Assignment 6 (Sol.)

Introduction to Machine Learning
Prof. B. Ravindran

1. Assume that you are given a data set and a neural network model trained on the data set. You are asked to build a decision tree model with the sole purpose of understanding/interpreting the built neural network model. In such a scenario, which among the following measures would you concentrate most on optimising?

   (a) Accuracy of the decision tree model on the given data set

   (b) F1 measure of the decision tree model on the given data set

   (c) Fidelity of the decision tree model, which is the fraction of instances on which the neural network and the decision tree give the same output

   (d) Comprehensibility of the decision tree model, measured in terms of the size of the corresponding rule set

   Sol. (c)
   Here the aim is not the traditional one of modelling the data well, rather it is to build a decision tree model which is as close to the existing neural network model as possible, so that we can use the decision tree to interpret the neural network model. Hence, we optimise the fidelity measure.

2. Which of the following properties are characteristic of decision trees?

   (a) High bias

   (b) High variance

   (c) Lack of smoothness of prediction surfaces

   (d) Unbounded parameter set

   Sol. (b), (c) & (d)
   Decision trees are generally unstable considering that a small change in the data set can result in a very different set of splits. This is mainly due to the hierarchical nature of decision trees, since a change in split points in the initial stages will affect all the subsequent splits.

   The decision surfaces that result from decision tree learning are generated by recursive splitting of the feature space using axis parallel hyper planes. They clearly do not produce smooth prediction surfaces such as the ones produced by, say, neural networks.

   Decision trees do not make any assumptions about the distribution of the data. They are non-parametric methods where the number of parameters depends solely on the data set on which training is carried out.

3. To control the size of the tree, we need to control the number of regions. One approach to do this would be to split tree nodes only if the resultant decrease in the sum of squares error exceeds some threshold. For the described method, which among the following are true?

   (a) It would, in general, help restrict the size of the trees

(b) It has the potential to affect the performance of the resultant regression/classification model

(c) It is computationally infeasible

Sol. (a) & (b)
While this approach may restrict the eventual number of regions produced, the main problem with this approach is that it is too restrictive and may result in poor performance. It is very common for splits at one level, which themselves are not that good (i.e., they do not decrease the error significantly), to lead to very good splits (i.e., where the error is significantly reduced) down the line. Think about the XOR problem.

4. Which among the following statements best describes our approach to learning decision trees

(a) Identify the best partition of the input space and response per partition to minimise sum of squares error

(b) Identify the best approximation of the above by the greedy approach (to identifying the partitions)

(c) Identify the model which gives the best performance using the greedy approximation (option (b)) with the smallest partition scheme

(d) Identify the model which gives performance close to the best performance (option (a)) with the smallest partition scheme

(e) Identify the model which gives performance close to the best greedy approximation performance (option (b)) with the smallest partition scheme

Sol. (e)
As was discussed in class we use a greedy approximation to identifying the partitions and typically use pruning techniques which result in a smaller tree with probably some degradation in the performance.

5. Having built a decision tree, we are using reduced error pruning to reduce the size of the tree. We select a node to collapse. For this particular node, on the left branch, there are 3 training data points with the following outputs: 5, 7, 9.6 and for the right branch, there are four training data points with the following outputs: 8.7, 9.8, 10.5, 11. What were the original responses for data points along the two branches (left & right respectively) and what is the new response after collapsing the node?

(a) 10.8, 13.33, 14.48

(b) 10.8, 13.33, 12.06

(c) 7.2, 10, 8.8

(d) 7.2, 10, 8.6

Sol. (c)
Original responses:
Left: $\frac{5+7+9.6}{3} = \frac{21.6}{3} = 7.2$
Right: $\frac{8.7,9.8,10.5,11}{4} = \frac{40}{4} = 10$
New response: $7.2 * \frac{3}{7} + 10 * \frac{4}{7} = 8.8$

6. Given that we can select the same feature multiple times during the recursive partitioning of the input space, is it always possible to achieve 100% accuracy on the training data (given that we allow for trees to grow to their maximum size) when building decision trees?

