Andrew Floyd

CS3001: Intro to Data Science

Dr. Fu

October 21st, 2018

HW #3

Github project link: https://github.com/afzm4/cs3001hw3

Task 1: *Describe the difference between classification and clustering?*

The main difference between classification and clustering is that classification is supervised, meaning that the training data instances and their features are labeled indicating the class. In general, with classification you have predefined classes and you are trying to figure out which class a new object belongs to. With clustering, one tries to group a set of objects and find whether or not there is some sort of relationship between the objects (the class labels of training data is unknown). Basically, (with clustering) given a set of attributes, the aim is to establish the existence of classes (or clusters).

Task 2: *Describe what is entropy?*

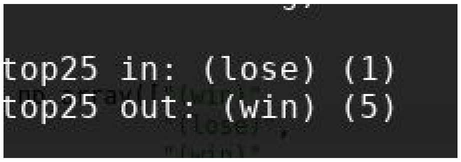
In data mining applications, entropy is used to calculate the homogeneity of a dataset (measure the level of impurity in a group). The more homogeneous the sample is, the lower the entropy is and vice-versa (meaning that if the sample is equally divided it has a entropy of one). This means that if a dataset is completely homogeneous, then the entropy would be zero. Information gain is based on the decrease in entropy after a dataset is split.

Task 3: *Describe and compare the following "feature selection measures" or called "splitting criteria": information gain and Gini index?*

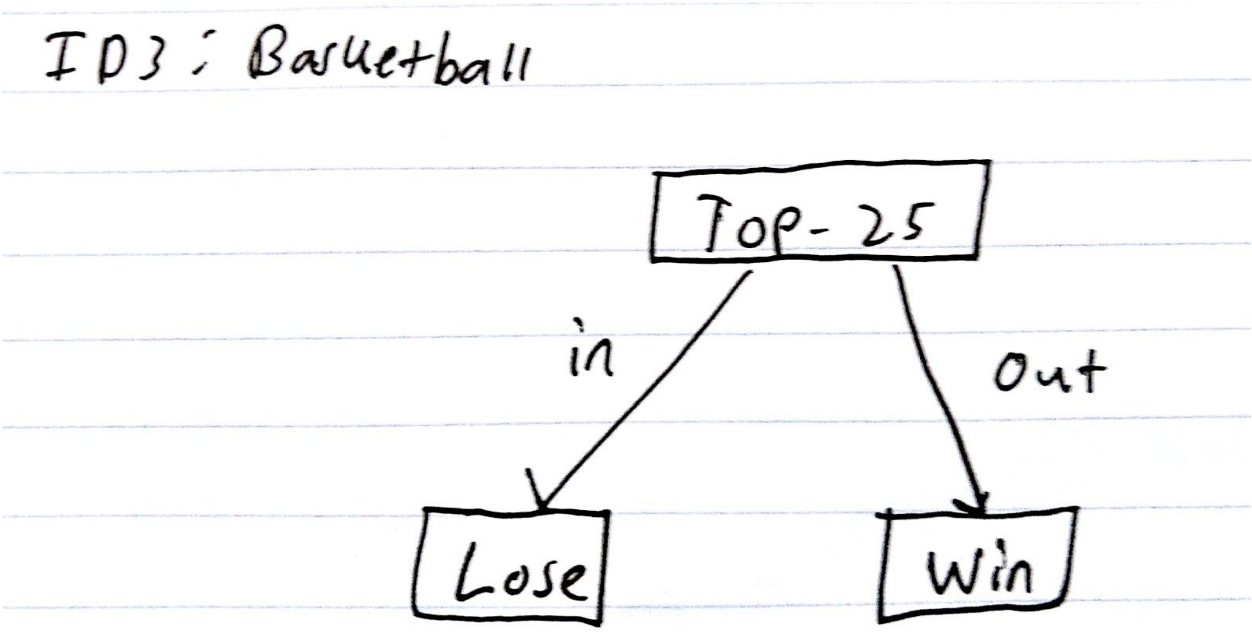
Information gain tells us how important a given attribute is. It can be used to decide the ordering of attributes in the nodes of a decision tree as well. It is calculated by subtracting the average entropy of the children from the entropy of the parent node, or by subtracting any of the three FILL IN.. Meanwhile, the Gini index is another measure of impurity which is lowest (zero) when all records belong to one class and largest (0.50) when records are equally distributed among all classes. This implies that most interesting information is when the gini index is low and that the information is less interesting as the index goes higher. They are both items that can be used to implement decision trees as information gain is used by ID3, C4.5 and C5.0 trees, while gini index is used by CART trees.

