+++---', '0+-+-+++-', '0+---']

1. L = 200 K = 101

No.of nodes in Rbest : 110

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.728333333333

pruned nodes , path or ID : ['0-+-++--', '0-+---++++', '0++-+--++', '0++-+ --', '0+-+-+++', '0+--++', '0++-+-+++', '0++--+', '0+-+-+-+', '0++-+-', '0+---++']

1. L = 800 K = 60

No.of nodes in Rbest : 98

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.738333333333

pruned nodes , path or ID : ['0-+---++', '0+-+---', '0--+++--', '0----++-', '0+-+-++-', '0++-+--++', '0----+--', '0+-+--', '0--+-++--+', '0+----++-', '0+--+-+---', '0+--+-+++', '0--+-++--', '0--+--+++++', '0+----++', '0-+-+', '0+-+-+++--']

1. L = 1500 K = 100

No.of nodes in Rbest : 58

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.761666666667

pruned nodes , path or ID : ['0+--', '0++-+-', '0--+++----', '0--+--+++++', '0+-+-+-+', '0--+---', '0-+-+', '0----+-', '0++--++--', '0-+---+++-', '0--+ --+', '0++----++', '0-+--++', '0+-+-++', '0-+---++', '0+-+--++', '0--+-++--+ ']

1. L = 400 K = 300

No.of nodes in Rbest : 142

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.723333333333

pruned nodes , path or ID : []

1. L = 2000 K = 12

No.of nodes in Rbest : 55

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.761666666667

pruned nodes , path or ID : ['0+-+-+++', '0--+-----', '0++-+-', '0+-+--+++-+ ', '0----++-', '0+---++', '0-+---++', '0+--++', '0+-+-+-+-', '0-+-++--', '0+ +----++', '0--+-', '0+-+-++--', '0-+-+', '0+-+--+++', '0----+--', '0+--+-+ ---', '0+--+-+++', '0--+++--', '0++--++', '0+-+-++-', '0+---', '0+-+-+-+']

1. L = 2500 K = 175

No.of nodes in Rbest : 63

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.758333333333

pruned nodes , path or ID : ['0-+-++--', '0++-+-', '0--+--+++', '0----++-+', '0+-+--+', '0+--+-+---', '0+-+-+-+', '0-+---++', '0+----++-+', '0--+-++-', '0+-+++', '0--+-', '0--+++', '0+-++', '0-+-+', '0+---', '0+-+-++']

1. L = 700 K = 180

No.of nodes in Rbest : 90

No.of nodes in Root : 142

Accuracy of original tree : 0.723333333333

Accuracy of pruned tree : 0.741666666667

pruned nodes , path or ID : ['0+---+', '0-+-++-', '0+--+-+', '0+--++--+', '0--+++----+--', '0--+--', '0-+-+', '0+--++--', '0++-+', '0++--++-', '0+--++ ', '0-+---+', '0+-+-+-+-+-']

sample output , Tree , Pruned Tree

L = 1000 , K = 10

train\_set = 'data\_sets1/training\_set.csv'

valid\_set = 'data\_sets1/validation\_set.csv'

test\_set = 'data\_sets1/test\_set.csv'

No.of nodes in Rbest : 51

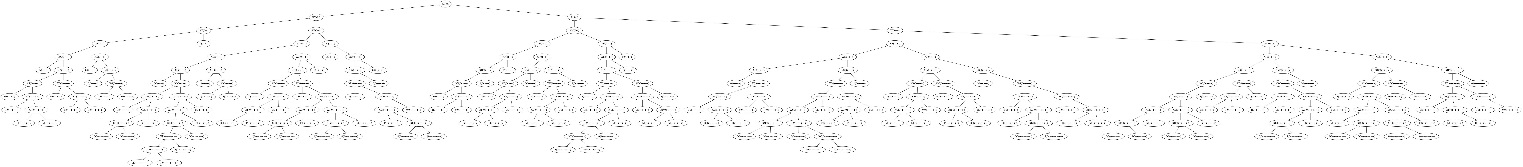
No.of nodes in Root : 137

Accuracy of original tree : 0.7585

Accuracy of pruned tree : 0.771

pruned nodes , path or ID : ['0+++-+-', '0-+---++-', '0++--+-', '0++++--+', '0+--++-+-', '0+-----+', '0-+-+-+', '0----+-++', '0+++++', '0+++---', '0++---+-', '0++++-', '0++++', '0+-+---', '0++--+', '0----++', '0-+---+', '0++-+-+', '0++---+', '0+-+--+-+', '0+--+-', '0++-+++-++', '0+-+--+', '0++-+', '0+--++-+']

Original Tree:



Pruned Tree:

