Business analytics and data mining Modeling using R: Week-8

19 March 2023 19:04

Name: Aakash

Assessment scores

Week 1 : Assignment 1: **95.0**

Week 2 : Assignment 2: **100.0**

Week 3 : Assignment 3: **100.0**

Week 4 : Assignment 4: **100.0**

Week 5 : Assignment 5: **100.0**

Week 6 : Assignment 6: **100.0**

Week 7 : Assignment 7: **100.0**

Week 8: Assignment 8: -

Week 9 : Assignment 9: -

1) Which of the following partitions are utilized by pruning steps of decision trees?

- ○ Training partition
- ○ Validation partition
- ○ Test partition
- ○ Any part of the data

(B)

Decision trees are a popular machine learning algorithm that can be used for both classification and regression tasks. When building a decision tree, it is important to split the available data into three different partitions:

A) Training partition: This partition is used to train the decision tree algorithm. The algorithm builds the tree based on the patterns and relationships it discovers in the training data.

B) Validation partition: This partition is used to evaluate the performance of the decision tree during training. The algorithm uses the validation partition to test the performance of the tree on data that it has not seen before. This step is important to prevent overfitting, which can occur when the algorithm becomes too specialized to the training data and performs poorly on new data.

C) Test partition: This partition is used to evaluate the final performance of the decision tree. After the algorithm has been trained and optimized on the training and validation data, it is tested on the test partition to obtain an unbiased estimate of its performance on new data.

In summary, the correct answer to the question is B) Validation partition, as this partition is specifically used during the pruning step of decision tree building to evaluate the tree's performance and make adjustments to reduce overfitting.

2) Which of the following statements is not true about recursive partitioning steps of decision trees?

- ○ Recursive partitioning results in non-overlapping multi-dimensional rectangles
- ○ Recursive partitioning results in smaller and smaller rectangular regions
- ○ Recursive partitioning continues till heterogeneous groups are reached
- ○ Recursive partitioning results in zero error

(C)

3) What will be the gini index value when there are two classes with equal representation of each class?

- ○ 0
- ○ 1
- ○ -1
- ○ 0.5

(D)

When there are two classes with equal representation, the Gini index value will be D) 0.5.

The Gini index is a measure of impurity or heterogeneity in a set of observations, and it ranges from 0 (complete purity, where all observations belong to the same class) to 1 (complete impurity, where the observations are equally distributed among all classes).

When there are two classes with equal representation, this means that the observations are evenly split between the two classes. In this case, the Gini index is given by:

Gini index = 1 - (p1^2 + p2^2)

where p1 and p2 are the proportions of observations in each class. Since both classes have equal representation, p1 = p2 = 0.5. Substituting these values into the equation, we get:

Gini index = 1 - (0.5^2 + 0.5^2) = 1 - 0.25 - 0.25 = 0.5

Therefore, the correct answer is D) 0.5.

4) How many terminal nodes will be estimated in a decision tree with 'n' number of decision nodes?

- ○ n
- ○ (n-1)
- ○ (n+1)
- ○ Can't say

(C)

The number of terminal nodes in a decision tree can vary based on the number of decision nodes, the complexity of the tree, and the sample size. However, the minimum number of terminal nodes that can be estimated in a decision tree with 'n' number of decision nodes is (n+1).

Each decision node in a decision tree creates two branches or child nodes, leading to a total of two possible outcomes for each decision. This means that the number of terminal nodes in a decision tree with 'n' decision nodes can be calculated as (2^n). However, this includes the root node, which is not a decision node but the starting point of the tree. Therefore, the number of terminal nodes in a decision tree can be calculated as (2^n - 1).

In addition to the above formula, a decision tree can also have additional terminal nodes due to ties in the split criteria or because some of the terminal nodes are not pure and need further splitting. Therefore, the actual number of terminal nodes in a decision tree can be (2^n - 1) + additional nodes.

Assuming there are no ties in the split criteria and all terminal nodes are pure, the minimum number of terminal nodes that can be estimated in a decision tree with 'n' number of decision nodes is (n+1). This occurs when the decision tree is a perfectly balanced binary tree, with each decision node splitting the data into two equal-sized subsets, resulting in (2^n) terminal nodes.

Therefore, the correct answer to the question is C) (n+1).

5) Which of the following R functions can be used to filter overlapping observations in training and validation partitions?

- ○ replace
- ○ intersect
- ○ rpart.control
- ○ split

(B)

The R function that can be used to filter overlapping observations in training and validation partitions is B) intersect.

The intersect function in R returns the common elements of two vectors. It can be used to filter the overlapping observations in training and validation partitions by identifying the common elements between the two partitions and removing them from one of the partitions. This ensures that the same observations are not used for both training and validation, which can lead to overfitting and biased model evaluation.