   (a) Yes

   (b) No

   Sol. (b)
   Consider a pair of data points with identical input features but different class labels. Such points can be part of the training data but will not be able to be classified without error.

7. Suppose on performing reduced error pruning, we collapsed a node and observed an improvement in the prediction accuracy on the validation set. Which among the following statements are possible in light of the performance improvement observed?

   (a) The collapsed node helped overcome the effect of one or more noise affected data points in the training set

   (b) The validation set had one or more noise affected data points in the region corresponding to the collapsed node

   (c) The validation set did not have any data points along at least one of the collapsed branches

   (d) The validation set did have data points adversely affected by the collapsed node

   Sol. (a), (b), (c) & (d)
   The first option is the kind of error we normally expect pruning to help us overcome. However, a node collapse which ideally should result in an increase in the overall error of the model may actually show an improvement due to a number of factors. Perhaps the points which should have been misclassified due to the collapse are mislabelled in the validation set (option (b)). Such points may also be missing from the the validation set (option (c)). Finally, even if the increased error due to the collapsed node is registered in the validation set, it may be masked by the absence of errors (existing in the training data) in other parts of the validation set (option (d)).

8. Consider the following data set:

price	maintenance	capacity	airbag	profitable
low	low	2	no	yes
low	med	4	yes	no
low	low	4	no	yes
low	high	4	no	no
med	med	4	no	no
med	med	4	yes	yes
med	high	2	yes	no
med	high	5	no	yes
high	med	4	yes	yes
high	high	2	yes	no
high	high	5	yes	yes

Considering 'profitable' as the binary values attribute we are trying to predict, which of the attributes would you select as the root in a decision tree with multi-way splits using the cross-entropy impurity measure?

(a) price
(b) maintenance
(c) capacity
(d) airbag

Sol. (c)
$cross\_entropy_{price}(D) = \frac{4}{11}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{4}{11}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{3}{11}(-\frac{2}{3}log_2\frac{2}{3} - \frac{1}{3}log_2\frac{1}{3}) = 0.9777$
$cross\_entropy_{maintenance}(D) = \frac{2}{11}(-\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2}) + \frac{4}{11}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{5}{11}(-\frac{2}{5}log_2\frac{2}{5} - \frac{3}{5}log_2\frac{3}{5}) = 0.8050$
$cross\_entropy_{capacity}(D) = \frac{3}{11}(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}) + \frac{6}{11}(-\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6}) + \frac{2}{11}(-\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2}) = \mathbf{0.7959}$
$cross\_entropy_{price}(D) = \frac{5}{11}(-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5}) + \frac{6}{11}(-\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6}) = 0.9868$

9. For the same data set, suppose we decide to construct a decision tree using binary splits and the Gini index impurity measure. Which among the following feature and split point combinations would be the best to use as the root node assuming that we consider each of the input features to be unordered?

(a) price - {low, med}|{high}
(b) maintenance - {high}|{med, low}
(c) maintenance - {high, med}|{low}
(d) capacity - {2}|{4, 5}

Sol. (c)
$gini_{price(\{low,med\}|\{high\})}(D) = \frac{8}{11} * 2 * \frac{4}{8} * \frac{4}{8} + \frac{3}{11} * 2 * \frac{2}{3} * \frac{1}{3} = 0.4848$
$gini_{maintenance(\{high\}|\{med,low\})}(D) = \frac{5}{11} * 2 * \frac{2}{5} * \frac{3}{5} + \frac{6}{11} * 2 * \frac{4}{6} * \frac{2}{6} = 0.4606$
$gini_{maintenance(\{high,med\}|\{low\})}(D) = \frac{9}{11} * 2 * \frac{4}{9} * \frac{5}{9} + \frac{2}{11} * 2 * 1 * 0 = \mathbf{0.4040}$
$ginit_{capacity(\{2\}|\{4,5\})}(D) = \frac{3}{11} * 2 * \frac{1}{3} * \frac{2}{3} + \frac{8}{11} * 2 * \frac{5}{8} * \frac{3}{8} = 0.4621$

10. Consider building a spam filter for distinguishing between genuine e-mails and unwanted spam e-mails. Assuming spam to be the positive class, which among the following would be more important to optimise?

