Andrew Taylor en605.645 10/17/2024

Self-Check 8: ID3 TayloDecision Trees

To construct a decision tree using the ID3 algorithm, we must compute the information gain of each feature at each level of the tree. The ID3 algorithm selects the feature with the highest information gain to split the data at each step.

1. *Entropy Calculation: For each feature (Shape, Size, and Color), I''ll calculate the entropy of the dataset based on how well it classifies the target feature, "Safe?".*
2. *Information Gain: I'll compute the information gain for each feature and select the one with the highest gain.*
3. *Tree Construction: I'll recursively repeat the process, constructing the tree at each node until all leaves represent a pure class.*

Let's begin by calculating the overall entropy for the target feature ("Safe?"):

Step 1: Calculate Entropy of the Dataset

The dataset contains 15 samples. The class label "Safe?" has the following distribution:

- *Yes: 7*
- *No: 8*

Entropy for the dataset is:

E(S) = $-(\frac{7}{15} log_2 \frac{7}{15} + \frac{8}{15} log_2 \frac{8}{15})$ *= 0.99679...*

Step 2: Calculate the Information Gain for Each Feature

To proceed with the ID3 algorithm, we need to calculate the information gain for each feature: Shape, Size, and Color. This involves:

The information gain for each feature is as follows:

- *Shape: 0.028*
- *Size: 0.288*
- *Color: 0.229*

The feature with the highest information gain is Size, so we split the dataset on Size.

Step 3: Split the Dataset on Size and Continue

Now, we split the dataset into two subsets based on the values of Size ("large" and "small") and recursively apply the same process to each subset.

- *Subset where Size = large*
- *Subset where Size = small*

Let's calculate the next step of entropy and information gain for both subsets

After splitting the dataset by Size, the results are:

- *Large subset (Entropy: 0.811):*
    - *Shape: Information Gain = 0.000*
    - *Color: Information Gain = 0.467*
- *Small subset (Entropy: 0.592):*
    - *Shape: Information Gain = 0.198*
    - *Color: Information Gain = 0.198*

For the large subset, the best split is on Color (highest information gain: 0.467). For the small subset, both Shape and Color provide the same information gain (0.198).

Tree Structure So Far:

1. *Size (root):*
    - *Large → split on Color*
    - *Small → split on either Shape or Color*

<u>Summary</u>

Step 1: Root Node (Split by Size)

We already determined that Size has the highest information gain (0.288). So, the first split will be on Size.

- *If Size = large → We need to split based on Color (since Color had the highest gain for the "large" subset).*
- *If Size = small → Both Shape and Color have the same gain, so we can choose either one (we'll go with Shape for simplicity).*

Step 2: Split on the "Large" Subset by Color

For the large subset, the split is based on Color.

- *If Color = green, all instances are Safe = yes (leaf).*

- *If Color = red, all instances are Safe = yes (leaf).*
- *If Color = blue, all instances are Safe = no (leaf).*

This completes the branch for the large subset.

Step 3: Split on the "Small" Subset by Shape

For the small subset, the split is based on Shape.

- *If Shape = round, all instances are Safe = no (leafd).*
- *If Shape = square, further split by Color.*
    - *If Color = green, all instances are Safe = no (leaf).*
    - *If Color = red, all instances are Safe = no (leaf).*
    - *If Color = blue, all instances are Safe = no (leaf).*

Final Tree

```
                        Size
                       /     \
                    large     small
                 Color       Shape
                /  | \            /    \
      blue red green    round    square
No      Yes      Yes        No          Color
                                       /   |  \
                               green red blue
                               No    No    No
```