Assignment 1: Decision Tree

August 2021

Instructions

- Use python programming language for your implementation.
- Use appropriate approach if you find some attribute is missing in your data.
- Report must contain step-wise description of your implementation and analysis of results. Since data analysis is a crucial task for any machine learning algorithm, report should demonstrate detailed analysis of results and conclusion. It should also clearly mention the steps to run your code.
- If the decision tree building is done using any package, penalty is -10 for that part.

Question 1

1. Build a decision-tree classifier by randomly splitting the dataset as 80/20 split. Use the impurity measures- 1) gini index and 2) information gain. Analyze the impact of using individual impurity measures on the prediction. Do not use package for building the tree and implement this part on your own. **15**
2. Provide the accuracy by averaging over 10 random 80/20 splits. Consider that particular tree which provides the best test accuracy as the desired one. **15**
3. What is the best possible depth limit to be used for your dataset. Provide a plot explaining the same. Also provide a plot of the test accuracy vs. the total number of nodes in the trees. **5+5**
4. Perform the pruning operation over the tree with the highest test accuracy in question 2 using a valid statistical test for comparison. **20**
5. Print the final decision tree obtained from question 3 following the hierarchical levels of data attributes as nodes of the tree. **10**
6. A brief report explaining the procedure and the results. **30**