Week 9: Daily Morning Challenge

Day 2: Thursday 5th March 2020

Question 1: Briefly illustrate the Gini Index and Gain Ratio as other option for attribute selection for decision tree classification model

The Gini was developed by the Italian statistician Corrado Gini in 1912, for the purpose of rating countries by income distribution. The maximum Gini Index = 1 would mean that all the income belongs to one country. The minimum Gini Index = 0 would mean that the income is even distributed among all countries. This index measures the degree of unevenness in the spread of values in the range of a variable. The theory is that variables with a relatively large amount of unevenness in the frequency distribution of values in its range (a high Gini Index value) have a higher probability to serve as a predictor variable for another related variable.

The empirical formula for the Gini score is
(5.1)G=n+1n−2∑1nn+1−ixin∑1nxi

where x_i is the value of I-variable, sorted from least to greatest

Gain ratio: This is a modification of information gain that reduces its bias and is usually the best option. Gain ratio overcomes the problem with information gain by taking into account the number of branches that would result before making the split. It corrects information gain by taking the intrinsic information of a split into account. The gain ratio is obtained by dividing the information gain for an attribute by its intrinsic information. Clearly attributes that have high intrinsic information (high uncertainty) tend to offer low gains upon splitting and, hence, would not be preferred in the selection process.

Question 2: Briefly illustrate how you would prune (pre or post) a decision tree classification model

In machine learning and data mining, pruning is a technique associated with decision trees. Pruning reduces the size of decision trees by removing parts of the tree that do not provide power to classify instances. Decision trees are the most susceptible out of all the machine learning algorithms to overfitting and effective pruning can reduce this likelihood.

Post-pruning
As the name implies, pruning involves cutting back the tree. After a tree has been built (and in the absence of early stopping discussed below) it may be overfitted. The

CART algorithm will repeatedly partition data into smaller and smaller subsets until those final subsets are homogeneous in terms of the outcome variable. In practice this often means that the final subsets (known as the leaves of the tree) each consist of only one or a few data points. The tree has learned the data exactly, but a new data point that differs very slightly might not be predicted well.
I will consider 3 pruning strategies,

- Minimum error. The tree is pruned back to the point where the cross-validated error is a minimum. Cross-validation is the process of building a tree with most of the data and then using the remaining part of the data to test the accuracy of the decision tree.
- Smallest tree. The tree is pruned back slightly further than the minimum error. Technically the pruning creates a decision tree with cross-validation error within 1 standard error of the minimum error. The smaller tree is more intelligible at the cost of a small increase in error.