DALARNA UNIVERSITY

Machine Learning

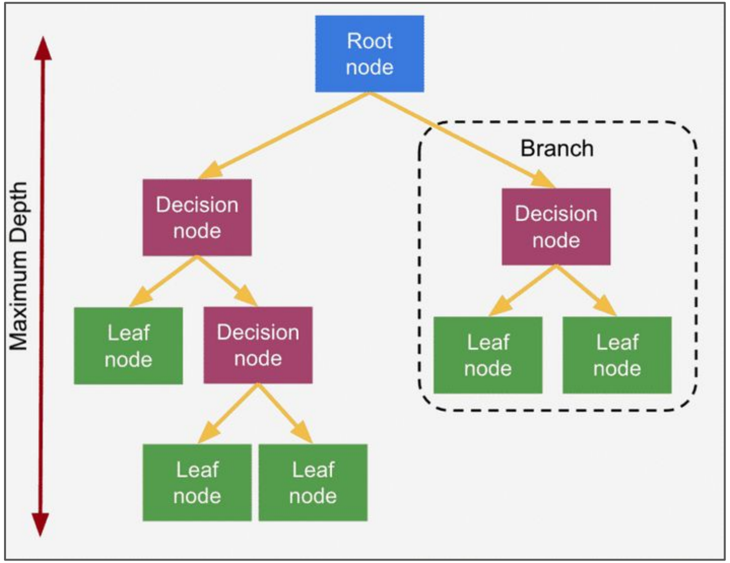
Supervised Learning

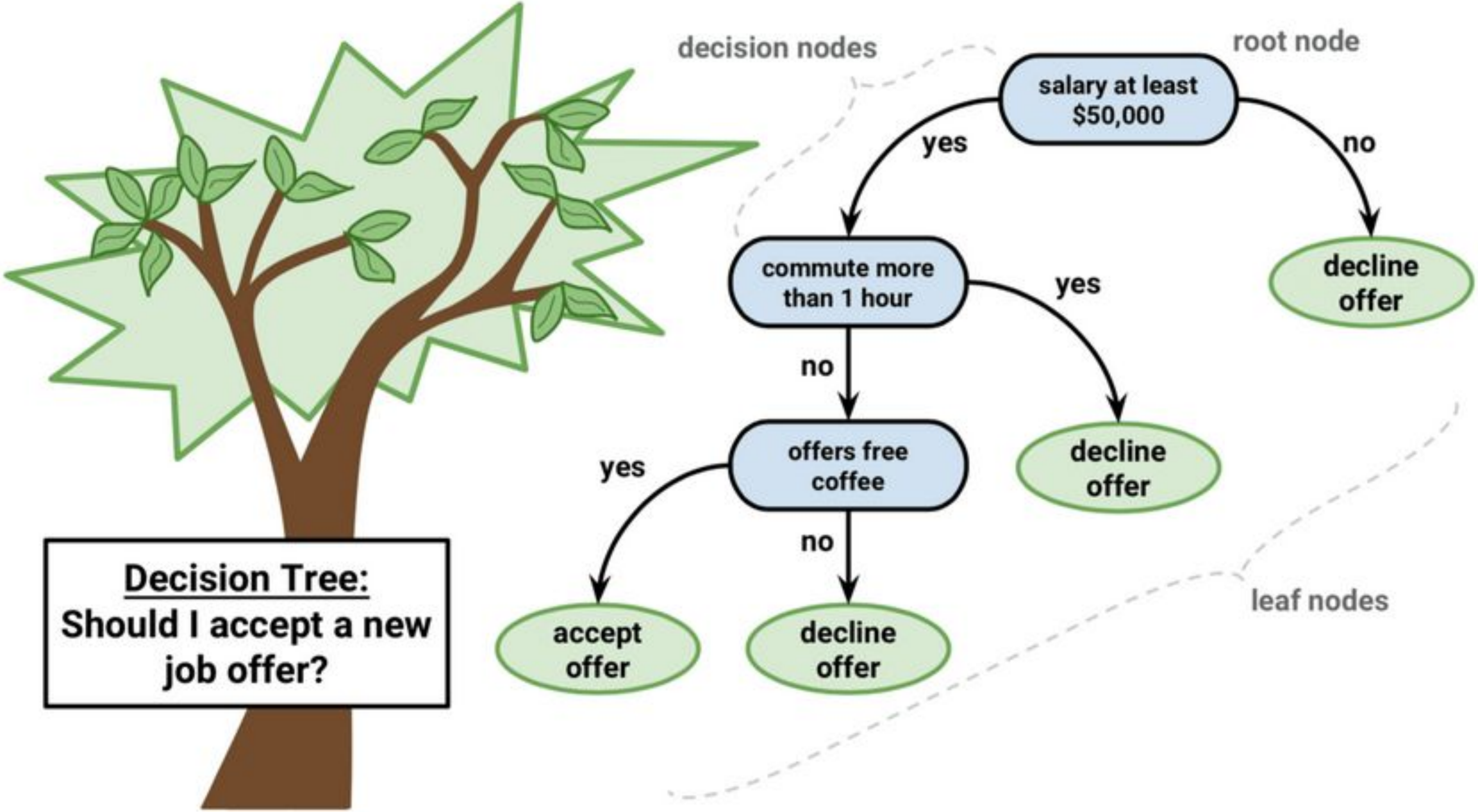
Decision Trees for **Classification**

Roger G. Nyberg

Decision Trees for Classification

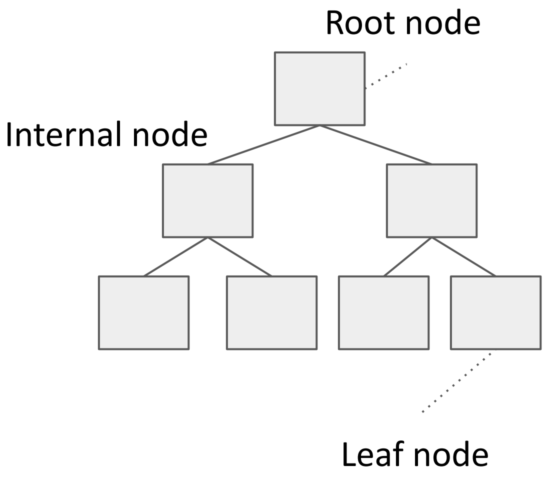
- A partitioning based technique
- Divides the search space into rectangular regions
- DT approaches differ in how the tree is built
- Some Algorithms: ID3, C4.5, CART, Hunt's, etc




decision nodes
root node
salary at least $50,000
yes
no
commute more than 1 hour
yes
decline offer
no
offers free coffee
decline offer
yes
no
accept offer
decline offer
leaf nodes
Decision Tree:
Should I accept a new job offer?

Classification and Regression Trees (CART)

- Introduced by Leo Breiman refers to Decision Tree algorithms and are used for classification or regression predictive modeling problems
    - Here focus is on using CART for classification
- The representation of the CART model is a binary tree, i.e each node can have zero, one or two child nodes
- A node represents a single input variable (X) and a split point on that variable, assuming the variable is numeric
- The leaf nodes of the tree contain an output variable (y) which is used to make a prediction



Very good explanation of decision trees

- **Decision and Classification Trees, Clearly Explained!!! [StatQuest]**
  https://www.youtube.com/watch?v=_L39rN6gz7Y (17 min)
  - Watch this at least twice.
  - You must be able to manually calculate which node should be set as the root node

Gini Index - Binary trees

- Gini index or **Gini impurity** measures the degree or probability of a particular variable being wrongly classified when it is randomly chosen
- Impurity
  - If all the elements belong to a single class, then it can be called pure. The degree of Gini index varies between 0 and 1, where
    - 0 denotes that all elements belong to a certain class, or if there exists only one class
    - 1 denotes that the elements are randomly z-distributed across various classes
    - 0.5 denotes uniformly distributed elements into some classes
- Gini Index Formula

$$Gini(P) = 1 - \sum_{i=1}^{n} (p_i)^2$$

, where $P=(p_1, p_2, p_3, \ldots, p_n)$, and $\boldsymbol{p_i}$ is the probability of an object being classified to a particular class

- While building the decision tree, choose the attribute/feature with the smallest Gini index as the root node

Predicting if someone loves the movie [Cool as Ice (1991)]

