EDA132

Assignment 2: Machine Learning

Induction of Decision Trees

Hannah Lindblad
elt13hli@student.lu.se

Therese Magnusson
dat13tma@student.lu.se

February 17, 2017

1 Induction of Decision Trees

In this assignment we have implemented the decision tree learning algorithm with an example set from Russell-Norvig, Section 18.3. The example set is for a decision regarding whether it's worthwhile to wait for a table at a restaurant or not. There are ten different input variables (attributes) to take into consideration for the decision.

2 Selected Improvement

We have implemented the suggested improvement of a generalized importance function. This means that the function can handle any number of values for the classification attribute.

3 Our Implementation

The program output is a decision tree in the format specified by the assignment description (see section 4 for a printout of the output). The algorithm uses an importance function which utilizes entropy to determine the next attribute for the algorithm to branch on. This in accordance with section 18.4 in Russel Norvig.

The data sets are parsed from a file where the data is stored in the ARFF-format. We have chosen Python 3 for the implementation and the program uses the data structures *OrderedDict* and *list* for storing the data sets. The program uses the regex package while parsing and the log2 function from the math package in the calculation of the entropy.

3.1 User guide

The program and input files used can be found at `/h/dk/v/dat13tma/AI/Assign2`. The program is named `decisionTree.py` and can be run in the terminal by standing in the correct directory and run:

```
python decisionTree.py
```

The program will then print one version for each of the two input files. The first has only two alternatives for the classification while the other has three.

4 Results

The example set used by the program can be seen in appendix A. The output from the program with that example set can be seen in figure 1.

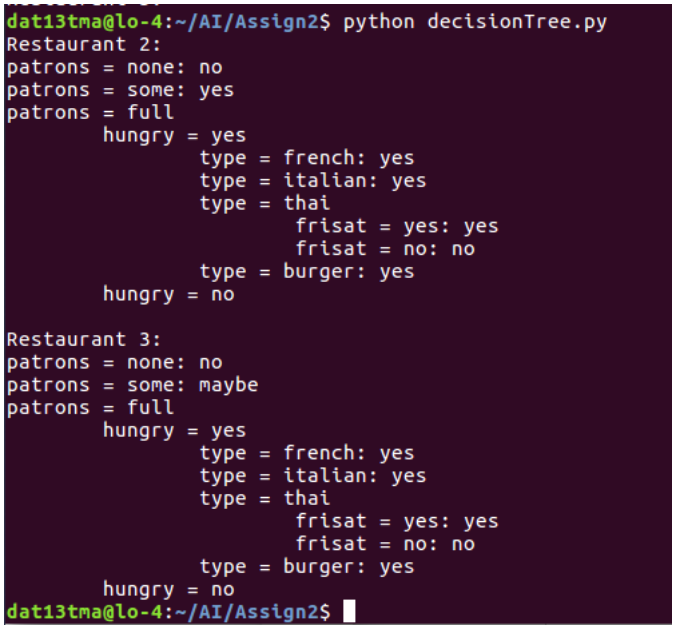

Figure 1: Program output

4.1 Comments on the results

It was hard to judge the validity of the importance function for more than two classifications. The test we performed was to change the original restaurant example set in Russell-Norvig to entail three classifications instead of two. We changed all examples with the attribute-result combination **Patrons : some** to be classified as the new, third classification. In this way the algorithm should return the same result as for the original example set (presented in Russell-Norvig in figure 18.6) except for a different classification of **Patrons : some**. This was also the output from the program in section 4.

References

Russell, S. & Norvig, P. (2010). Artificial Intelligence: A Modern Approach (3rd ed.). Upper Saddle River, New Jersey: Pearson Education, Inc.

Appendices

A Example Set

```
% 1. Title: Restaurant Domain

@RELATION restaurant

@ATTRIBUTE Alternate {yes,no}
@ATTRIBUTE Bar {yes,no}
@ATTRIBUTE FriSat {yes,no}
@ATTRIBUTE Hungry {yes,no}
@ATTRIBUTE Patrons {none,some,full}
@ATTRIBUTE Price {$,$$,$$$}
@ATTRIBUTE Raining {yes,no}
@ATTRIBUTE Reservation {yes,no}
@ATTRIBUTE Type {french,italian,thai,burger}
@ATTRIBUTE WaitEstimate {0-10,10-30,30-60,>60}
@ATTRIBUTE WillWait {yes,no,maybe}

@DATA
yes,no,no,yes,some,$$$,no,yes,french,0-10,maybe
yes,no,no,yes,full,$,no,no,thai,30-60,no
no,yes,no,no,some,$,no,no,burger,0-10,maybe
yes,no,yes,yes,full,$,yes,no,thai,10-30,yes
yes,no,yes,no,full,$$$,no,yes,french,>60,no
no,yes,no,yes,some,$$,yes,yes,italian,0-10,maybe
no,yes,no,no,none,$,yes,no,burger,0-10,no
no,no,no,yes,some,$,yes,yes,thai,0-10,maybe
no,yes,yes,no,full,$,yes,no,burger,>60,no
yes,yes,yes,yes,full,$$$,no,yes,itailan,10-30,no
no,no,no,no,none,$,no,no,thai,0-10,no
yes,yes,yes,yes,full,$,no,no,burger,30-60,yes
```