Business analytics and data mining Modeling using R: Week-9

25 March 2023 23:31

- Week 0
- Week 1
- Week 2
- Week 3
- Week 4
- Week 5
- Week 6
- Week 7
- Week 8
- Week 9
- Lecture 41 CLASSIFICATION AND REGRESSION TREES PART-6
- Lecture 42 PRUNING PROCESS
- Lecture 43 PRUNING PROCESS PART-2
- Lecture 44 PRUNING PROCESS PART-3
- Lecture 45 REGRESSION TREES
- Quiz: Week 9 : Assignment 9

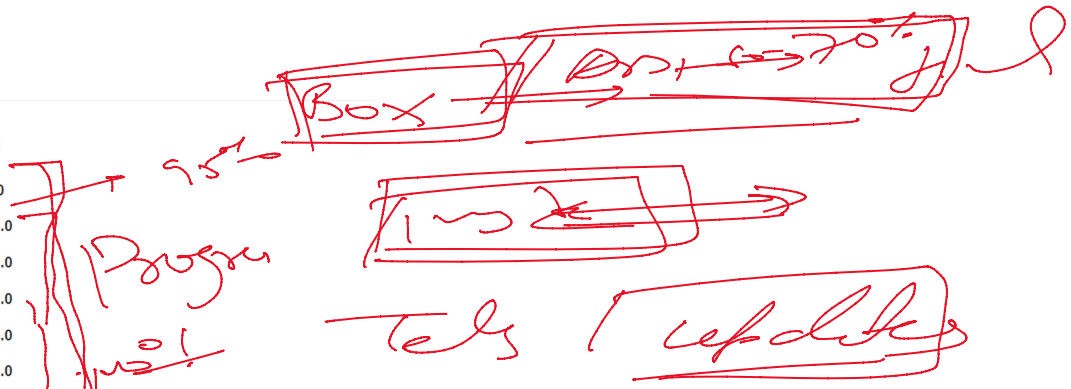
Name: Aakash

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: **100.0**
Week 9 : Assignment 9: -
Week 10 : Assignment 10: -

Announcement:
You are currently receiving course related emails. Click here to unsubscribe.

Discussion forum:



1) What is the range of gini values for the binary classification tasks?

- ○ {0, 0.5}
- ○ {0, (m-1)/m}
- ○ {0, 1}
- ○ {0, m/(m-1)}

(A)

The correct answer is {0, 0.5}.

The Gini index is commonly used as a measure of impurity in decision trees, particularly for binary classification problems. The Gini index ranges from 0 to 1, where 0 represents a completely homogeneous set (i.e., all members of the set belong to the same class) and 1 represents a completely heterogeneous set (i.e., the members of the set are evenly distributed across all classes).

In binary classification, there are only two classes, so the maximum heterogeneity occurs when the two classes are evenly split, which gives a Gini index value of 0.5. Therefore, the range of Gini values for binary classification tasks is {0, 0.5}.

The other options provided in the question are valid ranges for Gini index in different contexts, but not for binary classification tasks. For example, the range {0, (m-1)/m} is used for multiclass classification with m classes, and the range {0, 1} is used for regression problems. The range {0, m/(m-1)} is not a valid range for Gini index in any context, as it would allow the Gini index to exceed 1, which is not possible.

2) What is the range of entropy values for the binary classification tasks?

- ○ {0, log2(m)}
- ○ {0, 0.5}

2) What is the range of entropy values for the binary classification tasks?

- ○ {0, log2(m)}
- ○ {0, 0.5}
- ○ {0, 1}
- ○ None of the above.

(C)

The range of entropy values for binary classification tasks is {0, 1}.

Entropy is another measure of impurity that is commonly used in decision trees, particularly for binary classification problems. Entropy ranges from 0 to 1, where 0 represents a completely homogeneous set (i.e., all members of the set belong to the same class) and 1 represents a completely heterogeneous set (i.e., the members of the set are evenly distributed across all classes).

In binary classification, there are only two classes, so the maximum heterogeneity occurs when the two classes are evenly split, which gives an entropy value of 1. Therefore, the range of entropy values for binary classification tasks is {0, 1}.

The range {0, log2(m)} is used for multiclass classification with m classes, and the range {0, 0.5} is not a valid range for entropy values.

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

- ○ 0
- ○ 1
- ○ -1
- ○ 0.5

(B)

When there are two classes with equal representation of each class, the entropy measure value will be 1.

Entropy is a measure of impurity that is calculated using the following formula:

Entropy = -p1 * log2(p1) - p2 * log2(p2)

where p1 and p2 are the proportions of the two classes in the dataset.

When the two classes are equally represented, p1 = p2 = 0.5. Substituting these values in the formula, we get:

Entropy = -0.5 * log2(0.5) - 0.5 * log2(0.5)
= -0.5 * (-1) - 0.5 * (-1)
= 1

Therefore, the entropy measure value will be 1 when there are two classes with equal representation of each class.

4) Which of the following are true about both linear and logistic regression?

- ○ One or more independent variables
- ○ Same estimation method
- ○ Single dependent variable
- ○ None of the above

(A)

Linear regression is used to model the linear relationship between a dependent variable and one or more independent variables, whereas logistic regression is used to model the relationship between a binary dependent variable and one or more independent variables. Both linear and logistic regression models can have one or more independent variables.

However, C) Single dependent variable is not true for logistic regression, as it models the relationship between a binary dependent variable and one or more independent variables.