Task 4:

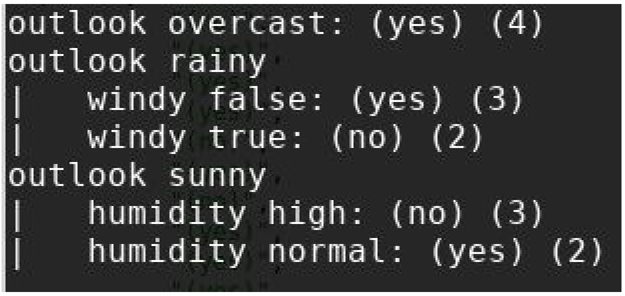
- ID3 Trees:
    - Basketball:



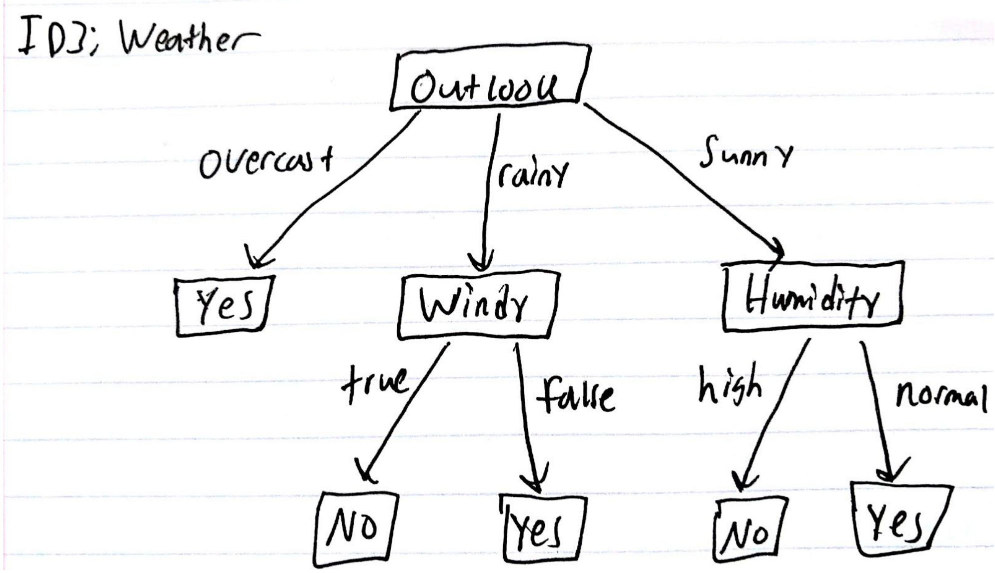
```
top25 in: (lose) (1)
top25 out: (win) (5)
```



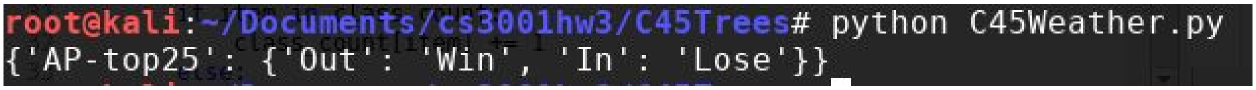
    - Weather:



