Business analytics and data mining Modeling using R:

Week-7

12 March 2023 05:09

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: **-**

1) Which technique has a step for finding training partition records that have the same predictor values as the new observation?

- ○ Naïve Bayes
- ○ k-NN
- ○ Complete Bayes
- ○ Multiple linear regression

(C)

In Complete Bayes, the algorithm estimates the probability distribution of the response variable given the predictor values using Bayes' theorem and assumes a prior probability distribution for the predictor values. When a new observation is presented, the algorithm identifies training partition records with the same predictor values as the new observation, and then uses those records to update the probability distribution of the response variable for the new observation. Therefore, finding training partition records with the same predictor values is a crucial step in the Complete Bayes algorithm.

2) Which of the following are correct about the statements given below?

Assertion (S): Complete Bayes model is not suitable for prediction task

Reason (R): It is difficult to find training records with exact values of predictors as in the new record

- ○ 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 assertion that "Complete Bayes model is not suitable for prediction task" (S) is true, as Complete Bayes suffers from the "curse of dimensionality" problem. As the number of predictors increases, the size of the space of possible predictor values grows exponentially, making it increasingly difficult to find training records with the same predictor values as the new record. This can lead to poor prediction performance.

The reason given for S, that "it is difficult to find training records with exact values of predictors as in the new record" (R), is also true, and it provides the correct explanation for why the assertion is true. The difficulty in finding training records with exact values of predictors as in the new record is a key reason why Complete Bayes is not suitable for prediction tasks.

3) A dataset consists of 800 records of whether a flight was delayed or not due to weather issues (clear weather or bad weather). Out of 800 flights, 600 flights were delayed and 200 flights were on time. Out of the total flights that got delayed, 450 flights faced bad weather. For the remaining flights which were on time, only 25 flights faced bad weather. Compute the conditional probability (exact Bayes) of the flights that got delayed given the weather was clear. ***1 point***

- ○ 0.25
- ○ 0.46
- ○ 0.75
- ○ 0.54

(B)

To compute the conditional probability of flights being delayed given the weather was clear, we need to apply Bayes' theorem.

Let D be the event that a flight is delayed, and C be the event that the weather is clear. We need to find P(D|C), i.e., the probability that a flight is delayed given that the weather is clear.

By Bayes' theorem:

P(D|C) = P(C|D) * P(D) / P(C)

We can compute the values of each term as follows:

- P(C|D): the probability of clear weather given that a flight is delayed. We don't have this information directly, but we can use the information given in the problem to compute it as follows:

P(C|D) = (total number of delayed flights with clear weather) / (total number of delayed flights)
= (150/600)
= 0.25

- P(D): the overall probability of a flight being delayed. This is given in the problem as 600/800 = 0.75.
- P(C): the overall probability of clear weather. We can compute this by considering all the cases where the weather was clear, whether the flight was delayed or not:

P(C) = (total number of flights with clear weather) / (total number of flights)
= (200 + 150) / 800
= 0.44

Note that we included the flights with clear weather that were on time (25 + 175 = 200) in the numerator.

Now we can plug these values into Bayes' theorem:

P(D|C) = P(C|D) * P(D) / P(C)
= 0.25 * 0.75 / 0.44
= 0.425

Therefore, the conditional probability of flights being delayed given the weather was clear is approximately 0.425, which is closest to option B) 0.46.

4) Which of the following statements related to Bayes model is incorrect?

- ○ Bayes model is more suitable for classification task rather than prediction task
- ○ When number of predictors is more, Exact Bayes model is more suitable
- ○ It is preferable to use categorical predictors for computing Bayes model
- ○ Numerical predictors should be preferably converted into categorical predictors for Bayes model

(B)

Bayesian models are statistical models that use Bayes' theorem to estimate probabilities of an event based on prior knowledge or information. In the context of machine learning, Bayesian models can be used for both classification and prediction tasks. However, the suitability of different types of Bayesian models may depend on various factors, such as the number of predictors, the type of data, and the availability of prior knowledge.

In the case of the statement "When the number of predictors is more, Exact Bayes model is more suitable", this is actually incorrect. The exact Bayesian model is based on Bayes' theorem and involves computing the joint probability of all the predictors given the outcome variable. As the number of predictors increases, the computation required for the exact Bayesian model becomes more complex and it may become impractical to compute. Therefore, when the number of predictors is high, approximations such as Naive Bayes or Bayesian network models are often used.

Naive Bayes is a popular and simple probabilistic algorithm that assumes independence between predictors. This assumption simplifies the computation required for the Bayesian model and makes it computationally efficient, even when the number of predictors is high. Naive Bayes is commonly used in text classification and spam filtering.

Bayesian network models are another type of Bayesian model that can handle a large number of predictors. Bayesian network models represent the relationships between the predictors as a directed acyclic graph (DAG) and use conditional probabilities to estimate the probability of the outcome variable. Bayesian network models are often used in medical diagnosis, image analysis, and expert systems.

In summary, the suitability of Bayesian models for a given task depends on various factors, and the exact Bayesian model may not be suitable when the number of predictors is high. Instead, approximate methods such as Naive Bayes or Bayesian network models are often used.

5) Which of the following are correct about the below given statements?

I: Naïve Bayes model is suitable for classification problems.

II: The probability values in Naïve Bayes, should be accurate in absolute terms.