DALARNA UNIVERSITY

- Data is collected using a survey strategy

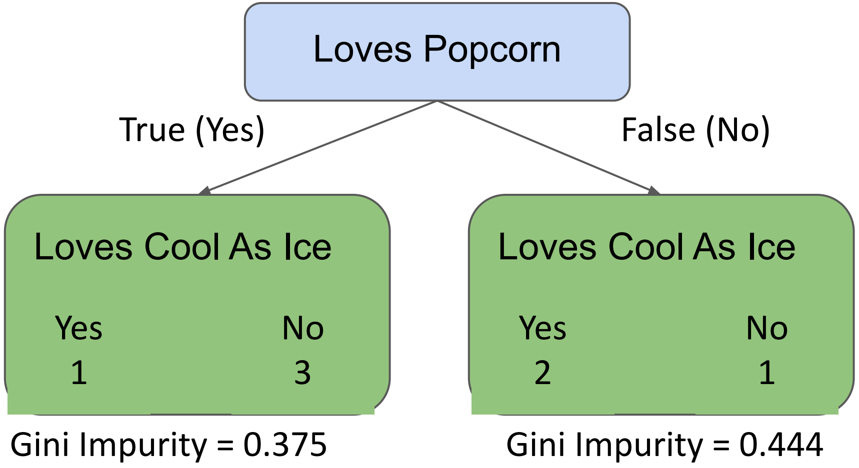
- Now, begin to build the tree!
  - Right now there is no tree at all
  - Where do I start!?
- Decide which feature (class) shall be the *root node*

classes/features/attributes — Target/response

Loves Popcorn	Loves Soda	Age	Loves "Cool as Ice"
Yes	Yes	7	No
Yes	No	12	No
No	Yes	18	Yes
No	Yes	35	Yes
Yes	Yes	38	Yes
Yes	No	50	No
No	No	83	No

- The feature (here a node) that *best predicts* if someone loves the movie “Cool As Ice (1991)” or not will become the root node
- The feature that has the ***lowest impurity*** is the best and will become the root node

The feature "Loves **popcorn**"



Loves Popcorn		Loves "Cool as Ice"
Yes		No
Yes		No
No		Yes
No		Yes
Yes		Yes
Yes		No
No		No

For each leaf compute: $Gini\ Impurity = 1 - P(\text{"}Yes\text{"})^2 - P(\text{"}No\text{"})^2$

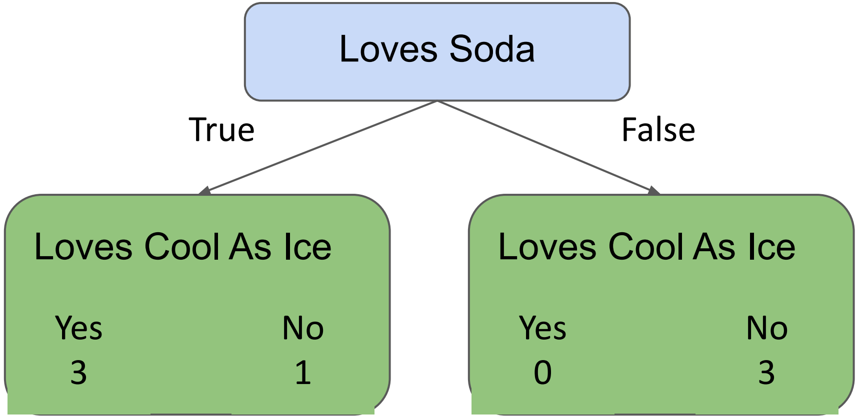
Left: $1\left(\frac{1}{1+3}\right)^2 - \left(\frac{3}{1+3}\right)^2 = 0.375$

Right: $1 - \left(\frac{2}{2+1}\right)^2 - \left(\frac{1}{2+1}\right)^2 = 0.444$

Total Gini Impurity = Weighted average of the leaf Gini Impurities:

$$\left(\frac{4}{4+3}\right) \cdot 0.375 - \left(\frac{3}{4+3}\right) \cdot 0.444 = 0.405$$

The feature "Loves **soda**"



Gini Impurity = 0.375

Gini Impurity = 0

Loves Soda	Loves "Cool as Ice"
Yes	No
No	No
Yes	Yes
Yes	Yes
Yes	Yes
No	No
No	No

For each leaf compute: $$Gini\ Impurity = 1 - P(\text{"}Yes\text{"})^2 - P(\text{"}No\text{"})^2$$