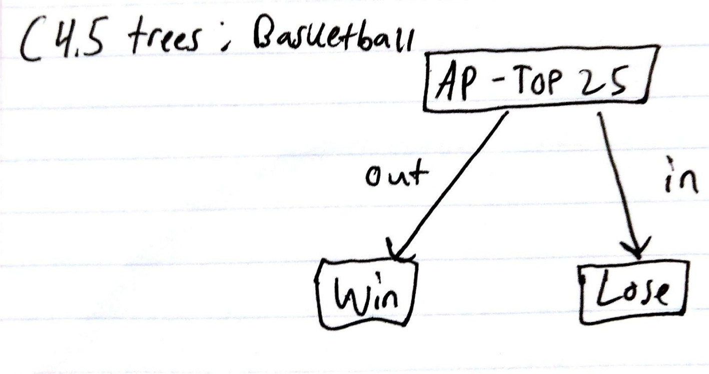
```
outlook overcast: (yes) (4)
outlook rainy
|   windy false: (yes) (3)
|   windy true: (no) (2)
outlook sunny
|   humidity high: (no) (3)
|   humidity normal: (yes) (2)
```



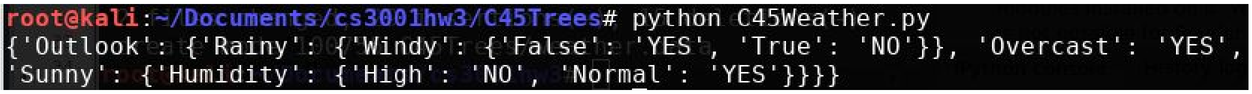
- C4.5 Trees
    - Basketball:



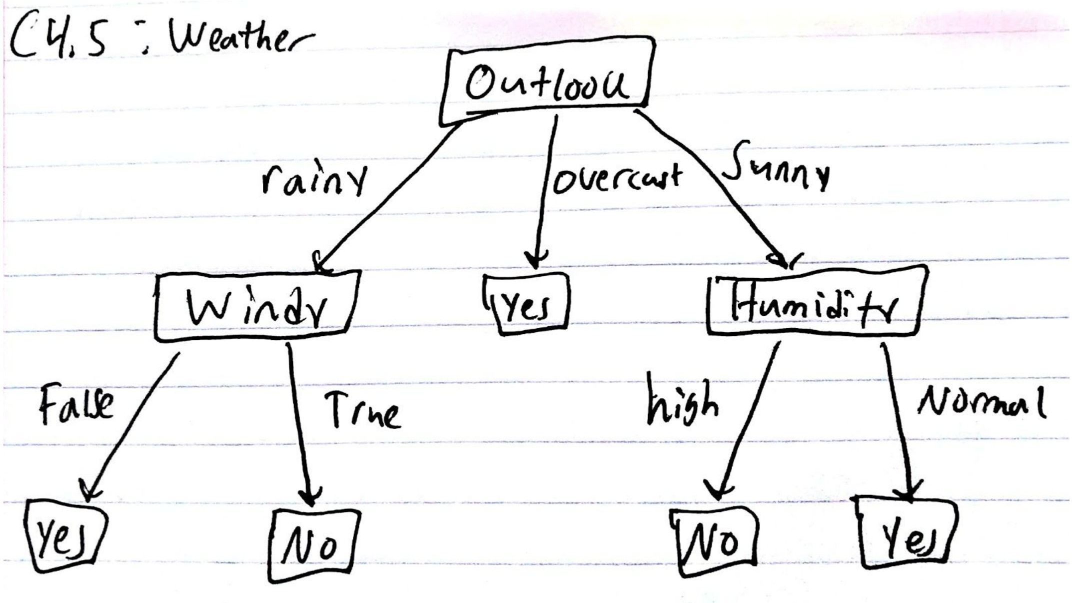
```
root@kali:~/Documents/cs3001hw3/C45Trees# python C45Weather.py
{'AP-top25': {'Out': 'Win', 'In': 'Lose'}}
```



