Decision tree

Using python.

Iván Andrés Trujilllo Abella

Facultad de Ingenieria
Pontificia Universidad Javeriana

trujilloiv@javeriana.edu.co

Supervised Machine learning

Definition

In this learning method exist the real label or target variable.

Example

Financial historical data that contain financial statements of healthy and bankrupt firms, we could develop an algorithm to predict of one year to another the insolvency state.

Medical Diagnosis!

...

- Set of questions
- Set of tests

$$\max P(Disease \mid Symptoms) \tag{1}$$

Reduce the risk of miss-classification...

Diagnosis quality...

The number of individuals classified correctly...

Medical Diagnosis!

...

The questions and tests not were randomly applied, unlike the order is important...

- The first question or test is the most important this allow us reduce the problem...
- According to the result or answer of the first test, the second or question, is the most informative for an correct diagnosis...
- Is repeated the same logic until the professional have the enough information to take a decision...

Assesment

...

For a total of N individuals where x have disease and $N - x$ are healthy assuming that the questions and the test are **informative** then the order work well if **after** you can

$$\hat{P}(Disease) = \frac{x}{N} \tag{2}$$

$$\hat{P}(Healthy) = \frac{N - x}{N} \tag{3}$$

Decision tree

...

The Decision tree make split (question) according to the data searching the major precision of the probability of classify correctly.

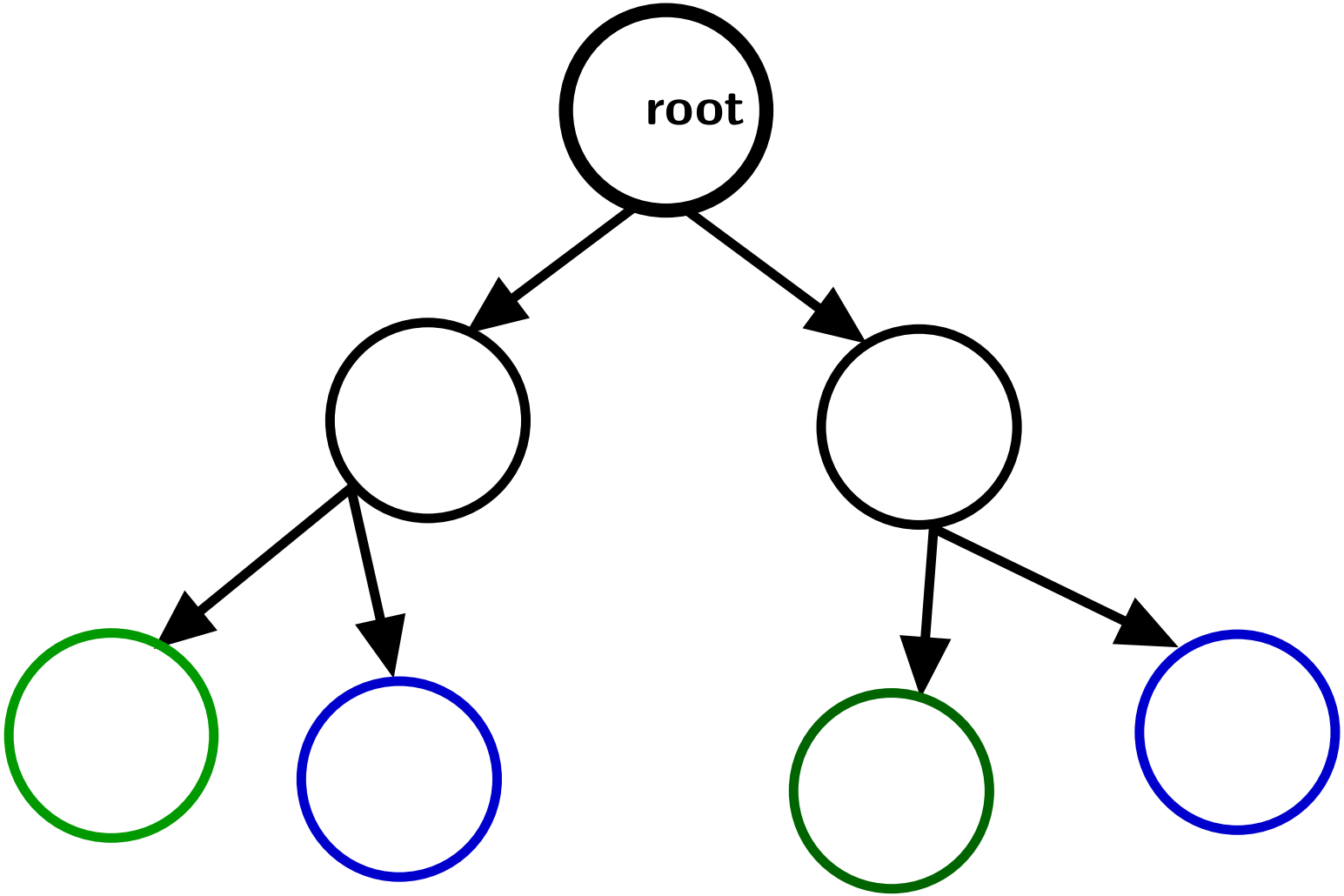
...

