

1. Which impurity heuristic (Entropy/Variance) yields the best classification accuracy? How does increasing the number of examples

and/or the number of clauses impact the (accuracy of the) two

impurity heuristics. Explain your answer.

According to the above table, Entropy heuristic yields better classification accuracy, if pruning is not done.

The more number of examples the model is trained on the greater it’s accuracy. This is because the model gets trained on a larger percentage of the population and so it gets more fine-tuned with increase in the number of examples. So it is able to classify new examples with more and more accuracy.

Increasing the number of clauses in the training data also increases the accuracy of prediction as the model becomes more expressive with more number of clauses and it is able to explore a larger hypothesis space to find the closest estimate of the true function.

2. Which over fitting avoidance method (reduced error pruning/ depth-

based pruning) yields the best accuracy? Again, how does increasing the number of examples and/or the number of clauses impact

the (accuracy of the) two over fitting avoidance methods. Explain

your answer.

According the above table Reduced error pruning yields the best accuracy as it takes the accuracy of the resulting tree into account while pruning each node and only prunes when the accuracy increases after pruning. Whereas in the case of depth-based pruning, it only considers the accuracy of the tree as a whole while pruning. It is equivalent to blindly chopping of branches beyond a certain depth which in most cases doesn’t ends up matching reduced error pruning in accuracy.

Increasing the number of examples and/or the number of clauses also increases the accuracy of the two over fitting avoidance methods as both increases the accuracy of the underlying naïve tree. Both pruning methods provide additional accuracy gain over and above that achieved by the naïve tree.

3. Are random forests much better in terms of classification accuracy

than your decision tree learners? Why? Explain your answer.

Yes. Random forests are much better in terms of classification accuracy than most of my decision trees. This is because the decision trees can only take a discrete decision at each node so even one incorrect classification at a particular node on the chosen path with mostly results in incorrect classification. Whereas random forests average the decision of multiple such trees so it is very difficult for a majority of trees to classify the data incorrectly. The probability of most of the trees being wrong is very less.