Assignment 5 (Sol.)

Introduction to Data Analytics

Prof. Nandan Sudarsanam & Prof. B. Ravindran

1. In a binary classification scenario where x is the independent variable and y is the dependent variable, logistic regression assumes that the conditional distribution $y|x$ follows a

   (a) Bernoulli distribution
   (b) binomial distribution
   (c) normal distribution
   (d) exponential distribution

   Sol. (a)
   The dependent variable is binary, so a Bernoulli distribution is assumed.

2. 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.

3. For a K-NN classification model, as the value of K increases

   (a) bias increases and variance decreases
   (b) bias decreases and variance increases

   Sol. (a)
   As K increases, the decision boundaries become smoother and the overall model becomes less complex, hence bias is increasing. Similarly, as K increases, due to considering more neighbours to make decisions, small changes in the data set do not result in drastic differences in predictions, i.e., variance is decreasing.

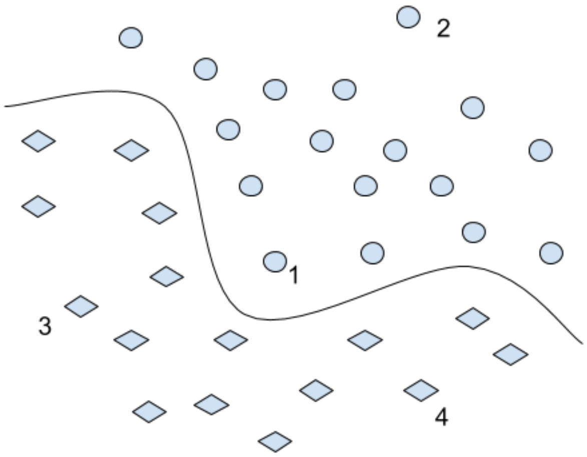
4. For a supervised learning problem, we have seen that the data used in the testing phase, i.e., the test set, is not used for training or building the model. Does this mean that we do not require labels for data points in the test set?

   (a) no

   (b) yes

   Sol. (a)
   To give an assessment of how good the learned model performs, we do need labels for data points in the test set so that we can compare the learned model's predictions to the actual labels.

5. For a two-class classification problem, we use an SVM classifier and obtain the following separating hyperplane.



   We have marked 4 instances of the training data. Identify the point which will have the most impact on the shape of the boundary on it's removal.

   (a) 1

   (b) 2

   (c) 3

   (d) 4

   Sol. (a)
   We need to identify support vectors on which the hyperplane is supported. In the figure above, data point 1 is a support vector. Removal of data point 1 will have an impact on the decision boundary.

6. For a linearly non-separable data set, are the points which are misclassified by the SVM model support vectors?

   (a) no

(b) yes

Sol. (b)
Since such points are involved in determining the decision boundary, they (along with points lying on the margins) are support vectors. Note that eliminating (or not considering) any such point will have an impact on the decision boundary.

7. In the linearly non-separable case, what effect does the C parameter have on the SVM model?

   (a) it determines the count of support vectors

   (b) it is a count of the number of data points which do not lie on their respective side of the hyperplane

   (c) it determines how many data points lie within the margin

   (d) it allows us to trade-off the number of misclassified points in the training data and the size of the margin

   Sol. (d)
   A high value of the C parameter results in more emphasis being given to the penalties arising out of points lying on the wrong sides of the margins. This results in reducing the number of such points being considered in deciding the decision boundary by reducing the margin.

8. Suppose that we use a radial basis function kernel with appropriate parameters to perform classification on a particular two class data set where the data is not linearly separable. In this scenario

   (a) the decision boundary in the original feature space is linear

   (b) the decision boundary in the original feature space is non-linear

   (c) the decision boundary in the transformed feature space is linear

   (d) the decision boundary in the transformed feature space is non-linear

   Sol. (b), (c)