For example, if 'train' and 'val' are two data frames containing the training and validation partitions respectively, the following code can be used to filter the overlapping observations:

bash Copy code

```
common_obs <- intersect(train$ID, val$ID)
val_filtered <- val[!(val$ID %in% common_obs), ]
```

This code first identifies the common observations between the two partitions based on a common ID variable, and then removes these observations from the validation partition. The resulting 'val_filtered' data frame contains only the non-overlapping observations that can be used for unbiased model evaluation.

Therefore, the correct answer to the question is B) intersect.

6) Which of the following is used as the basis for finding a suitable split-value of a predictor?

- ○ Impurity increase
- ○ Impurity reduction
- ○ size
- ○ None of the above

(B)

The basis for finding a suitable split-value of a predictor in decision trees is B) Impurity reduction.

When constructing a decision tree, the goal is to find the best split for each predictor that maximally separates the classes. To accomplish this, the decision tree algorithm evaluates different split points for each predictor and selects the one that maximizes the impurity reduction of the resulting child nodes.

The impurity reduction measures the decrease in impurity achieved by splitting the data at a particular value of the predictor. Different impurity measures can be used, including Gini index, entropy, and classification error. The impurity reduction is calculated as the difference between the impurity of the parent node and the weighted average of the impurity of the child nodes, where the weights are proportional to the number of observations in each child node. The split value that achieves the highest impurity reduction is selected as the optimal split point for the predictor.

Therefore, the correct answer to the question is B) Impurity reduction.

7) Which of the following data mining tasks can be modeled using decision trees?

- ○ Classification task only
- ○ Prediction task only
- ○ Both classification and prediction task
- ○ None of the above

(C)

Both classification and prediction tasks can be modeled using decision trees. Therefore, the correct answer is C) Both classification and prediction tasks.

In a classification task, the goal is to predict a categorical target variable based on a set of predictor variables. Decision trees can be used to create a model that predicts the class label of a new observation based on its predictor variables. The tree structure consists of decision nodes that split the data based on the predictor variables and leaf nodes that represent the predicted class labels.

In a prediction task, the goal is to predict a continuous target variable based on a set of predictor variables. Decision trees can be used to create a model that predicts the value of a new observation based on its predictor variables. The tree structure consists of decision nodes that split the data based on the predictor variables and leaf nodes that represent the predicted value of the target variable.

Therefore, decision trees can be used for both classification and prediction tasks.

8) Which of the following steps is used to fit the decision tree to the predictors' information and not to the noise?

- ○ Splitting
- ○ Pruning
- ○ Partitioning
- ○ None of the above

(B)

Pruning is a technique used to prevent overfitting of the decision tree to the training data by reducing its complexity. It involves removing branches or nodes from the tree that do not improve its performance on the validation data. Pruning helps to generalize the tree better to new data and improves its predictive accuracy.

During the growing phase of the decision tree, the tree is grown to its maximum depth, which may lead to overfitting. Pruning is then performed by removing some of the branches or nodes from the tree, which may not be necessary for classification or prediction. Pruning is done by evaluating the performance of the tree on a separate validation dataset or by using cross-validation techniques.

Therefore, the correct answer is B) Pruning.

9) Compute the entropy measure value for the following case:
There are 12 students who had taken a test. But only 9 students could pass the test and the remaining failed.

- ○ 0.727
- ○ 1.20
- ○ 0.189
- ○ 0.811

(D)

To compute the entropy measure value for the given case, we can use the formula:

Entropy = -p(success) log2 p(success) - p(failure) log2 p(failure)

where p(success) is the proportion of students who passed the test, and p(failure) is the proportion of students who failed the test.

In this case, there are 12 students in total, and 9 of them passed the test, so the proportion of students who passed the test is:

p(success) = 9/12 = 0.75

Similarly, the proportion of students who failed the test is:

p(failure) = 3/12 = 0.25

Now, we can substitute these values in the entropy formula:

Entropy = -0.75 log2 0.75 - 0.25 log2 0.25

Using a calculator, we can simplify this expression as:

Entropy ≈ 0.811

Therefore, the entropy measure value for the given case is approximately 0.811, and the correct answer is D) 0.811.

10) While growing decision trees, nodes are formed for each recursive partition within the multi-dimensional space of predictors. The final nodes of fully grown decision tree corresponding to the final homogenous groups are referred as:

- ○ Decision nodes
- ○ Root node
- ○ Terminal nodes
- ○ None of the above

(C)

The final nodes of a fully grown decision tree, corresponding to the final homogeneous groups, are referred to as terminal nodes. Therefore, the correct answer is C) Terminal nodes.

Decision nodes are the internal nodes of a decision tree that split the data into smaller subgroups based on a selected predictor. Root node is the topmost decision node of a decision tree. It represents the entire dataset before any splitting has occurred.