

Decision tree

Using python.

Iván Andrés Trujillo Abella

Facultad de Ingenieria
Pontificia Universidad Javeriana

`trujilloiv@javeriana.edu.co`



Supervised Learning Machine

In this case the algorithm will learn about the real label.



Decision tree

What is a np-hard problem?
Gini impurity.



The worse situation is when in a leaf each class have the same probability.
due of there are not domination.

We need choose the better variable of the data set as the root node.

Which is the better variable



The idea is reduce the entropy in all tree, then if there are k classes we have

$$H(T) = \sum_{i=1}^k p_i H(i) \quad (1)$$

train and test data.

information gain:

why entropy is convex? it is a theorem?



insights about question

For instance when one person need guess a object to another person, need ask more informative question eliminating a large of options for instance it is animal?, is more informative that ask for color.



overfitting



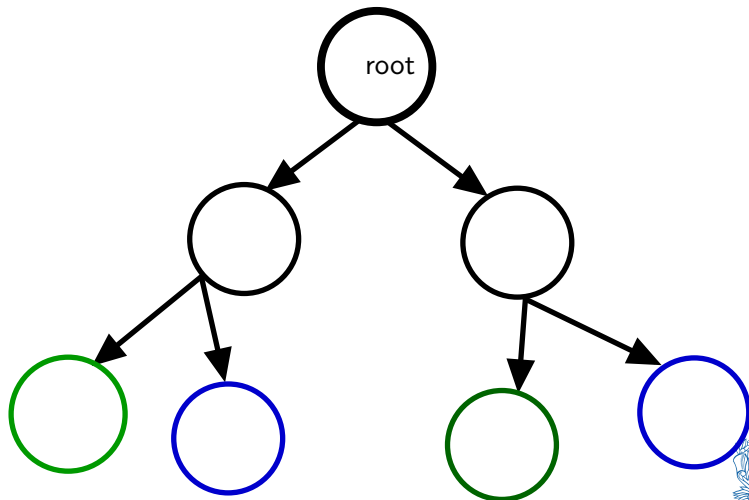
DT is very interpretable.



bagging



representation



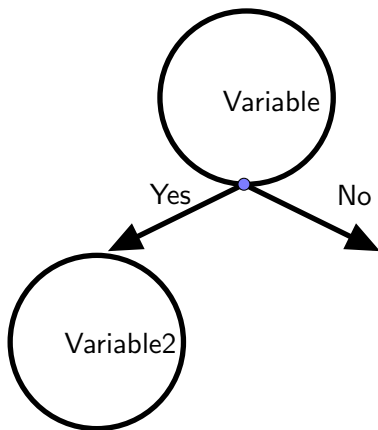
Nodes, leaves

Each node is a variable and the links are the possible values that could be take the variable.

The green and blue circles are the leaves of the tree, and inside of them are the class to predict



Nodes, Links



The node **variable** have two possible values **Yes** and **No**.

Pseudo code

```
Choose the better variable
split the data according the attributes
of the better variable,
for each instance apply the same process recursively.
```



Sector	Income	Size	Bankruptcy
Financial	High	Medium	No
Financial	High	Medium	No
Financial	Low	Small	Yes
Agricultural	Low	Small	No
Agricultural	Low	Medium	No
Agricultural	High	Small	Yes
Agricultural	High	Small	Yes

Table: Complete data set



Split data by Sector

sector	income	size	bankruptcy
Financial	High	Medium	No
Financial	High	Medium	No
Financial	Low	Small	Yes

Table: Financial

sector	income	size	bankruptcy
Agricultural	Low	Small	No
Agricultural	Low	Medium	No
Agricultural	High	Small	Yes
Agricultural	High	Small	Yes

Table: Agricultural



Process

Until now we suppose that the better variable of the data set is **sector** after, by each instance suppose that we find the better variable assuming for instance that for *financial* is *income* and for *agriculture* is *size*.



Stopping criteria

income	size	bankruptcy
Low	Small	Yes
High	Small	Yes
High	Small	Yes

Table:

The algorithm will stop when the label is the same for all rows, then return a leaf with the attribute of class.



Stopping criteria

Income	Size	Bankruptcy
High	Medium	Yes
High	Medium	No
High	Medium	Yes

Table:

The algorithm will stop when attributes are the same for all variables, then return a leaf with the most common.



How select the better variable?

Entropy

The better variable will be those that is able to discriminate among the classes, for instance to select a variable and split all belong to the same class.

This lead to homogeneity concept: for instance we select **Sector** and split **financial** and **agricultural** and of N patterns

$$\left[\begin{array}{ll} \text{Financial} & \text{Yes} = \frac{N}{2} - 4 \\ & \text{No} = 4 \\ \text{Agricultural} & \text{Yes} = 4 \\ & \text{No} = \frac{N}{2} - 4 \end{array} \right]$$

this mean to split the data in *Financial* there are $\frac{N}{2} - 4$ rows with yes and 4 with No.



Bad quality

A variable with bad quality not let us discriminate and therefore the proportion could be equally in each class.



Entropy

Bankruptcy

Yes

No

Yes

Yes

No

No

$$Entropy(Bankruptcy) = -\frac{3}{6} \log_2 \frac{3}{6} - \frac{3}{6} \log_2 \frac{3}{6}$$

Note here that is the probability of occur yes = $\frac{3}{6}$.

Generally $Entropy(var) = -\sum_i^{classes} P_i \log_2 P_i$.



Gain of information



if-else

Root is the first node, the circle with colors are also known as leaves.
Literally, each node is a question, *the x value is greater or not than k ?*



Predict a new observation

To predict a new pattern, the tree only follow the path relative to the attributes that have the new data.



Python

```
from sklearn import
```