(a) Precision
(b) Recall

Sol. (a)
If we optimise recall, we may be able to capture more spam e-mails, but in the process, we may also increase the number of genuine mails being predicted as spam. On the other hand, if we optimise precision, we may not be able to capture as many spam e-mails as in the previous

approach, but the percentage of e-mails being classified as spam actually being spam will be high.

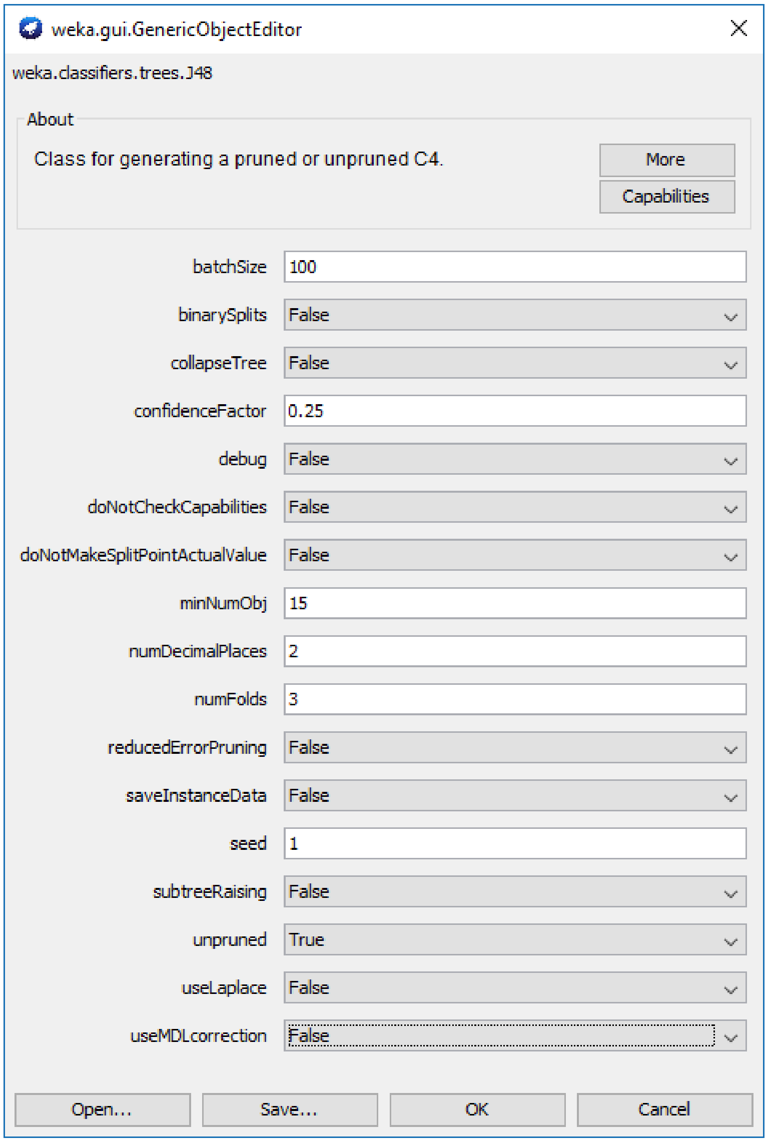
Of the two possible trade-offs, we would prefer the later (optimising precision), since in the former approach, we will be risking more genuine e-mails being classified as spam, which is a costlier error to make in a spam filtering application (as compared to missing out a few spam e-mails which the user may have to deal with manually).

Weka-based assignment questions

In this assignment, we will use the UCI Mushroom data set available here.

We will be using the J48 decision tree algorithm which can be found in Weka under classifiers/trees.

We will consider the following to be the default parameter settings:



weka.gui.GenericObjectEditor

weka.classifiers.trees.J48

About

Class for generating a pruned or unpruned C4.

batchSize	100
binarySplits	False
collapseTree	False
confidenceFactor	0.25
debug	False
doNotCheckCapabilities	False
doNotMakeSplitPointActualValue	False
minNumObj	15
numDecimalPlaces	2
numFolds	3
reducedErrorPruning	False
saveInstanceData	False
seed	1
subtreeRaising	False
unpruned	True
useLaplace	False
useMDLcorrection	False

Note the following:

The class for which the prediction model is to be learned is named 'class' and is the first attribute in the data.