Using measures of inequality asses if each question or split preserve an correct classification.

$$H(T) = \sum_{i=1}^{k} p_i H(i) \tag{4}$$

This expression measure the work of the tree, k is the number of classes.

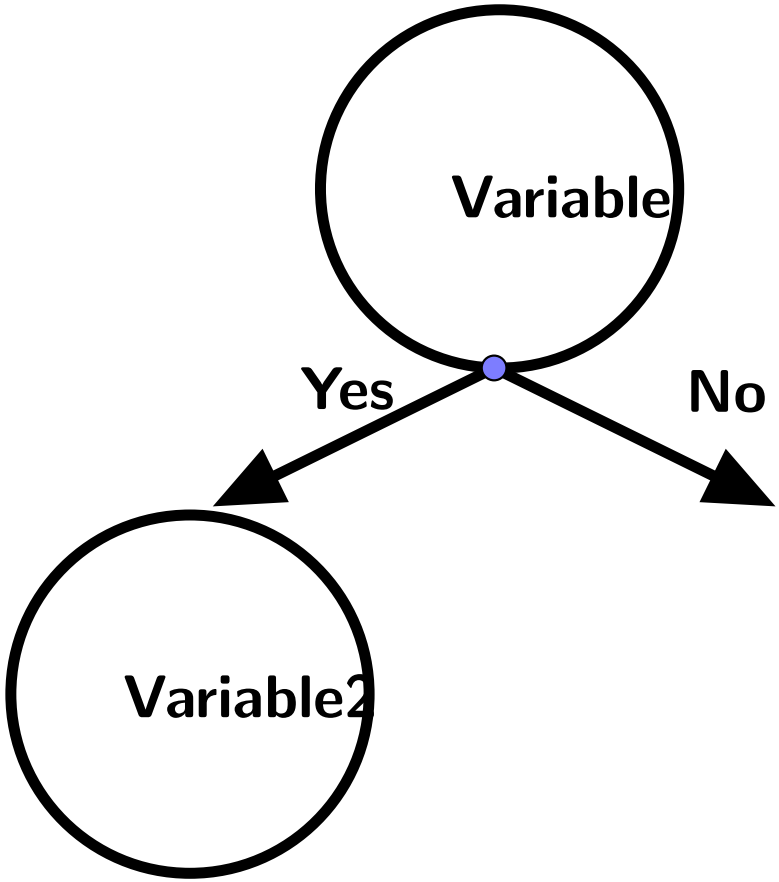
representation



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 withe 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

$$\begin{bmatrix} Financial & Yes = \frac{N}{2} - 4 \\ & No = 4 \\ Agricultural & Yes = 4 \\ & No = \frac{N}{2} - 4 \end{bmatrix}$$

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

Information gain

for instance N is the number of observations in dataset

cat	Bankruptcy
cat1	Yes
cat1	No
cat1	Yes
cat1	Yes
cat2	No
cat2	No
cat2	Yes
cat2	Yes

Table:

Information gain

Assume that a variable have *cat*1, *cat*2 and each one could produce the split of bankruptcy we need take a average of homogeneity:

cat1	**cat2**
Bankruptcy	**Bankruptcy**
Yes	No
No	No
Yes	Yes
Yes	Yes

Table:

Information gain

According to the above table $IG(Bankrupcy)$ is therefore $Entropy(Bankruptcy) - (\frac{n(cat_1)}{N} Entropy(bankruptcy|cat1) + \frac{n(cat_2)}{N} Entropy(bankruptcy|cat2))$ where $n(cat_i)$ is the number of rows in $i - th$ class.

Predict a new observation

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