5) In binary classification, a cutoff value of 0.5 means that cases with an estimated probability, P(Y=1) > 0.5 are classified to:

- ○ Class 1
- ○ Class 0
- ○ Both a and b
- ○ None of the above

(A)

The answer is (A) Class 1.

In binary classification, the predicted output is either 0 or 1, representing the negative and positive class, respectively. The output is obtained by comparing the estimated probability, P(Y=1), to a cutoff threshold. If P(Y=1) is greater than or equal to the cutoff threshold, the output is 1 (positive class); otherwise, the output is 0 (negative class).

In this case, the cutoff value is set to 0.5. Therefore, if P(Y=1) > 0.5, the output will be 1 (positive class), and if P(Y=1) < 0.5, the output will be 0 (negative class). Hence, option (A) is the correct answer.

6) Which of the following are not true about logistic regression?

- ○ Least squares method is used
- ○ Maximum likelihood method is used
- ○ Instead of using outcome variable (y) in the model, a function of y, called logit is used
- ○ None of the above

(A)

The answer is (A) Least squares method is used.

Logistic regression is a statistical method used for binary classification problems, where the outcome variable (y) takes only two values (0 or 1). The goal is to find a relationship between the input variables (X) and the probability of the outcome variable being 1. Unlike linear regression, logistic regression does not use the least squares method for parameter estimation. Instead, it uses the maximum likelihood method to estimate the parameters of the logistic regression model.

In logistic regression, the logit function of the probability of y=1 is modeled as a linear function of the input variables. The logit function is defined as the natural logarithm of the odds ratio of the probability of y=1 to the probability of y=0. Therefore, option (C) is true.

Therefore, the correct answer is (A) Least squares method is used.

7) Which of the following can be true when we add a new variable in the linear regression model?

- ○ R-squared and adjusted R-squared both increase
- ○ R-squared increases and adjusted R-squared decreases
- ○ R-squared decreases and adjusted R-squared decreases
- ○ R-squared decreases and adjusted R-squared increases

(A)

When a new variable is added to a linear regression model, the R-squared value of the model generally increases because the model now has additional information to explain the variation in the dependent variable. This increase in R-squared value can be interpreted as an improvement in the goodness of fit of the model.

The adjusted R-squared value also increases when a new variable is added to the model if the increase in R-squared value is significant enough to compensate for the penalty for including an additional variable. The adjusted R-squared value is designed to avoid overfitting, and it will increase if the new variable improves the model significantly.

Therefore, the correct answer is (A) R-squared and adjusted R-squared both increase when a new variable is added in the linear regression model.

8) Which of the following is correct about the training of linear regression models based on below given statements?
I: Overfitting is more likely if we have less data
II: Overfitting is more likely when the set of all possible mappings of inputs to outputs is small.

- ○ Both are False
- ○ I is False and II is True
- ○ I is True and II is False
- ○ Both are True

(C)

The correct answer is (C) I is True and II is False.

Overfitting occurs when a model is too complex and captures the noise in the training data, which leads to poor performance on new data. Training a linear regression model can be affected by overfitting, and the risk of overfitting depends on various factors.

Statement I: Overfitting is more likely if we have less data.
This statement is true. With less data, the model may capture the noise in the training data and perform well on the training data but poorly on new data. The model may fit the training data too closely and not capture the underlying patterns in the data, resulting in overfitting.

Statement II: Overfitting is more likely when the set of all possible mappings of inputs to outputs is small.
This statement is false. In fact, overfitting is more likely when the set of all possible mappings of inputs to outputs is large, as the model has more flexibility to fit the noise in the training data. A small set of possible mappings may limit the complexity of the model and reduce the risk of overfitting.

Therefore, the correct answer is (C) I is True and II is False.

9) Which type of data is typically modeled by regression trees?

- ○ Linear
- ○ Nonlinear
- ○ Can't say
- ○ None of the above

(B)

Regression trees are used to model nonlinear relationships between a dependent variable and one or more independent variables. Therefore, the correct answer is (B) Nonlinear.

Regression trees are a type of decision tree model that recursively splits the data into subsets based on the values of the independent variables, with the goal of minimizing the variance of the dependent variable within each subset. The splits can be based on categorical or continuous variables and can result in a tree-like structure of decision rules that can be used to predict the value of the dependent variable for new data.

While regression trees can also model linear relationships, they are particularly useful for modeling nonlinear relationships where the effect of one variable on the dependent variable may change depending on the values of other variables.

10) Which type of relationship between the input attribute and output attribute is assumed in simple regression?

- ○ Linear
- ○ Quadratic
- ○ Inverse
- ○ None of the above

(A)

In simple linear regression, a linear relationship is assumed between the input (independent) attribute and the output (dependent) attribute. Therefore, the correct answer is (A) Linear.

Simple linear regression is a statistical method used to model the relationship between two quantitative variables by fitting a linear equation to the observed data. The equation takes the form of $Y = a + bX$, where Y is the dependent variable, X is the independent variable, a is the intercept, and b is the slope. The slope (b) represents the change in Y for a unit change in X and indicates the strength and direction of the linear relationship between X and Y.

While simple linear regression assumes a linear relationship between X and Y, other types of regression models can be used to model nonlinear relationships, such as polynomial regression for quadratic or higher-order relationships, and exponential or logarithmic regression for inverse relationships.