Business analytics and data mining Modeling using R:

Week-6

05 March 2023 16:08

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: -

Week 7 : Assignment 7: -

1) Which of the following is correct about the below given statements?
Assertion (S): The value of adjusted R2 is always less than the value of R2
Reason (R): Adjusted R2 accounts for the number of predictors in multiple linear regression model

- ○ Both S and R are true and R is the correct explanation of S
- ○ Both S and R are true but R is not the correct explanation of S
- ○ S is true but R is false
- ○ S is false but R is true

(A)

The correct option is: Both S and R are true and R is the correct explanation of S.

Explanation:
Adjusted R2 is a modified version of R2 that adjusts for the number of predictors in a multiple linear regression model. Adjusted R2 is always less than or equal to R2. This is because adjusted R2 penalizes the addition of irrelevant predictors to the model, which reduces its value compared to R2. Therefore, Assertion (S) is true.

Reason (R) is also correct because adjusted R2 takes into account the number of predictors in the model, whereas R2 does not. The formula for adjusted R2 includes a penalty term that increases as the number of predictors in the model increases, thereby providing a more accurate estimate of the model's predictive power. Therefore, Reason (R) is the correct explanation of Assertion (S).

Hence, the correct option is Both S and R are true and R is the correct explanation of S.

2) Which of the following is true about the best value of 'k' in kNN when working with data having complex and irregular structures?

- ○ Value of 'k' should be on the higher side
- ○ Value of 'k' should be on the lower side
- ○ Value of 'k' should be equal to the total number of observations in the dataset
- ○ The value of 'k' has no impact

(B)

When working with data having complex and irregular structures, the value of 'k' in kNN should be on the lower side.

Explanation:

The k-nearest neighbor (kNN) algorithm is a non-parametric algorithm used for both classification and regression tasks. The value of k represents the number of nearest neighbors that are considered while making a prediction for a new data point.

When the data has complex and irregular structures, the local neighborhoods can be quite different from each other, and a smaller value of k is better suited to capture this local heterogeneity. A smaller value of k will allow the algorithm to better capture the local structure of the data and make more accurate predictions.

On the other hand, when the data has simple structures, a higher value of k can be used without losing too much accuracy. In general, it is always recommended to choose an odd value of k to avoid ties in the voting process.

Therefore, the correct answer is that the value of 'k' in kNN should be on the lower side when working with data having complex and irregular structures.

3) Which of the following statements is incorrect with respect to adjusted R-squared value?

- ○ Higher the number of predictors, higher the adjusted R-squared value
- ○ Adjusted R-squared uses a penalty on the number of predictors
- ○ Higher values of adjusted R-squares indicate better fit
- ○ None of the above

(A)

The statement "Higher the number of predictors, higher the adjusted R-squared value" is incorrect with respect to adjusted R-squared value.

Explanation:

Adjusted R-squared is a modified version of R-squared that adjusts for the number of predictors in a multiple linear regression model. The formula for adjusted R-squared is:

Adjusted R-squared = 1 - [(1 - R-squared) * (n - 1) / (n - p - 1)]

where n is the number of observations and p is the number of predictors.

The penalty term in the formula for adjusted R-squared increases as the number of predictors increases. Therefore, as the number of predictors increases, the adjusted R-squared value will increase only if the increase in R-squared due to the addition of new predictors is more than offset by the penalty term.

Hence, the statement "Higher the number of predictors, higher the adjusted R-squared value" is incorrect. The correct statement is that adjusted R-squared uses a penalty on the number of predictors and higher values of adjusted R-squares indicate a better fit. Therefore, the incorrect statement is the first statement "Higher the number of predictors, higher the adjusted R-squared value".

4) Which of the following linear regression algorithms can be used for variable selection and dimension reduction?

- ○ Exhaustive search
- ○ Partial iterative search
- ○ Both A and B
- ○ None of the above

(C)

Both Exhaustive search and Partial iterative search linear regression algorithms can be used for variable selection and dimension reduction.

Explanation:

Variable selection is the process of selecting a subset of the original features to use in a model. Dimension reduction is the process of transforming the original features into a lower-dimensional space while retaining as much information as possible.

Exhaustive search is a linear regression algorithm that involves testing all possible combinations of predictors to identify the best subset of predictors for the model. It is a variable selection technique that can also be used for dimension reduction.

Partial iterative search is another linear regression algorithm that involves iteratively selecting and removing predictors from the model based on their statistical significance. This is also a variable selection technique that can be used for dimension reduction.

Therefore, both Exhaustive search and Partial iterative search linear regression algorithms can be used for variable selection and dimension reduction. Hence, the correct option is (C) Both A and B.

5) Which of the following partial iterative search algorithms start with the full model?

- ○ Forward selection
- ○ Backward selection
- ○ Exhaustive search
- ○ Stepwise regression

(B)

Backward selection is the partial iterative search algorithm that starts with the full model.

Explanation:

Backward selection is a variable selection technique that starts with the full model (i.e., all predictors included) and iteratively removes the least significant predictor until a stopping criterion is met. In each iteration, the predictor with the highest p-value (i.e., least significant) is removed from the model.

