Esercitazione 2 16/10/2023

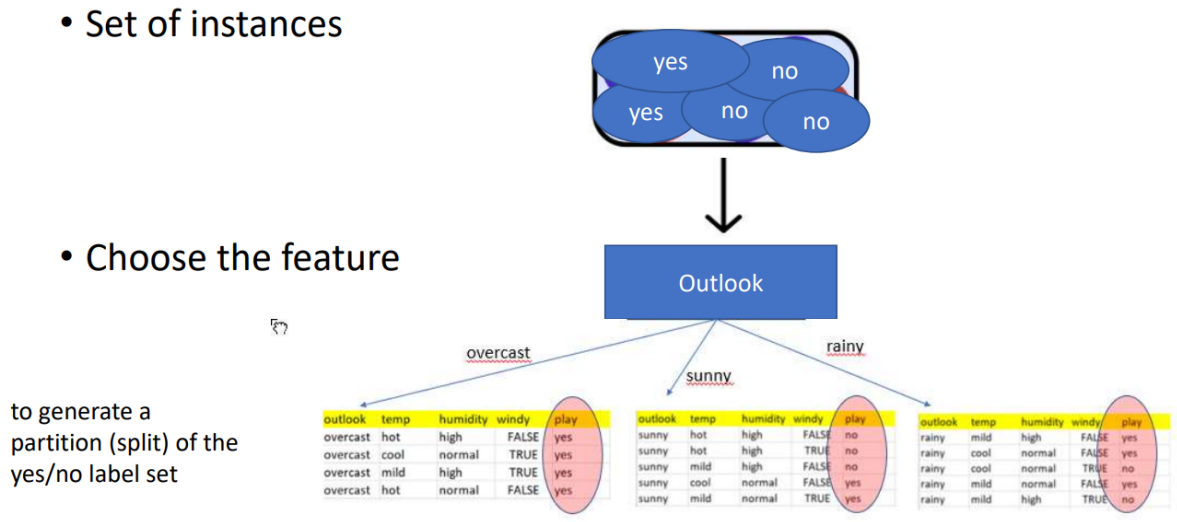
Seconda parte dell'esercitazione, prima parte su carta

L'albero di decisione è un albero con una radice e dei percorsi. Si scendono i rami composti da nodi arrivando ai nodi che sono foglie dell'albero.

Come fare crescere un albero di decisione

(simil ID3) steps qua sotto tradotti un po' male

- Inizia con un albero vuoto
- Scegli la feature da usare per dividere i dati
- Per ogni separazione:
  - Se non c'è altro da fare, predici con il nodo foglia
  - Altrimenti, vai allo step 2 e effettua la ricorsione



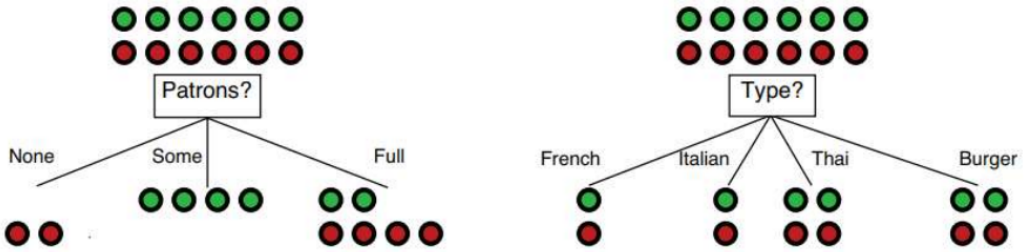
Nel primo caso: 4 si, 2 no. Nel secondo caso: 2 si, 2 no. Nel terzo caso: 3 si, 2 no.

Il primo caso è quindi il più puro, il secondo invece non lo è per niente perché il target divide il set a metà.

Quindi dove vado a splittare? (considerando anche che qui è scelto Outlook senza motivo) Sono contento di avere overcast come foglia dell'albero perché se Outlook assume il valore overcast posso dire sempre di si, mentre invece nel caso di sunny non saprei come predire.

Bisogna quindi capire qual è il modo migliore di fare splitting.

Idea: if we can test only one more attribute, which one results in the smallest empirical loss on examples that come through this branch?



Assuming we pick the majority value in each new leaf,

$0+0+2=2$ for $Patrons$ $1+1+2+2=6$ for $Type$

Gini Index

Simile ad information gain

The Gini index for a given set of instances is calculated as:

$$Gini(D) = 1 - \sum_{i=1}^{c} p_i^2$$

Where:

- D is the set of instances.
- c is the number of classes (2 in our case: "Fruit" and "Non-Fruit").
- p_i is the proportion of instances in class i within set D.

0 : purity,
1 : no purity (the elements are evenly distributed among different classes).

