Trang Do

COM 307 Machine Learning

Professor Stephen Douglass

A not too official write-up:

The steps I am following to build my decision tree:

1. I split the data set into the training and testing dataset as a ratio of 70:30.
2. I calculate the Gini Index, a metric to measure how often a randomly chosen element would be incorrectly identified, with this formula:

$$\text{Gini Index} = 1 - \sum_j P_j^2$$

   In this case, an attribute with lower gini index would be preferred

3. Then I calculate Entropy. This is the measure of uncertainty of a random variable and it characterizes the impurity of an arbitrary collection of examples.

   The higher the entropy the more the information content, using this formula:

$$H(x) = -\sum_{i=1}^{N} p(x_i) \log_2 p(x_i)$$

4. After my model makes predictions, I calculate the accuracy of the trained classifier. In this case, the accuracy is 0.87, which is pretty high.
5. The last step, I print out the decision tree as a pdf.

   My Decision Tree Depth = 7, the level 1 gini = 0.196, which is pretty low.