Forward selection, on the other hand, starts with the null model (i.e., no predictors included) and iteratively adds the most significant predictor until a stopping criterion is met.

Stepwise regression is a combination of both forward and backward selection. It starts with a null model and iteratively adds and removes predictors based on their statistical significance until a stopping criterion is met.

Exhaustive search, as mentioned in the previous answer, involves testing all possible combinations of predictors to identify the best subset of predictors for the model.

Therefore, the correct option is (B) Backward selection is the partial iterative search algorithm that starts with the full model.

6) Which of the following algorithms overlooks the pairs or groups of predictors that perform well together but perform poorly as single predictors?

- ○ Forward selection
- ○ Backward selection
- ○ Exhaustive search
- ○ None of the above

(A)

Forward selection is a variable selection technique that starts with the null model and iteratively adds the most significant predictor until a stopping criterion is met. This technique can identify pairs or groups of predictors that perform well together, as it considers all possible combinations of predictors.

In each iteration, the predictor with the highest increase in R-squared value is added to the model. This allows for the identification of pairs or groups of predictors that work well together, as the increase in R-squared value may be due to the interaction between predictors.

Backward selection, on the other hand, starts with the full model and iteratively removes the least significant predictor until a stopping criterion is met. As I mentioned earlier, this technique can overlook the pairs or groups of predictors that perform well together but perform poorly as single predictors.

Exhaustive search is another variable selection technique that involves testing all possible combinations of predictors to identify the best subset of predictors for the model. This technique can also identify pairs or groups of predictors that perform well together.

Therefore, the correct answer is (A) Forward selection.

7) What would be the Euclidean distance between the following data points with 4 predictors: S (3,5,2,8) and T (1,4,6,2)

- ○ 16.15
- ○ 7.54
- ○ 5
- ○ 13

(B)

To calculate the Euclidean distance between two data points with 4 predictors, we can use the following formula:

d = sqrt((x2-x1)^2 + (y2-y1)^2 + (z2-z1)^2 + (w2-w1)^2)

where:

- x1, y1, z1, and w1 are the values of the first data point for the 4 predictors
- x2, y2, z2, and w2 are the values of the second data point for the 4 predictors
- d is the Euclidean distance between the two data points

Using this formula, we can calculate the Euclidean distance between S (3,5,2,8) and T (1,4,6,2) as follows:

d = sqrt((1-3)^2 + (4-5)^2 + (6-2)^2 + (2-8)^2)
= sqrt(4 + 1 + 16 + 36)
= sqrt(57)
≈ 7.55

Therefore, the Euclidean distance between S (3,5,2,8) and T (1,4,6,2) is approximately 7.55.

8) Which of the following is highly likely when using a high value of k in k-NN technique?

- ○ Fitting to local patterns
- ○ Fitting to global patterns
- ○ Fitting to noise
- ○ None of the above

(B)

When using a high value of k in the k-NN technique, the algorithm is highly likely to fit to global patterns. This is because a higher value of k means that the algorithm considers a larger number of neighboring data points in the classification or regression task. This can result in a smoother decision boundary and a more generalized model that captures the overall patterns in the data, rather than overfitting to local variations or noise.

On the other hand, using a low value of k can lead to overfitting to local patterns or noise in the data. This is because the algorithm may be too sensitive to the exact positions of nearby data points, rather than the overall trends in the data.

Therefore, the correct answer is (B) Fitting to global patterns.

9) Which of the following scenario is regarded as a naïve rule in k-NN?

- ○ When k = 1
- ○ When k > 1
- ○ When k = n (where 'n' is the number of total observations)
- ○ When 1 < k < n (where 'n' is the number of total observations)

(C)

When k = n, the algorithm simply assigns the class or label of the new data point to the most common class or label in the entire dataset, since all data points are equally distant from the new data point. This approach completely ignores the features and patterns in the data and can result in poor generalization performance.

Using smaller values of k (such as 1 < k < n) can help to better capture the local patterns in the data and improve the performance of the model. However, choosing a value of k that is too small can also result in overfitting to the noise or outliers in the data.

Therefore, the correct answer is (C) When k = n (where 'n' is the number of total observations).

10) Which of the following is true when k-NN is used for prediction tasks rather than classification tasks?

- ○ Computation of distance between the new observation and training partition records is different
- ○ Value of new record is determined using weighted average of all the k-nearest records
- ○ Value of new record is determined using weighted average of the records belonging to the dominant class
- ○ Overall misclassification error is used as performance metric

(B)

When k-NN is used for prediction tasks rather than classification tasks:

- The computation of distance between the new observation and training partition records is the same as in the classification task.
- The value of the new record is determined using the weighted average of the numerical values of the k-nearest records. The weight assigned to each record is proportional to its distance from the new record, such that closer records have a higher weight and more influence on the prediction than farther records.
- There is no concept of dominant class in regression tasks, so the weighted average of the numerical values of the k-nearest records is used to estimate the value of the new record.
- The performance metric used for regression tasks is typically the mean squared error (MSE) or the root mean squared error (RMSE), which measure the difference between the predicted values and the true values of the test set.

Therefore, the correct statement is: the value of the new record is determined using the weighted average of the numerical values of the k-nearest records.