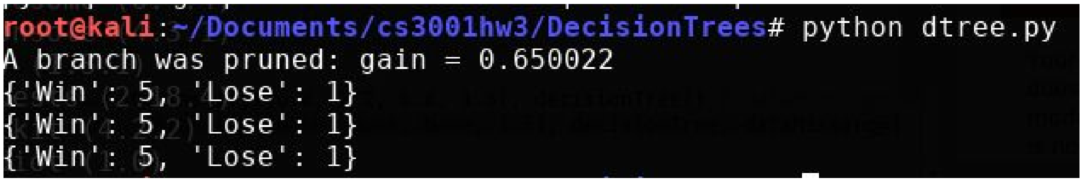
- Weather:



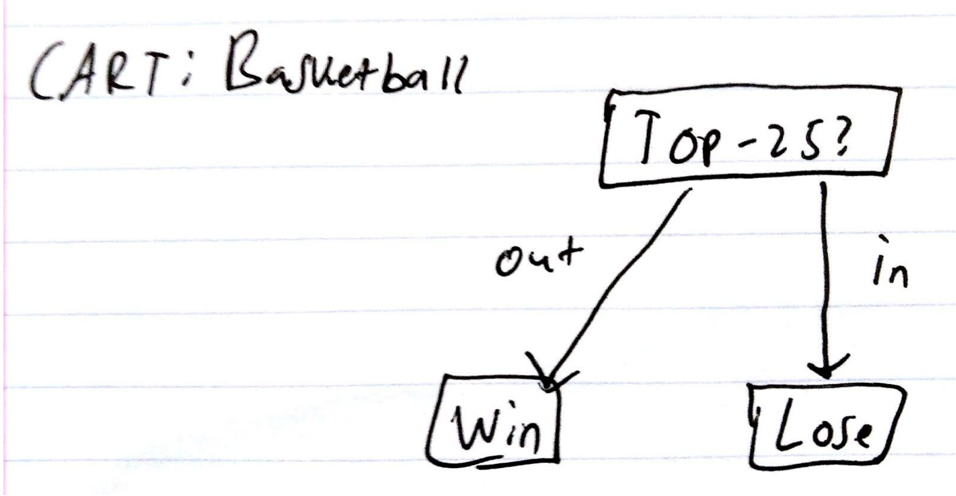
```
root@kali:~/Documents/cs3001hw3/C45Trees# python C45Weather.py
{'Outlook': {'Rainy': {'Windy': {'False': 'YES', 'True': 'NO'}}, 'Overcast': 'YES',
'Sunny': {'Humidity': {'High': 'NO', 'Normal': 'YES'}}}}
```



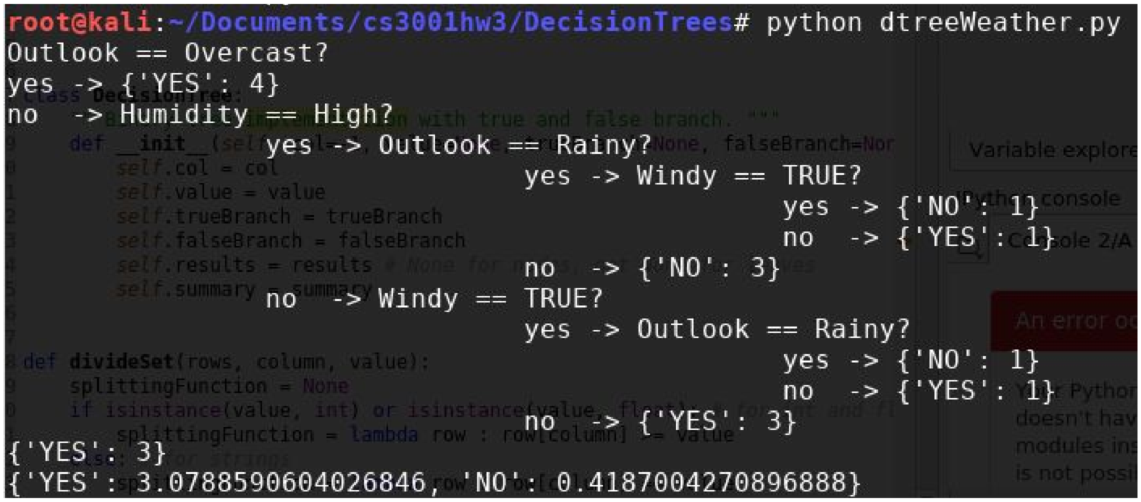
- CART Trees
  - Basketball:



