Decision Trees

Decision Trees

- A decision tree is a flowchart-like tree structure where:
  - internal node = feature (or attribute)
  - the branch = a decision rule
  - leaf node= the outcome

The Basic Idea

1. Select the best attribute using Attribute Selection Measures(ASM) to split the records.

2. Make that attribute a decision node and breaks the dataset into smaller subsets.

3. Starts tree building by repeating this process recursively for each child until one of the condition will match:

    - All the tuples belong to the same attribute value

    - There are no more remaining attributes.

    - There are no more instances.

Atribute Selection Measure

- Attribute selection measure is a way for selecting the splitting criterion that partition data in the “best” possible way.

- **Information Gain:** the decrease in entropy
  - entropy: measures the impurity of the input set
  - computes the difference between entropy before split and average entropy after split of the dataset based on given attribute values
  - the attribute with the highest information gain is chosen as the splitting attribute at node

- **Gain ratio:** normalizing the information gain
    - The attribute with the highest gain ratio is chosen as the splitting attribute

- **Gini index:** considers a binary split for each attribute

$$\text{Gini}(D)=1-\sum_{i=1}^{m} Pi^2$$

Pi is the probability that a tuple belongs the class

- Discrete-valued attribute: the subset that gives the minimum gini index for that attribute is selected as a splitting attribute.

- Continuous-valued attributes: the strategy is to select each pair of adjacent values as a possible split-point and the point with smaller gini index is chosen as the splitting point.

References

- https://www.datacamp.com/community/tutorials/decision-tree-classification-python