9. 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	4	yes	yes
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)
$cross\_entropy_{price}(D) = \frac{4}{12}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{4}{12}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{4}{12}(-\frac{3}{4}log_2\frac{3}{4} - \frac{1}{4}log_2\frac{1}{4}) = 0.9371$
$cross\_entropy_{maintenance}(D) = \frac{2}{12}(-\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2}) + \frac{4}{12}(-\frac{2}{4}log_2\frac{2}{4} - \frac{2}{4}log_2\frac{2}{4}) + \frac{6}{12}(-\frac{3}{6}log_2\frac{3}{6} - \frac{3}{6}log_2\frac{3}{6}) = 0.8333$
$cross\_entropy_{capacity}(D) = \frac{3}{12}(-\frac{1}{3}log_2\frac{1}{3} - \frac{2}{3}log_2\frac{2}{3}) + \frac{7}{12}(-\frac{4}{7}log_2\frac{4}{7} - \frac{3}{7}log_2\frac{3}{7}) + \frac{2}{12}(-\frac{2}{2}log_2\frac{2}{2} - \frac{0}{2}log_2\frac{0}{2}) = \mathbf{0.8043}$
$cross\_entropy_{airbag}(D) = \frac{5}{12}(-\frac{3}{5}log_2\frac{3}{5} - \frac{2}{5}log_2\frac{2}{5}) + \frac{7}{12}(-\frac{4}{7}log_2\frac{4}{7} - \frac{3}{7}log_2\frac{3}{7}) = 0.9793$

10. 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)
$gini_{price(\{low,med\}|\{high\})}(D) = \frac{8}{12} * 2 * \frac{4}{8} * \frac{4}{8} + \frac{4}{12} * 2 * \frac{3}{4} * \frac{1}{4} = 0.4583$
$gini_{maintenance(\{high\}|\{med,low\})}(D) = \frac{6}{12} * 2 * \frac{3}{6} * \frac{3}{6} + \frac{6}{12} * 2 * \frac{4}{6} * \frac{2}{6} = 0.4722$
$gini_{maintenance(\{high,med\}|\{low\})}(D) = \frac{10}{12} * 2 * \frac{5}{10} * \frac{5}{10} + \frac{2}{12} * 2 * 1 * 0 = \mathbf{0.4167}$
$gini_{capacity(\{2\}|\{4,5\})}(D) = \frac{3}{12} * 2 * \frac{1}{3} * \frac{2}{3} + \frac{9}{12} * 2 * \frac{6}{9} * \frac{3}{9} = 0.4444$

Weka-based programming assignment questions

For the questions on SVMs, you will need to download the LibSVM package and use it along with Weka. You can install the package using Weka's package manager (in the Tool's menu of Weka's GUI Chooser window), or follow the instructions given here.

Once successfully installed, you will find the option LibSVM under 'functions' when choosing the classifier. The options that we will need to considered are debug (set to True), degree (for use with polynomial kernel), and kernelType (for choosing the kernel). Note that you will need to launch Weka from the command line (or using the 'Weka 3.8 (with console)' shortcut) to view the output from the optimiser where the parameter nSV will indicate the number of support vectors. Make use the visualise options to view the data. As in the last assignment, we have given both train and test sets for each data set.

11. For data set 1, train linear and radial basis function kernel SVMs (default parameter values). What are the number of support vectors in each case?

    (a) 100, 100
    (b) 3, 104
    (c) 3, 116
    (d) 550, 100

    Sol. (c)

12. For data set 2, train 5 degree polynomial (5 degree, coef0 = 0), 10 degree polynomial (10 degree, coef0 = 0) and radial basis kernel functions (default parameter values). What are the number of support vectors in each case?

    (a) 230, 20, 27
    (b) 324, 20, 27
    (c) 425, 21, 25
    (d) 362, 20, 25

    Sol. (a)