- ○ Statement I is true and Statement II is false
- ○ Statement II is true and Statement I is false
- ○ Both the statements are true
- ○ Both the statements are false

(A)

The correct option is A) Statement I is true and Statement II is false.

I: Naïve Bayes model is suitable for classification problems.
This statement is true. Naive Bayes is a popular algorithm for classification tasks. It is a probabilistic algorithm that calculates the probability of each class given a set of predictor variables. Naive Bayes is commonly used for text classification, spam filtering, sentiment analysis, and other classification tasks.

II: The probability values in Naïve Bayes, should be accurate in absolute terms.
This statement is false. The probability values in Naive Bayes are not required to be accurate in absolute terms. Naive Bayes algorithm is based on the Bayes theorem, which estimates the probability of the class given the predictor variables. Naive Bayes uses the prior probability and likelihood to estimate the posterior probability of the class. The accuracy of the probability values depends on the quality and quantity of the data used to train the model.

In summary, Naive Bayes is suitable for classification problems and does not require absolute accuracy of probability values. The correct option is A) Statement I is true and Statement II is false.

6) Which of the following statements are true with respect to Naïve Bayes model?

- ○ Predictor values must not be independent of each other for a given class
- ○ Absolute accuracy in actual probability values is essential to classify a new record
- ○ Reasonable accuracy in rank ordering of probability values is required to classify a new observation
- ○ Computation of denominator in Naïve Bayes formula impacts the rank ordering of probability values

(C)

C) Reasonable accuracy in rank ordering of probability values is required to classify a new observation.

In the Naive Bayes model, the algorithm calculates the posterior probability of each class given the predictor variables for a new observation. The classifier assigns the new observation to the class with the highest posterior probability. Therefore, it is crucial to have a reasonable accuracy in the rank ordering of the probability values to accurately classify a new observation.

A) Predictor values must not be independent of each other for a given class.
This statement is incorrect. The Naive Bayes algorithm assumes that the predictor variables are conditionally independent given the class variable. Although this assumption is not strictly true, the Naive Bayes algorithm often performs well in practice even when the assumption is violated.

B) Absolute accuracy in actual probability values is essential to classify a new record.
This statement is also incorrect. Naive Bayes algorithm does not require absolute accuracy in actual probability values to classify a new record. Instead, the algorithm focuses on relative probabilities to make decisions. Therefore, the rank ordering of probability values is more important than their absolute values.

D) Computation of denominator in Naive Bayes formula impacts the rank ordering of probability values.
This statement is incorrect. The computation of the denominator in the Naive Bayes formula, which is a normalization constant, scales the posterior probabilities such that they sum to one. It does not affect the rank ordering of probability values. Instead, the likelihood of the predictor variables for each class is what determines the rank ordering of probability values.

In summary, only option C is correct, and options A, B, and D are incorrect.

7) For which of the following scenarios, it is acceptable to misclassify a few truthful financial reports as fraudulent reports?

- ○ When the goal is to reduce overall misclassification error
- ○ When the goal is to accurately identify records belonging to a specific class of interest
- ○ When the goal is to maximize overall classification accuracy
- ○ None of the above

(B)

8) Which of the following about cut-off value is suitable for identifying maximum records belonging to a specific class of interest?

- ○ Below 0.5
- ○ Above 0.5
- ○ Equal to 0.5
- ○ None of the above

(A)

9) Which of the following tools are suitable for computing conditional probability?

- ○ Pivot Table
- ○ Conditional Formatting
- ○ Pie chart
- ○ None of the above

(A)

Pivot Table is a feature in spreadsheet applications like Excel and Google Sheets that allows users to summarize and analyze data. It can be used to calculate conditional probabilities by specifying the appropriate row and column labels, as well as the appropriate aggregate function (e.g., sum, count, average).

For example, suppose we have a dataset of customer purchases that includes information on the customer's gender and the product category. We can use Pivot Table to calculate the conditional probability of a customer purchasing a product in a specific category given that the customer is male or female.

Therefore, the correct answer to the question is "A) Pivot Table."

10) Which of the following are limitations of Naïve Bayes method?

- [ ] Performs well mainly for classification of records
- [ ] Estimation of actual probabilities of a class
- [ ] Requires large number of records to obtain good results
- [ ] None of the above

(A,B,C)

Naïve Bayes is a popular machine learning algorithm used for classification tasks. However, it has several limitations that can affect its predictive performance in certain situations.

The limitations of Naïve Bayes are:

A) Performs well mainly for classification of records: Naïve Bayes works well for classification problems, especially when there are many predictors and the data is high-dimensional. However, it may not be as effective in situations where there is a small sample size, or when the predictors are highly correlated with each other.

B) Estimation of actual probabilities of a class: Naïve Bayes assumes that the predictors are independent of each other, which may not always be the case in real-world scenarios. As a result, the actual probabilities of a class may be poorly estimated, especially when there are interactions between predictors.

C) Requires large number of records to obtain good results: Naïve Bayes requires a sufficiently large sample size to obtain good results. If the sample size is too small, then the model may suffer from overfitting or underfitting, resulting in poor predictive performance.

In summary, Naïve Bayes is a useful algorithm for many classification problems. However, it is important to keep in mind its limitations and to assess whether it is the appropriate algorithm to use for a particular data set and task.