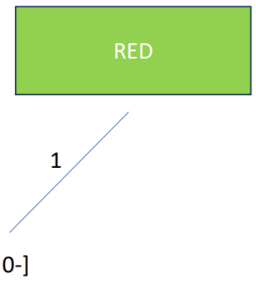
calcolo del gini index dell'intero dataset

Example	Is Red	Is Round	Is Fruit
1	1 (Yes)	1 (Yes)	1 (Fruit)
2	1 (Yes)	0 (No)	1 (Fruit)
3	0 (No)	1 (Yes)	1 (Fruit)
4	0 (No)	0 (No)	0 (Non-Fruit)
5	1 (Yes)	0 (No)	1 (Fruit)

Number of fruit samples (class 1) = 4
Number of non-fruit samples (class 0) = 1

$$Gini(D) = 1 - [(4/5)^2 + (1/5)^2] = 1 - [16/25 + 1/25] = 1 - 17/25 = 8/25$$

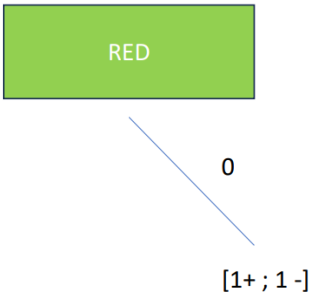
Gini Index of "Is Red=1":



1. Subset 1: "Is Red" = 1 (Yes)
   - Number of fruit samples (class 1) = 3
   - Number of non-fruit samples (class 0) = 0 (No non-fruit samples in this subset)
   - Gini Index for Subset 1 (Is Red = 1):
     $Gini(D_1) = 1 - [(3/3)^2 + (0/3)^2] = 1 - [1 + 0] = 0$

Example	Is Red	Is Round	Is Fruit
1	1 (Yes)	1 (Yes)	1 (Fruit)
2	1 (Yes)	0 (No)	1 (Fruit)
3	0 (No)	1 (Yes)	1 (Fruit)
4	0 (No)	0 (No)	0 (Non-Fruit)
5	1 (Yes)	0 (No)	1 (Fruit)

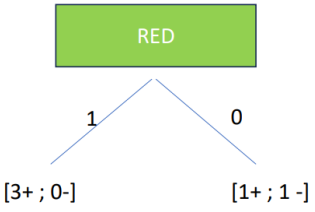
Gini Index of "Is Red=0":



2. Subset 2: "Is Red" = 0 (No)
   - Number of fruit samples (class 1) = 1
   - Number of non-fruit samples (class 0) = 1
   - Gini Index for Subset 2 (Is Red = 0):
     $Gini(D_2) = 1 - [(1/2)^2 + (1/2)^2] = 1 - [1/4 + 1/4] = 1/2$

Example	Is Red	Is Round	Is Fruit
1	1 (Yes)	1 (Yes)	1 (Fruit)
2	1 (Yes)	0 (No)	1 (Fruit)
3	0 (No)	1 (Yes)	1 (Fruit)
4	0 (No)	0 (No)	0 (Non-Fruit)
5	1 (Yes)	0 (No)	1 (Fruit)

Evaluate the split of «Is Red» USING Gini



1. Subset 1: "Is Red" = 1 (Yes)
   - Number of fruit samples (class 1) = 3
   - Number of non-fruit samples (class 0) = 0 (No non-fruit samples in this subset)
   - Gini Index for Subset 1 (Is Red = 1):
     $Gini(D_1) = 1 - [(3/3)^2 + (0/3)^2] = 1 - [1 + 0] = 0$

2. Subset 2: "Is Red" = 0 (No)
   - Number of fruit samples (class 1) = 1
   - Number of non-fruit samples (class 0) = 1
   - Gini Index for Subset 2 (Is Red = 0):
     $Gini(D_2) = 1 - [(1/2)^2 + (1/2)^2] = 1 - [1/4 + 1/4] = 1/2$

Example	Is Red	Is Round	Is Fruit
1	1 (Yes)	1 (Yes)	1 (Fruit)
2	1 (Yes)	0 (No)	1 (Fruit)
3	0 (No)	1 (Yes)	1 (Fruit)
4	0 (No)	0 (No)	0 (Non-Fruit)
5	1 (Yes)	0 (No)	1 (Fruit)

"Is Red = 1" provides pure separation between fruit and non-fruit samples) for the root node (attribute IsRed) using Gini index. In this case: Gini Index of 0

Evaluate the split of «Is Red» USING Gini

1. Subset 1: "Is Round" = 1 (Yes)
   - Number of fruit samples (class 1) = 2
   - Number of non-fruit samples (class 0) = 0 (No non-fruit samples in this subset)
   - Gini Index for Subset 1 (Is Round = 1):
     $Gini(D_1) = 1 - [(2/2)^2 + (0/2)^2] = 0$
2. Subset 2: "Is Round" = 0 (No)
   - Number of fruit samples (class 1) = 2
   - Number of non-fruit samples (class 0) = 1
   - Gini Index for Subset 2 (Is Round = 0):
     $Gini(D_2) = 1 - [(2/3)^2 + (1/3)^2] = 1 - [4/9 + 1/9] = 4/9$

Example	Is Red	Is Round	Is Fruit
1	1 (Yes)	1 (Yes)	1 (Fruit)
2	1 (Yes)	0 (No)	1 (Fruit)
3	0 (No)	1 (Yes)	1 (Fruit)
4	0 (No)	0 (No)	0 (Non-Fruit)
5	1 (Yes)	0 (No)	1 (Fruit)

Now, let's compare the Gini Index for the "Is Round" and "Is Red" splits:

- Gini Index for "Is Round" split (Is Round = 1): 0
- Gini Index for "Is Round" split (Is Round = 0): 4/9

As calculated, the optimal split for the root node based on "Is Round" results in a Gini Index of 0, indicating pure separation between fruit and non-fruit samples in this subset. Therefore, "Is Round = 1" is the best split for the root node.

Una domanda per l'esame qui potrebbe essere calcolare il valore di Gini quando il valore di isRed è 1.