How can the AUC (Area Under Curve) be calculated in R?

Authored by

stats writer

April 28, 2024

The AUC (Area Under Curve) is a commonly used metric for evaluating the performance of a predictive model. In R, the AUC can be calculated using the "auc" function from the "pROC" package. This function takes in the predicted probabilities and the actual outcome values as inputs and calculates the AUC value. The AUC value represents the overall ability of the model to classify observations correctly, with a higher value indicating better performance. The AUC can also be visualized using a ROC (Receiver Operating Characteristic) curve, where the AUC is equal to the area under the curve. Overall, the AUC calculation in R provides a reliable and efficient way to assess the performance of predictive models.

Calculate AUC (Area Under Curve) in R

Logistic Regression is a statistical method that we use to fit a regression model when the response variable is binary. To assess how well a logistic regression model fits a dataset, we can look at the following two metrics:

Sensitivity: The probability that the model predicts a positive outcome for an observation when indeed the outcome is positive. This is also called the "true positive rate."Specificity: The probability that the model predicts a negative outcome for an observation when indeed the outcome is negative. This is also called the "true negative rate."

One way to visualize these two metrics is by creating a ROC curve, which stands for "receiver operating characteristic" curve.

This is a plot that displays the sensitivity along the y-axis and (1 - specificity) along the x-axis. One way to quantify how well the logistic regression model does at classifying data is to calculate AUC, which stands for "area under curve."

The closer the AUC is to 1, the better the model.

The following step-by-step example shows how to calculate AUC for a logistic regression model in R.

Step 1: Load the Data

First, we'll load the Default dataset from the ISLR package, which contains information about whether or not various individuals defaulted on a loan.

```
#load dataset
data <- ISLR::Default

#view first six rows of dataset
head(data)

default student balance income
1 No No 729.5265 44361.625
2 No Yes 817.1804 12106.135
```

3 No No 1073.5492 31767.139

4 No No 529.2506 35704.494

5 No No 785.6559 38463.496

6 No Yes 919.5885 7491.559

Step 2: Fit the Logistic Regression Model

Next, we'll fit a logistic regression model to predict the probability that an individual defaults:

```
#make this example reproducible
set.seed(1)

#Use 70% of dataset as training set and remaining 30% as testing set
sample <- sample(c(TRUE, FALSE), nrow(data), replace=TRUE, prob=c(0.7,0.3))
train <- data
test <- data

#fit logistic regression model
model <- glm(default~student+balance+income, family="binomial", data=train)
```

Step 3: Calculate the AUC of the Model

Next, we'll use the auc() function from the pROC package to calculate the AUC of the model. This function uses the following syntax:

auc(response, predicted)

Here's how to use this function in our example:

```
#calculate probability of default for each individual in test dataset
predicted <- predict(model, test, type="response")

#calculate AUC
library(pROC)
auc(test$default, predicted)
```

```
Setting levels: control = No, case = Yes
Setting direction: controls < cases
Area under the curve: 0.9437
```

Since this value is close to 1, this indicates that the model does a very good job of predicting whether or not an individual will default on their loan.