Statistical Machine Learning Spring 2022

Decision Tree

1 Algorithms

1.1 ID3 Algorithm

1. Calculate the Information Gain (IG) of each feature.
2. Make a decision tree node using the feature with the maximum IG.
3. Split the data based on all possible outcomes of the chosen feature.
4. Within each subset of data, if they all belong to the same class, make the current node as a leaf node with the class as its label.
5. Repeat for the remaining features until we run out of all features, or the decision tree has all leaf nodes.

1.2 Improvements in C4.5 from ID3

C4.5 made a number of improvements to ID3. Some of these are:

- Handling both continuous and discrete attributes - In order to handle continuous attributes, C4.5 creates a threshold and then splits the list into those whose attribute value is above the threshold and those that are less than or equal to it.
- Handling training data with missing attribute values - C4.5 allows attribute values to be marked as ? for missing. Missing attribute values are simply not used in gain and entropy calculations.
- Handling attributes with differing costs.
- Pruning trees after creation - C4.5 goes back through the tree once it's been created and attempts to remove branches that do not help by replacing them with leaf nodes.

  How does this last property related to Occam's Razor?

1.3 Improvements in C5.0 from C4.5

- C5.0 is significantly faster than C4.5.
- C5.0 is more memory efficient than C4.5.
- C5.0 gets similar results to C4.5 with considerably smaller decision trees.
- C5.0 allows you to weight different cases and misclassification types.

2 Properties

Please answer these following questions yourself.

1. Generative or Discriminative classifier?
2. Overfitting? If overfitting may occur, how to overcome it? Can C4.5 overcome the overfitting problem that may or may not occur when using ID3?
3. Unique outcome from the same set of data? In HW1 Q5, how many decision trees you can have if restrict the depth of the tree by 2?