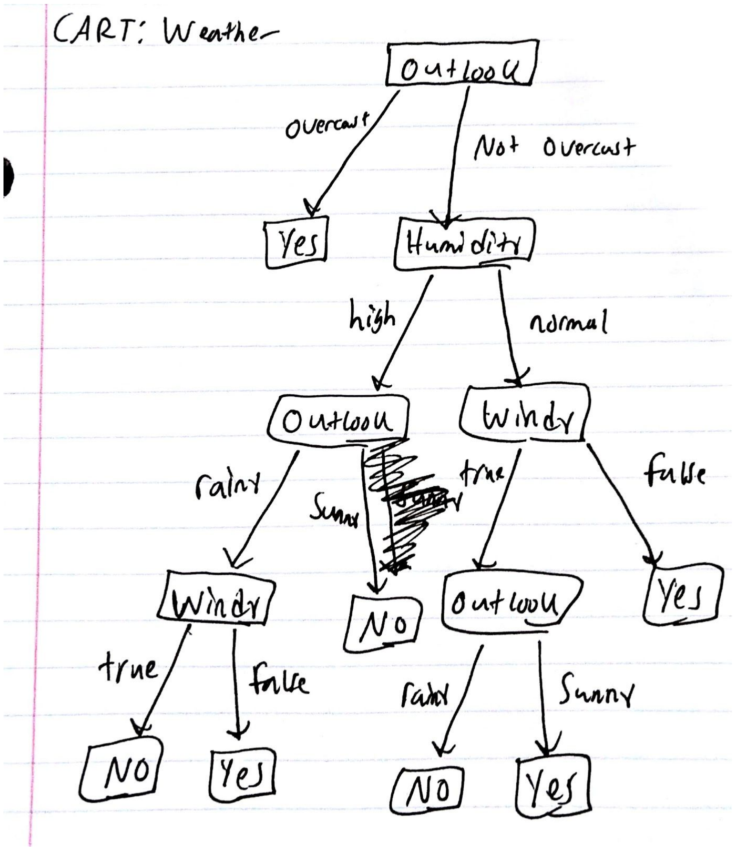
```
root@kali:~/Documents/cs3001hw3/DecisionTrees# python dtree.py
A branch was pruned: gain = 0.650022
{'Win': 5, 'Lose': 1}
{'Win': 5, 'Lose': 1}
{'Win': 5, 'Lose': 1}
```



- Weather:



```
root@kali:~/Documents/cs3001hw3/DecisionTrees# python dtreeWeather.py
Outlook == Overcast?
yes -> {'YES': 4}
no  -> Humidity == High?
                yes -> Outlook == Rainy?
                                yes -> Windy == TRUE?
                                                yes -> {'NO': 1}
                                                no  -> {'YES': 1}
                                no  -> {'NO': 3}
                no  -> Windy == TRUE?
                                yes -> Outlook == Rainy?
                                                yes -> {'NO': 1}
                                                no  -> {'YES': 1}
                                no  -> {'YES': 3}
{'YES': 3}
{'YES': 3.0788590604026846, 'NO': 0.4187004270896888}
```


CART: Weather
Outlook
overcast
Not Overcast
Yes
Humidity
high
normal
Outlook
Windy
rainy
Sunny
true
false
Windy
No
Outlook
Yes
true
false
rainy
Sunny
No
Yes
No
Yes