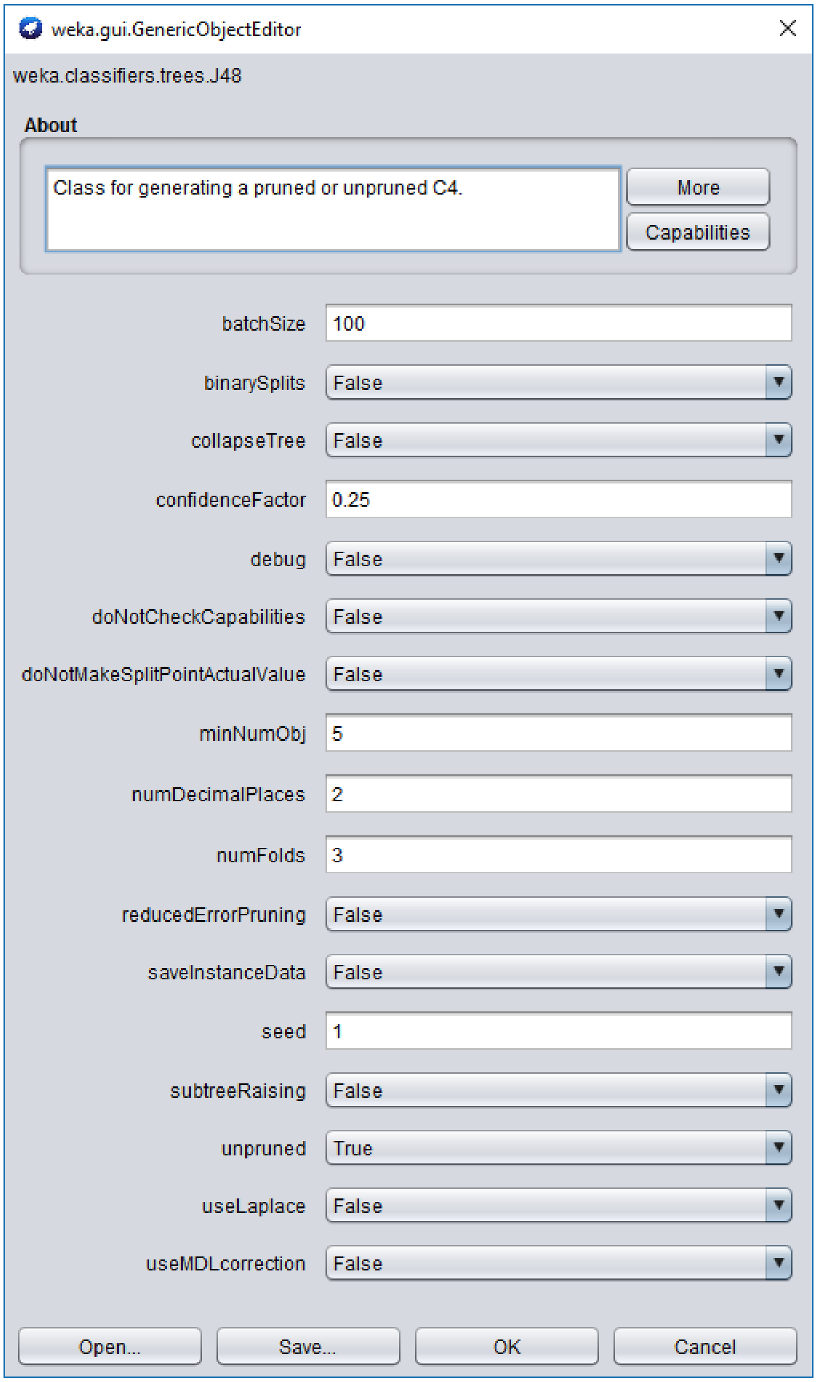
13. In the previous question, which classifier among the three would you prefer to use for unseen data?

    (a) 5 degree polynomial
    (b) 10 degree polynomial
    (c) radial basis function

    Sol. (b)
    While both the 10 degree polynomial and the radial basis function kernel result in zero error, the 10 degree polynomial model is preferred since the number of support vectors in this model is comparatively less, which may indicate less overfitting.

    For the questions on decision tree, we will be using the UCI Tic-Tac-Toe Endgame data set (Dataset5). Use the decision tree model J48 under the trees folder. Make use of 10-fold cross validation for all experiments with this data set. Also, make use of the visualise tree option (by right clicking in the results list) once a decision tree model is built.

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



weka.classifiers.trees.J48

Class for generating a pruned or unpruned C4.

batchSize: 100
binarySplits: False
collapseTree: False
confidenceFactor: 0.25
debug: False
doNotCheckCapabilities: False
doNotMakeSplitPointActualValue: False
minNumObj: 5
numDecimalPlaces: 2
numFolds: 3
reducedErrorPruning: False
saveInstanceData: False
seed: 1
subtreeRaising: False
unpruned: True
useLaplace: False
useMDLcorrection: False

14. 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) 10, 6

(b) 6, 8

(c) 8, 6

(d) 6, 10

Sol. (d)

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

    (a) 10, 6

    (b) 6, 8

    (c) 8, 6

    (d) 6, 10

    Sol. (b)

16. Considering the weighted average F-measure as the performance indicator (higher values indicate better performance), do the pruned trees result in increased performance or decreased performance in the multi-way and binary split cases respectively?

    (a) decrease, increase

    (b) increase, decrease

    (c) increase, increase

    (d) decrease, decrease

    Sol. (a)
    While a decrease is observed in the multi-way split case, we observe a slight increase in the binary split case. This is indicative of the ability of pruning to produce a better tree by reducing overfitting.

17. Which among the following attributes seems to be the most important for this particular classification task?

    (a) top-left-square

    (b) top-right-square

    (c) middle-middle-square

    (d) bottom-middle-square

    Sol. (c)
    Since the middle-middle-square attribute occurs as the root node in the different decision tree models, this attribute appears to be the most discriminative for this particular classification task.