Decision Trees - HW1

Karthik Ravindra Rao

February 6, 2017

1 ID3

1.1 Implementation

Following functions are written in the file titled HW1_ID3ṗy

entropy($attribute, classification\_attribute$) : function to calculate entropy of a given attribute. Classification attribute is the attribute for which prediction is being made.

seperateBranchesOfAttribute($example, attribute$) : function which returns dataframes seperated by the mentioned attribute.

prepareData() : prepares the data in the desired format. If outlook can take 3 different values, it is converted into numericals in alphabetical order

findRootNode($target, labels$) : if table and labels are passed, the attribute with the highest information gain is returned.

id3($examples, classification_attribute, attributes$) : main algorithm which constructs decision tree.

class **DecisionTree**(*object*) : Data-structure of a decision tree

```
def __init__(self):
self.label = None
self.children = [ ]
```

printTree($tree$) : Prints tree using level order traversal

The following output is obtained for the given data. This output, is depicted in form of a chart in figure 1. The attribute values which the branches represent can be obtained by un-commenting lines 219-224. These lines were commented to generate output as desired by the assignment.

```
Occupied

Location Size Location

yes yes no yes no no Price no yes yes Size yes Size

no no yes yes no yes no yes yes
```

1.2 Prediction

Tree in Figure 1 was obtained for the given data.
It predicts **'YES'** for the below mentioned test case, which can be traced out using Figure 1.

```
(size = Large; occupied = Moderate; price = Cheap; music = Loud; location = City-Center;
VIP = No; favorite beer = No).
```

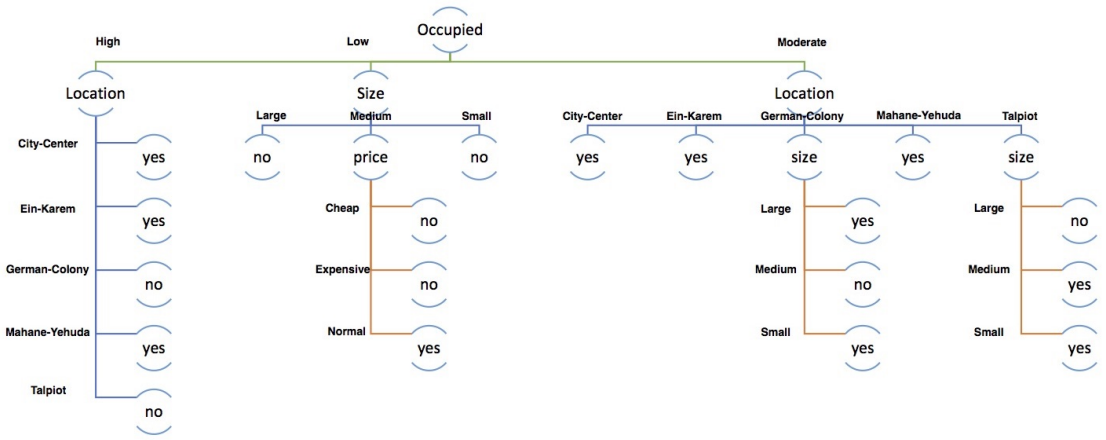

Figure 1: Decision tree obtained for good night-out in Jerusalem for the New Year's Eve. data

2 Software Familiarization

2.1 Scikit-learn

Decision tree using scikit is implemented in file HW1.py

Data is initially pre-processed using pandas and label pre-processor. Then, using decision tree classifier the decision tree is implemented by passing the dataframe and the classifying attribute, which can then used to predict for new values

```
clf = tree.DecisionTreeClassifier(criterion = "entropy")
clf = clf.fit(target, lables)
print(clf.predict([[0,2,0,0,0,0,0]]))
```

2.2 Improvements for Implemented Algorithm

- For this algorithm to be used as function, it should be capable of accepting any ċsv file and make predictions without hard-coded pre-processing that is currently employed
- A predict function needs to be written where test data can be predicted without manual-tracing which is currently employed
- This algorithm cannot process continuous values or missing values. C4.5 - an improved version of ID3, performs a better job in this regard

3 Applications of Decision Trees

3.1 Predicting Library Book Use

[AKA] Decision trees can be used to arrange books in easily accessible areas and not so easily accesiable areas. In an experiment, book usage records for 80,000 titles during the period July 1975 to June 1984 were used to prepare examples classified using six attributes - namely, checkout history, last use, publication date,language, country, and alphabetic prefix of the Library of Congress classification. They were labeled using classes based on how often they are checked-out in "future". The quality of a given choice policy is evaluated using a measure called the "expected advantage over random" (EAR) was found to be 73.12% in this case.

3.2 Characterization of Leiomyomatous Tumors

[AKA] Decision trees can be used to determine the group of a tumor based on features. Decaestecker applied decision tree learning to the difficult problem of leiomyomatous tumor diagnosis. Decision

tree was constructed using 23 cases, each evaluated by three different pathologists. Decaestecker found that Decision trees are better than logistic regression and neural networks for the studied task

3.3 Star/Cosmic-Ray Classification in Hubble Space Telescope Images

[AKA]Decision trees are used to differentiate between cosmic rays and stars in the images collected by the Hubble telescope. In the experiments conducted by Salzberg, a set of 2211 pre-classified images were used as training sample for decision tree construction. A seperate 2282 pre-classified images were used to test the performance. An accuracy of 95% was achieved.

References

[AKA] Hussein Almuallim, Shigeo Kaneda, and Yasuhiro Akiba. Development and application of decision trees.