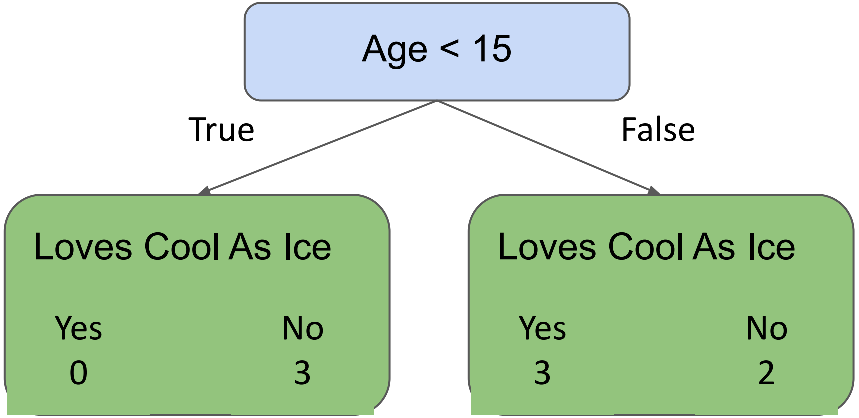
Left: $$1 - \left(\frac{3}{3+1}\right)^2 - \left(\frac{1}{3+1}\right)^2 = 0.375$$

Right: $$1 - \left(\frac{0}{0+3}\right)^2 - \left(\frac{3}{0+3}\right)^2 = 0$$

Total Gini Impurity = Weighted average of the leaf Gini Impurities:

$$\left(\frac{4}{4+3}\right) \cdot 0.375 + \left(\frac{3}{4+3}\right) \cdot 0 = 0.214$$

The feature “Age” → Age < 15



Age	Loves "Cool as Ice"
7	No
12	No
18	Yes
35	Yes
38	Yes
50	No
83	No

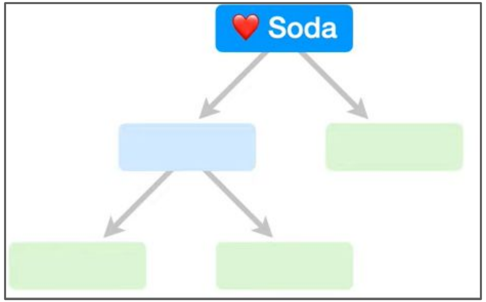
- How were the computations made for numerical values?
- How was the conclusion to use “Age < 15” reached?
- **Total Gini Impurity** for (Age < 15) = 0.343

Watch the previous video

Compare and decide which node should be set as the root

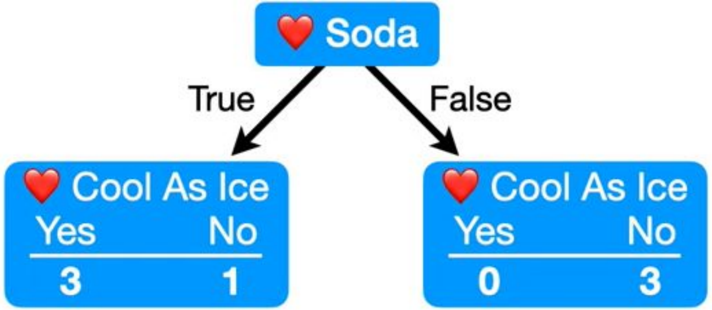
- From the three given features (here represented as nodes), we now have computed:
  - "Loves Popcorn": Total Gini impurity = 0.405
  - "Loves Soda": Total Gini impurity = 0.214
  - "Age": Total Gini impurity = 0.343

- The feature (node) "Loves Soda" has the lowest total Gini impurity and thus **gives us the best initial prediction** among the three features (nodes).
  - This is why this node shall become the root node of the tree!



Continuing building the tree

- Now that we have the root node we know that the question to decide on is “Do you love soda?”
- The right child node has the Gini impurity = 0, so this will become a leaf node.
- The left child node is impure, so we need to investigate if we can reduce the impurity based upon “Loves Popcorn” and “Age”.



References

Bishop, C. M., ***Pattern Recognition and Machine Learning*** (Information Science and Statistics), 2006. Springer-Verlag New York, Inc., Secaucus, NJ, USA.

Duda, R., ***Pattern Classification***, 2nd ed., 2001, Wiley-Interscience

Classification And Regression Trees for Machine Learning
https://machinelearningmastery.com/classification-and-regression-trees-for-machine-learning/

What is Information Gain and Gini Index in Decision Trees?
https://www.analyticssteps.com/blogs/what-gini-index-and-information-gain-decision-trees