We will use the default 'Cross-validation' test option with 'Folds = 10'.

Once a decision tree model has been built, you can right click on the corresponding entry in the 'Result list' pane on the bottom left, and select 'Visualize tree' to see a visual representation of the learned tree.

11. How many levels does the unpruned tree contain considering multi-way and binary splits respectively, with the other parameters remaining the same as above?

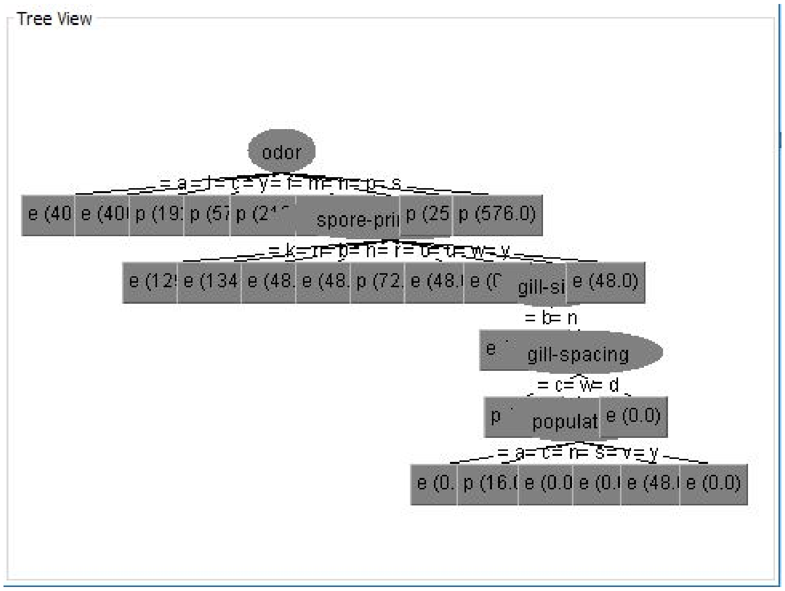
(a) 6, 8

(b) 6, 7

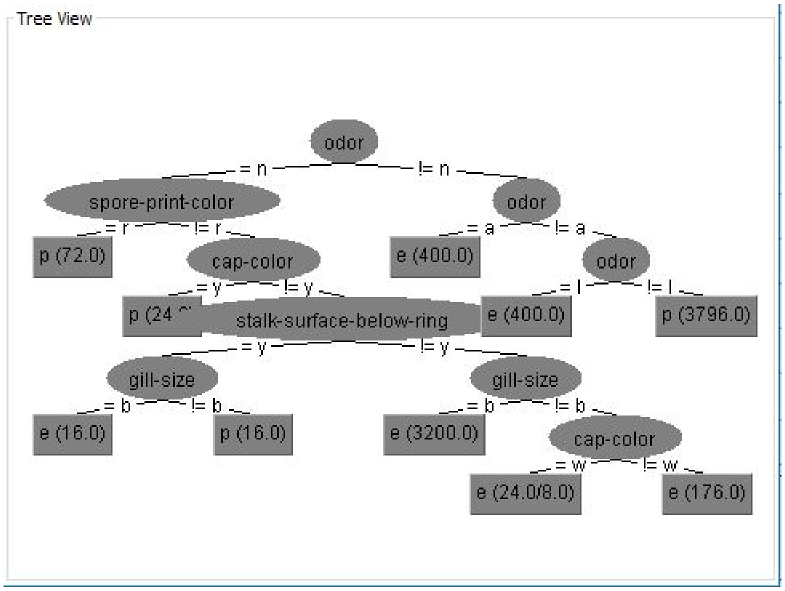
(c) 5, 7

(d) 5, 8

Sol. (b)



Multi-way split



Binary split

12. How many levels does the pruned tree (unpruned = false, reducedErrorPruning = false) contain considering multi-way and binary splits respectively?

