

Decision Tree

~ Build the classification model by implementing the decision tree ~

Professor: Sang-Wook Kim

Department of Computer Science

2012003716 Kwangil Cho

2017. 04. 20

0. Introduction

Decision tree learning uses a decision tree as a **predictive model** which maps **observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves)**. It is one of the predictive modelling approaches used in statistics, data mining and machine learning. Tree models where the target variable can take a finite set of values are called classification trees; in these tree structures, **leaves represent class labels and branches represent conjunctions of features that lead to those class labels**. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees.

1. Objectives and Environment

The goal of this project is to build the classification model with decision tree which has been learned by training data in advance. After building the decision tree, classify the test data to find out expected result.

I have been working on the following environment:

- Operating System: Ubuntu 14.05 LTS 64-bit
- Compilation: g++ compiler
- Language: C/C++ with C++ standard 11
- Source code editor: Vim
- Source code version control: Git
- Followed HYU coding convention faithfully
- In the source code, there is sufficient description of each function by comments.

2. Program Structures

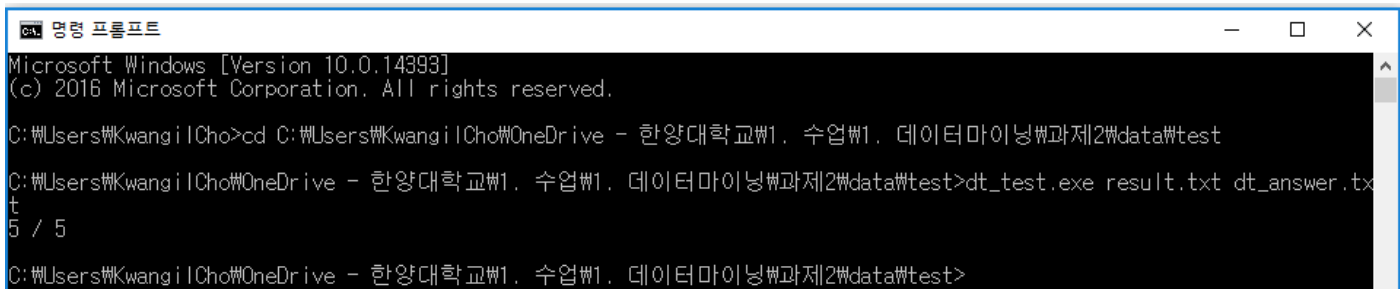
- (0) The program expects 4 arguments(executable file, training data file, test data file, result file) for building the decision tree. When the arguments are successfully ready, input handling routine works. First, handler opens input and output files. Second, input handler parses and saves the training data as global variable.
- (1) When the environment settings and input handling procedure is done, the tree is being built recursively by greedy manner. the following procedure to grow the decision tree.
 - A. The constructor of the Decision Tree receives three input arguments: training data for that node, attributes of the training data, and all the classes of each attribute.
 - B. I implemented the classification model using the information ratio. Therefore, the first thing the model needs is to get the information entropy of the node without doing anything. Information entropy is obtained by counting the number of target attributes of tuple of training data.
 - C. There are three cases when the decision tree stops growing. In the first case, there is no attribute to divide. In the second case, there is no sample data (tuple). In the last case, there is already made decision because all data has been classified into one class.
 - D. check the stop condition and if it continues to grow the tree, select the attribute which is the index of decision tree node.
 - E. Information gain is obtained for each attribute. And if information gain is used only, if there are many branches, distortion may occur. Therefore, information about split is obtained and the value is normalized.
 - F. Using information gain and split information, the tree can obtain the information ratio. By comparing the information ratios of each attribute, the largest attribute is set as the index attribute of the corresponding tree node.
 - G. When the corresponding attribute is set, the children of the node can be divided through the attribute's detailed classes. Each child recursively performs the above tree building process with the training data filtered by the detailed classes.
- (2) Once the decision tree building process is complete, a classification model is created and trained.
- (3) Open the test data file, read each tuple and search the decision tree for each data tuple. The result of searching tree (the leaf attribute value) is the classification result.
- (4) If there is no more test data to be dealt with, terminates the program successfully.