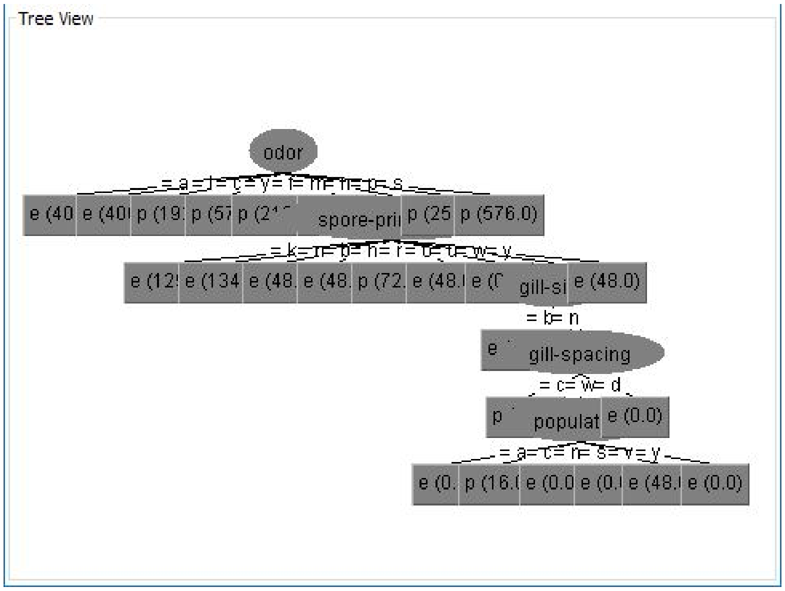
    (a) 6, 6

    (b) 6, 7

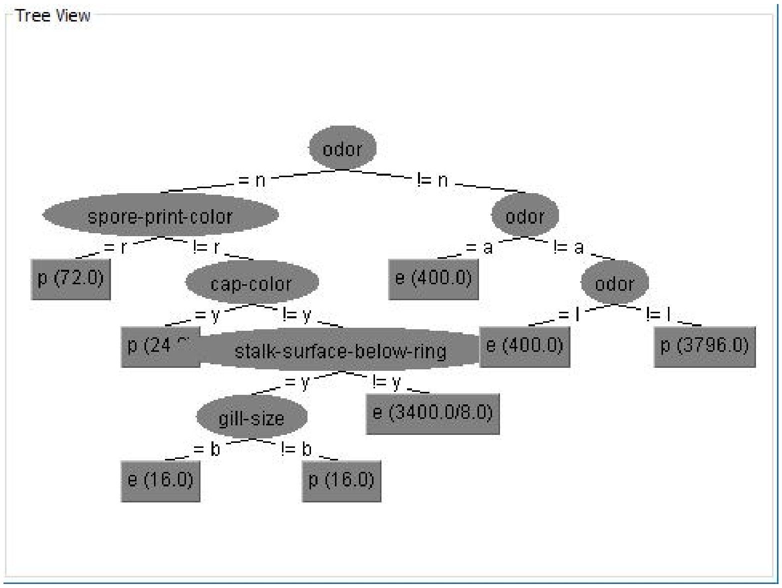
    (c) 5, 7

    (d) 5, 6

    Sol. (a)



Multi-way split



Binary split

13. Consider the effect of the parameter minNumObj. Try modifying the parameter value and observe the changes in the performance values. For example, considering respectively, multi-way and binary splits with the default parameters as discussed before, at what (maximum) value of the parameter do we start to see zero error?

(a) 11, 7

(b) 11, 6

(c) 10, 6

(d) 10, 7

Sol. (c)
This can be observed by trying out the listed values with the default parameters. The idea is for you to appreciate the significance of the minNumObj parameter.

14. Which among the following pairs of attributes seem to be the most important for this particular classification task?

(a) population, gill-spacing

(b) stalk-surface-below-ring, cap-color

(c) gill-spacing, ring-number

(d) odor, spore-print-color

Sol. (d)
From multiple models built using binary/multi-way splits and with and without pruning, we observe that the attributes, odor and spore-print-color always appear at higher levels in the trees indicating their importance for the classification task.