3. Execution

```
ki@ubuntu:~/DataMining/assignment2$ make
g++ -g -Wall -O3 -std=c++11 -I./include/ -o ./bin/decisiontree src/decisiontree.cc -L./lib/ -lpthread
ki@ubuntu:~/DataMining/assignment2$ ./bin/decisiontree ./data/dt_train.txt ./data/dt_test.txt result.txt
ki@ubuntu:~/DataMining/assignment2$
```

```
result.txt
1 |age      income  student credit_rating  Class:buys_computer
2 <=30    low      no      fair      no
3 <=30    medium  yes     fair      yes
4 31...40 low      no      fair      yes
5 >40     high     no      fair      yes
6 >40     low      yes     excellent no
```

- (1) Build and execute the program with the first given training & test set from TA.



```
Microsoft Windows [Version 10.0.14393]
(c) 2016 Microsoft Corporation. All rights reserved.

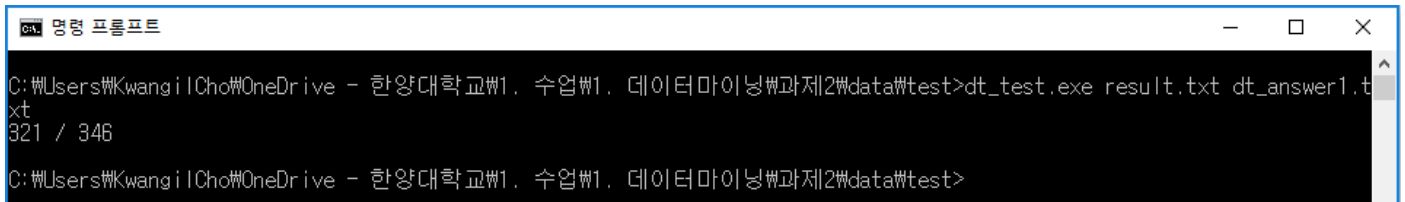
C:\Users\KwangilCho>cd C:\Users\KwangilCho\OneDrive - 한양대학교\1. 수업\1. 데이터마이닝\과제2\data\test
C:\Users\KwangilCho\OneDrive - 한양대학교\1. 수업\1. 데이터마이닝\과제2\data\test>dt_test.exe result.txt dt_answer.txt
5 / 5
C:\Users\KwangilCho\OneDrive - 한양대학교\1. 수업\1. 데이터마이닝\과제2\data\test>
```

- (2) Find the result by the comparison program. (5/5, 100%)

```
ki@ubuntu:~/DataMining/assignment2$ make
g++ -g -Wall -O3 -std=c++11 -I./include/ -o ./bin/decisiontree src/decisiontree.cc -L./lib/ -lpthread
ki@ubuntu:~/DataMining/assignment2$ ./bin/decisiontree ./data/dt_train1.txt ./data/dt_test1.txt result.txt
ki@ubuntu:~/DataMining/assignment2$
```

```
result.txt
1 buying  maint  doors  persons lug_boot  safety  car_evaluation
2 med     vhigh  2      4      med     med     acc
3 low     high   4      4      small  low     unacc
4 high    vhigh  4      4      med     med     acc
5 high    vhigh  4      more   big     low     unacc
6 low     high   3      more   med     low     unacc
7 med     high   2      more   small  high    acc
8 vhigh   low    3      2      med     high    unacc
9 med     high   2      4      small  low     unacc
10 med    low    5more  4      small  med     good
```

- (3) Build and execute the program with the second given training & test set from TA.



```
C:\Users\Kwangil\OneDrive - 한양대학교\1. 수업\1. 데이터마이닝\과제2\data\test>dt_test.exe result.txt dt_answer1.txt
321 / 346
C:\Users\Kwangil\OneDrive - 한양대학교\1. 수업\1. 데이터마이닝\과제2\data\test>
```

(4) Find the result by the comparison program. (321/346, 92.7%)

4. Epilogue

Through this project, I was able to learn how to build the decision tree which is the classification model. The process of training the decision tree with training data and improving the correctness of the tree were a big food for thought. Thanks to professor and assistant who prepared for class and practices.

5. Reference

- [Wikipedia:Decision Tree](https://en.wikipedia.org/wiki/Decision_tree)(https://en.wikipedia.org/wiki/Decision_tree)